Draft Guidelines for Preparing Assessment Reports for the Medical Services Advisory Committee

Draft Version 4.0

August 2020

Record of updates

Date	Version	Summary of changes
March 2016	2.0	Technical Guidelines for Preparing Assessment Reports for the Medical Services Advisory Committee – Medical Service Type: Therapeutic
July 2017	3.0	Technical Guidelines for Preparing Assessment Reports for the Medical Services Advisory Committee – Service Type: Investigative
August 2020	Draft 4.0	Draft Guidelines for Preparing Assessment Reports for the Medical Services Advisory Committee

Table of Contents

Preamble

These Guidelines have been developed to provide advice to applicants and assessment groups on the health technology assessment (HTA) **methods** which are used throughout the Medical Services Advisory Committee (MSAC) assessment pathway for requests for public funding that fall within the remit of MSAC (e.g. Medicare Benefits Schedule (MBS) services, national screening programmes, blood products for the National Product List (NPL), and technologies which may be funded via other mechanisms). The current Guidelines are to be used for all requests for new public funding (i.e. first time applications and subsequent re-applications).

For information on **processes** relating to the preparation and assessment of requests for public funding via the MSAC pathway, the reader is referred to the MSAC website[a] for further details. The **Application Form** and **Templates** to be used when preparing MSAC applications and assessment reports, respectively, are also provided on the MSAC website, together with the Commonwealth **HTA Glossary**. To facilitate completion, the Application Form and Templates are cross-referenced directly to the current Guidelines.

Purpose and role of MSAC

The Medical Services Advisory Committee (MSAC) is a non-statutory committee established by the Australian Government Minister for Health in 1998. MSAC appraises medical services, health technologies and/or programs proposed for public funding, and provides advice to Government about the level and quality of evidence relating to the comparative safety, clinical effectiveness, cost-effectiveness, and total cost of providing such services. Amendments and reviews of existing services funded by the Medicare Benefits Schedule (MBS) or other programs (for example, blood products or screening programs) are also considered by MSAC.

MSAC advises the Minister for Health on medical services in relation to:

- the strength of evidence about the comparative safety, clinical effectiveness, cost-effectiveness and total cost of the medical service;
- whether public funding should be supported for the medical service and, if so, the circumstances under which public funding should be supported;
- the proposed MBS item descriptor and MBS fee for the service, where funding is supported through the MBS; and
- other matters related to the public funding of health services, referred by the Minister for Health.

There is no obligation on Government to accept or implement the advice MSAC provides.

Membership of MSAC

MSAC is an independent expert committee comprising professionals from the fields of clinical medicine, health economics and consumer matters. The Minister for Health determines the size and composition of MSAC. Members are drawn from a wide range of experts, constituted from time-to-

[a] msac.gov.au

time to address the likely type of applications for the committee's consideration. The current membership of MSAC is available on the MSAC website http://www.msac.gov.au.

MSAC sub-committees

MSAC is supported by two sub-committees: the PICO Advisory Sub-Committee (PASC) and the Evaluation Sub-Committee (ESC). MSAC also has an Executive Committee (made up of the chairs of MSAC, ESC and PASC, and also the Deputy Chair of MSAC) to manage MSAC activities between formal committee meetings.

Key factors influencing decision making by MSAC

MSAC provides advice to inform the circumstances under which health technologies should be funded, subsidised or made available in the Australian health care system. This advice is to the Minister for Health and relates to the listing of a health technology on the MBS, although the remit of MSAC is broader and includes providing advice for other funding arrangements.

In its considerations, MSAC is primarily influenced by strength and quality of the evidence of the following quantifiable factors:

- Comparative health gain: Assessed in terms of the magnitude and clinical importance of effect. The comparative health gain includes both the effectiveness and the safety of the health technology (Section 2).
- Comparative cost-effectiveness: Results derived typically from a cost-effectiveness or cost-utility analysis (presented as incremental cost-effectiveness ratios), or from a cost-minimisation approach (Section 3).
- Predicted use in practice and financial implications: Presented as the projected annual cost per year to the Australian Government and/or to other funding bodies as relevant to the application (Section 4).

The impact of health technologies on the Australian population may not be limited to quantifiable impacts on health. MSAC decision making is also influenced by additional, less-readily quantifiable factors:

- Equity: The advice to subsidise a health technology may have an impact on the equitable access tothe health technology or health resources by different groups, such as those categorised by age, socioeconomic status or geographical location.
- Personal or other utility: Value derived from the use of a health technology that may not be characterised by improvements in health. For example, value (harms and benefits) associated with a knowledge of a prognosis or diagnosis Technical Guidance 28().
- Presence of effective alternatives: This helps to determine the clinical need for the health technology.
- Other relevant considerations: including organisations impacts, ethical concerns and social aspects (Technical Guidance 29).

In making its decision, MSAC considers the best available evidence. This includes evidence as it is provided in the assessment report, provided by experts and as informed by consumer evidence and perspectives.

Purpose of the guidelines

An assessment report is a document that captures technical details relevant to the assessment of a technology for consideration by MSAC. The Guidelines for Preparing Aassessment Reports for MSAC have been developed, in order to assist the drafting of an assessment report.

While MSAC decision making is influenced by a range of factors, an assessment report is not intended to capture all of these factors. The MSAC process involves multiple inputs, of which the assessment report is the primary source for technical, typically quantifiable evidence.

The technical components of an assessment of a health technology include:

- The derivation of a clinical conclusion that focuses on evidence supporting comparative health impacts.
- An estimate of cost-effectiveness, as informed by the clinical conclusion.
- An estimate of the utilisation of a technology, and the financial implications for the Australian Government or funder.

These are the core elements of the assessment report and the Guidelines, and reflect the key information relevant to MSAC's decision making.

There are other aspects of value that are more difficult to quantify, that are also considered by MSAC in decision making.

- For investigative technologies, there may be circumstances where value in addition to that provided by the clinical conclusion may be required to support a positive MSAC recommendation. These other aspects of value include the benefits and harms associated with the knowledge provided by a test (such as a prognosis or diagnosis), may or may not be health related, and are often qualitative. Guidance for this additional value is provided in these current Guidelines.
- For all technologies, additional relevant factors that may influence MSAC decision making include issues such as equity, implementation issues, organisational issues, social impacts and ethics. Guidance for other relevant factors is provided in these current Guidelines.

These Guidelines do not contain guidance on incorporation of consumer, patient or public engagement. Published results of such engagement, where available, may influence the HTA process outlined in the Guidelines, and may inform an assessment of "other relevant factors". Engagement with consumers, patients or the public for the purpose of MSAC deliberations occurs through mechanisms separate to the drafting of the assessment report.

Notes on these draft guidelines

Previously, methodological advice regarding the preparation and assessment of technologies and services by MSAC were published as two separate documents: *Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee – Medical Service Type: Therapeutic (Version 2.0) March 2016* (referred to as the **Therapeutic Guidelines**); and *Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee – Service Type: Investigative (Version 3.0) July 2017* (referred to as the **Investigative Guidelines**).

In the previous and these (draft) current Guidelines, the following technology definitions apply:

- **Health technology:** A technology used in a health care system — for example, therapeutic services (such as medicines and procedures), medical devices, investigative medical services (such as diagnostic tests and imaging services), equipment and supplies, and organisational and managerial systems. For the purposes of some definitions of this glossary, particularly in relation to existing health technologies, this usual definition is extended to include any medical service, placebo or watchful waiting instead of an active health technology. *For ease of reading the word 'technology' is used throughout the document but applies to all types of technology or services.*
- **Therapeutic technologies:** A type of technology that is claimed to directly improve the health of people receiving it. Nothing else needs to be rendered to achieve the improvement in health outcomes. Examples of therapeutic technologies are devices, medicines, vaccines, procedures, programs or systems.
- **Investigative technologies:** A type of health technology that is claimed to generate clinically relevant information about the individual to whom the service is rendered. To achieve an improvement in health outcomes, this information must result in a change in the clinical management of an intermediate intervention. In this sense, investigative procedures can only indirectly improve health outcomes. Examples of investigative technologies are imaging, pathology, genetic testing, and clinical assessments for diagnosis, prognosis, staging, monitoring, prediction of treatment response, surveillance and cascade screening. *For ease of reading, the word 'test' is used throughout the document as an alternative term for 'investigative technology', but is intended to reflect the broad range of investigative technologies available.*

Following a comprehensive review of the two documents, many areas of duplicated advice were identified, particularly related to defining the PICO (Population Intervention Comparator Outcomes) for an application, and advice for developing economic and budget impact analyses. These (draft) current Guidelines have sought to reduce duplication, by merging the Therapeutic and Investigative Guidelines into one document.

Previous versions of the MSAC Guidelines have been organised into Sections and Subsections, that directly relate to Sections and Subsections (of the same number) in the assessment report templates. This convention has not been continued in these (draft) current Guidelines, for two key reasons:

- Components of these (draft) current Guidelines may be used across multiple Sections of an assessment report, and many components may not be relevant during the assessment of a technology; and
- The assessment report templates have an abbreviated Section and Subsection structure to prompt focused reviews, such that some guidance may be required for assessment of a technology, while results of the evaluation would be provided in a technical report, and not the assessment report.

These (draft) current Guidelines have retained the Section Structure (as presented in Figure 1). Within each Section, there are Technical Guidance Subsections, which are abbreviated to TG throughout the draft Guidelines.

Users of these (draft) current Guidelines will not need to refer to all TG subsections in the Guidelines to prepare a technology application or assessment. Where information is specific to a particular type of technology, this is clearly identified as such. The relevant TG subsections (according to type of technology) are indicated in Figure 1.

Other key changes in the (draft) current Guidelines compared to earlier versions of the Technical Guidelines are:

- a renewed focus on including the most applicable evidence in an assessment report,
- additional advice regarding the framing and presentation of direct or linked evidence to support requests for investigative services,
- incorporation of advice regarding the presentation of evidence and economic modelling for requests for genetic or genomic testing of heritable diseases,
- advice regarding the concept of personal utility as a consideration for investigative technologies, and
- introduction of the concept of an MSAC Reference Case for economic modelling of all technologies and services.
- a restructure of the Guidelines that included Sections A though F to Sections 1 through 5 (consistent with the 2016 PBAC Guidelines structure). The key structural difference is the removal of Section C (translation issues), which are now addressed within the relevant subsections of Section 3 (Economic Evaluation).
- Addition of more detailed guidance to address personal and other utility and other relevant considerations (Section 5).

Navigation of the guidelines

The types of health technologies considered by MSAC are varied, though broadly categorised into therapeutic technologies (consultative services, interventions, devices etc) and investigative technologies (medical tests).

For an assessment of a therapeutic technology, the relevant clinical TG subsections are located in Section 2A (Technical Guidance 6 to Technical Guidance 8). All of the included subsections will be required to perform an assessment. For an assessment of an investigative technology, the relevant clinical TG subsections are located in Section 2B (Technical Guidance 9 to Technical Guidance 16). Not all TG subsections in Section 2B will be relevant for the assessment of a test. The relevant TG subsections will be dependent upon the nature of the test, and the evidence that is available, and is informed by an Assessment Framework (Technical Guidance 9). A guide to the relevant clinical TG subsections for assessing an investigative technology is presented in Figure 4 of Section 2B. Section 5 should be considered for all applications and assessments. Advice that evidence is not available regarding personal, other utility and other relevant considerations, if relevant should be provided.

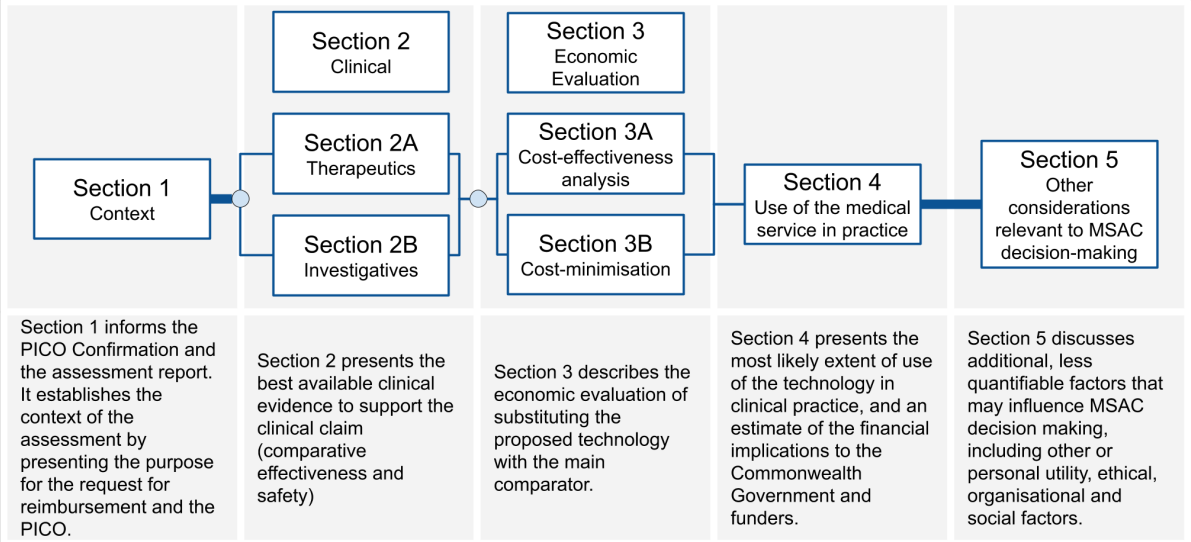


Figure 1 Sections of the guidelines

Glossary of terms used in the guidelines

The purpose of this glossary is to define terms used within these draft Guidelines for public consultation. Additional terminology may also be defined in the HTA advisory committee glossary (http://www.pbs.gov.au/info/industry/useful-resources/glossary). Following public consultation, this glossary may partly or wholly be incorporated in an updated version of the existing HTA glossary.

Algorithm, clinical management	The set of possible clinical management options for a defined population over time, presented according to the subpopulations which receive each option. Often presented in simplified form as a flow diagram, or in more precise form as a decision analysis.
Assessment framework	The analytic framework or logic diagram that is used to illustrate the necessary steps that link the use of an investigative technology (commonly a test) in the target population and the consequences that this may have on health outcome gains.
Assessment questions	The questions addressed by the health technology assessment to inform the overall public funding question.
Biomarker	A characteristic (usually measured by a test) by which a pathological or physiological process (disease, response to treatment etc) can be identified. A biomarker may be defined by the presence or absence of a characteristic, or it may be defined by a quantity of a parameter above or below a specified threshold.
Clinical utility	The net health benefit/harm derived from an investigative health technology
Clinical utility standard	The test and method of interpretation (assay, sample type, thresholds used etc) used to demonstrate clinical utility by being used to allocate patients to alternative options in the key clinical studies generating direct evidence of

	health outcome gains. This replaces the previous term of 'evidentiary standard'.
Direct randomised trial	*Compare with indirect comparison.* A trial in which participants are randomly allocated to groups that receive either the proposed health technology or its main comparator.
Direct from test to health outcomes evidence	Evidence which shows the impact the test has on health outcomes. The alternative to direct from test to health outcomes evidence is a linked evidence approach.
Exchangeability	*Compare with transitivity.* An assessment of exchangeability in an indirect comparison or network meta-analysis considers whether there are any differences with respect to the distribution of any characteristics across the relevant clinical studies that may confound the results of the comparison.
Exemplar	*Compare with facilitated.* The combination of intervention and population for which sufficient evidence is likely to be available as the basis for MSAC to decide its advice on public funding.
Facilitated	*Compare with exemplar.* The combination of intervention(s) and population(s) which is close enough to the exemplar to not require a full HTA. Instead MSAC could decide its advice on public funding based on accepting sufficient similarities between the facilitated combination(s) and the exemplar combination.
Germline	Mutations which occur in the germ cells (eggs and sperm) and are heritable.
Indirect comparison	*Compare with direct comparison*. An analysis that indirectly compares the proposed health technology to its main comparator by comparing one set of trials, in which participants were randomised to receive the proposed health technology or a common reference, with another set of trials, in which participants were randomised to receive the main comparator or the common reference.
Linked evidence	*Compare with direct from test to health outcomes evidence.* When evidence from studies of test accuracy is linked to evidence of change in management and evidence of treatment effectiveness to derive an estimate of the clinical utility of the test.
Multifactorial algorithms	Algorithms which combine multiple factors to determine a person's risk of a future event. Algorithms may be static (learning occurs prior to the dissemination of the technology) or dynamic (learning continues to occur following the dissemination of the technology). Also, machine learning algorithms. Mathematical models built on training data that are used to discover structure within data and/or to predict an output.
Number needed to test	Number of people tested for one person to undergo the intended change in clinical management.
Other utility	Any consequence for the health and well-being of a patient, family members or carers which does not arise from changes in health outcomes quantified in

	the clinical utility evidence. This includes concepts such as the value (or benefits and harms) of knowing, the value of naming, and the impact on patient or family members / carers well-being through being able to plan non-health resources which come at a cost to the person, and sometimes funders, such as transport, accommodation, education, community care (self and others including children), equipment, income loss and insurance. If the benefit is for the patient, this can be termed 'personal utility.
Penetrance	The proportion of individuals for whom traits or characteristics associated with a particular genetic variant will manifest in the phenotype within a specified period of time.
Personal utility	See other utility.
Proband	Individual (index case) in a family who is affected with the disease and has a relevant known germline mutation.
Somatic	Mutations that occur after conception, and are neither inherited nor passed on to offspring.
Standard, clinical reference	*Compare with standard, non-clinical reference.* A reference standard that detects a clinical disorder or clinical outcome of interest.
Standard, non-clinical reference	*Compare with standard, clinical reference.* A reference standard that detects a biomarker, parameter or analyte.
Streamlined	An abbreviated HTA approach used for facilitated population and intervention combinations.
Test	A simplified term for investigative health technology.
Test, cascade	A test of family members of a proband for the identified germline mutation.
Test, diagnostic	A test used to inform or identify a disease, condition or injury.
Test, monitoring	A test used to observe a disease, condition or parameter over time.
Test, predictive	*Compare with prognostic test*. A test which estimates differences in the proportions of individuals in a tested population developing a disease or experiencing a clinical event over time according to different test results (such as test positive and test negative) after altering clinical management in response to one or more of these different test results.
Test, prognostic	*Compare with predictive test*. A test which estimates differences in the proportions of individuals in a tested population developing a disease or experiencing a clinical event over time according to different test results (such as test positive and test negative) without altering clinical management.
Test, screening	A test which is used to detect disease, abnormalities or associated risk factors in asymptomatic members of a population at risk.
Test, staging	A test which is used to classify the severity of a disease.

Test, triage	A test which is used to determine which patients require further tests.
Transitivity	*Compare with exchangeability.* An assessment of transitivity in an indirect comparison or network meta-analysis considers whether there are important differences with respect to the distribution of known treatment effect modifiers across the relevant clinical studies that are likely to confound the results of a comparison.

Presenting an assessment report

The information provided in these guidelines is intended to inform the creation of a PICO Confirmation and an assessment of a health technology. It does not prescribe the most appropriate presentation of an assessment report. The format of the assessment report, and the technical report, is provided in a template accessible on the MSAC website.

MSAC reconsiderations of a health technology

Health technologies that are not recommended for funding may be reconsidered by MSAC if new evidence is provided to address the main concerns raised by MSAC. Subsequent assessment reports should address the concerns raised by MSAC in response to the previous assessment report. Tabulating this, as described in Technical Guidance 4. The information requests in the Guidelines should be followed when providing the new information to support the reconsideration. , Assessment reports for the purpose of a reconsideration should avoid presenting Information that is not disputed. Delete sections of the template that are not required.

Care should be taken when presenting new information that alters the interpretation of results from previous reports.

A key outcome of an assessment report for the reconsideration is a clear presentation and discussion of how the new information addresses the main matters of concern to MSAC.

Section 1 Context

Introduction

This Section provides guidance for establishing the context of an assessment of a medical service. This includes describing the purpose of the application for funding of the technology, developing the Population / Intervention / Comparator / Outcomes (PICO) criteria for use of the technology and the associated assessment questions, and justifying the proposed Medicare Benefits Schedule (MBS) descriptor and fee (where applicable) and addressing personal and other utility and any other considerations.

The MSAC application process may involve different pathways. For most applications that progress to an assessment report, a PICO Confirmation is considered by PASC to focus the assessment report and ensure it is clinically relevant. There are some circumstances under which a PICO Confirmation may not be considered by PASC prior to the development of an assessment report (e.g. re-applications, some referrals from the MBS Reviews Taskforce).

Regardless of whether PASC formally considers a PICO Confirmation, the development of the PICO and assessment questions is a pivotal part of the development of an assessment report. The instructions below will:

- Act as a reference for the details that may be required for an application form;
- Provide guidance for preparing a PICO Confirmation;
- Assist with establishing the PICO in an assessment report where no PICO Confirmation has been required;
- Act as a reference for HTA groups performing a critical appraisal of an assessment report;
- Provide guidance on how to justify changes to an agreed PICO, should the assessment report be required to deviate from the Ratified PICO Confirmation.

Preparing an Assessment Report

In general, if an agreed PICO Confirmation is available this will be reflected in Section 1 of the Assessment Report. The Assessment Report should not deviate from the agreed PICO Confirmation. If changes to the agreed PICO Confirmation items are made, both the agreed PICO Confirmation item and the variation should be presented. The need for the new approach should be justified, noting that any deviations from the agreed PICO Confirmation may affect confidence in the applicability of the evidence and/or analyses presented.

Technical Guidance 1 Purpose of application

TG 1.1 The request for public funding

A clear purpose for the application is necessary to enable meaningful interpretation of the evidence presented in an assessment report. The purpose for the application should be precise and include:

1. The intended use and outcomes of administering the proposed health technology. This statement should include a clear description of the technology and of the proposed purpose of the technology.
2. A clinical claim for the proposed health technology in terms of its impact on health outcomes i.e. whether use of the technology results in superior effectiveness and/or safety compared to current management of the same condition, or whether it is claimed to be non-inferior.
3. A statement justifying the need for the health technology if there is no claimed health advantage. This may occur when:
    a. the medical service involves an investigative technology that is claimed to have other benefits for an individual, family members or carers (Technical Guidance 28 – Other utility and Technical Guidance 29- Other relevant considerations).
    b. the medical service involving the technology produces non-inferior health outcomes, but has additional practical and/or cost advantages over current clinical management.
4. A statement clarifying whether funding is sought under the MBS or another funding source.
5. A statement indicating whether other applications relating to the proposed health technology are in progress (eg, an application to the Therapeutic Goods Administration or Pharmaceutical Benefits Advisory Committee or Prostheses List Advisory Committee).

TG 1.2 Defining the clinical claim

The advice provided by MSAC is primarily based on both the clinical effectiveness and the cost-effectiveness of a health technology compared with current practice. This is referred to as comparative clinical effectiveness, and comparative cost-effectiveness. Acceptable cost-effectiveness means that the additional health or other benefits derived from the health technology are considered to be sufficient to justify any additional costs associated with use of the technology. In making its determination, MSAC is more influenced by the health outcomes and healthcare spending associated with the technology within the healthcare system than impacts outside of the healthcare system (given its role and remit).

In some circumstances, non-health related outcomes may provide additional context for decision making (for example, patient preferences and organisational issues that will affect implementation and use). For more information on other relevant considerations, see Technical Guidance 28 and Technical Guidance 29.

Clinical claims for health technologies

The aim of the majority of health technologies is to have an impact on the health of patients. A service involving a technology, that is intended to replace an existing service, will usually have a positive or neutral impact on health. In some circumstances health technologies will result in a loss of health and will require the consideration of other factors for MSAC to make a favourable funding recommendation.

Health outcomes in this context refers to the aggregate of the patient relevant health outcomes; that is, the net clinical benefit. Separate claims may be reasonable for outcomes typically considered to measure *effectiveness* and for outcomes that measure *safety* or *patient harms*. However, it is important to consider the net benefit of the technology in terms of its effectiveness and safety.

Appropriate clinical claims for health outcomes are:

- The use of the proposed technology results in superior health outcomes compared to the comparator / standard practice.
- The use of the proposed technology results in non-inferior health outcomes compared to the comparator / standard practice.
- The use of the proposed technology results in inferior health outcomes compared to the comparator / standard practice.

Claims of inferior health outcomes are uncommon. Two examples where a claim of inferior health outcomes may be considered are:

- The health technology is less costly, and the magnitude of the health benefit lost is small
- The health technology is more acceptable, or addresses equity issues, such that the uptake is expected to be greater than the current medical service. In this circumstance, it may be reasonable to explore a comparison against "no medical service" for those that would not access the current service.

Establishing an appropriate claim for an investigative technology

An investigative technology generates information about an individual in the target test population. This information is then interpreted, categorised and used for clinical decision making. Clinical decisions may ultimately impact on the health of the patient.

While the impact of a test may ultimately be on health outcomes, the benefits of a test are often described in other ways, such as:

1. An increase in the efficiency or ease of use of a test, with no change to the information derived.
2. An improvement in the efficiency or timing of information provided by a new test that replaces several sequential tests, but in comparison to the multiple tests, no new information would be provided.
3. A reduction in the adverse events associated with testing as a consequence of the new test, which may result in an improvement in health or may support a claim of non-inferiority.
4. An increase in the acceptability or accessibility of a test, such that a broader population would access the test.
5. An improvement in the information provided, such that patients are more accurately categorised with regard to the medical condition (which may or may not lead to improved health outcomes, depending on how each patient is managed).

The clinical claim can be informed by understanding the benefits of the test (relative to current care), and how this is likely to affect patient management, and ultimately their health outcomes (see Table 1). All health technologies, including investigative health technologies (tests) are required to establish a claim that relates to health outcomes, as described above.

Table 1 The possible benefits of tests and suitable claims associated with these benefits

Comparative function	Possible benefits	Effect on management of patient	Health outcomes	Suitable clinical claims (health outcome gains)	Supportive evidence[a]
Test detects the same parameter as the comparator	**Replaces some or all current tests for the same condition** ***May replace no testing for some patients if new test increases coverage*** Faster, cheaper, or more convenient Smaller sample required, reduced re-biopsy rate, possibly safer More accurate More definitive or earlier result, cease or reduce further testing More feasible (panel vs sequential testing)	**No change in management** **OR** **Change in management** Increase in coverage Increase in compliance Change in the patients identified eg earlier in disease process Reduction in subsequent testing Reduction in treatments for adverse events	**No change**	Non-inferior to comparator	Evidence of no change in management *The downstream management will be the same for the same patients as the comparator.*
			If no change in net health benefit, a claim of non-inferiority is appropriate.	Non-inferior to comparator OR	Evidence that overall health is non-inferior
			Average improvement in health	Superior to comparator	Evidence that overall health is superior
Test detects new parameter	**New test replaces current test** Replaces test that detected a different parameter for the same purpose	**No change in management** **OR** **Change in management** Better targets patients to appropriate treatments / subsequent management	**No change**	Non-inferior to comparator	Evidence of no change in management *The downstream management will be the same for the same patients as the comparator.*
	New test, or additional test[b] Confirms diagnosis Confirms diagnosis and provides additional information (e.g. predicts response to treatment) Provides prognostic information (eg, allows treatment planning, resource allocation or value of knowing) New diagnosis or disease state Monitors disease course		**Average improvement in health**	Superior to comparator	Evidence that overall health is superior

a – Particularly for claims of non-inferiority, evidence will be required to support no change in health AND some evidence will be required to support the potential benefits of the test (such as faster, more accurate etc). In the absence of some additional benefit of the test, the purpose for using the test rather than the comparator is unclear. This may help avoid a proliferation of tests that do not provide additional information.

b – For a new test (or an additional test) that results in increased costs there would typically need to be an improvement in health demonstrated. However, if the new test results in a change in management such that there are downstream cost-offsets, a claim of non-inferiority may be possible. Another possibility is that a new test may claim to be non-inferior and the assessment of the test would include other or personal utility considerations, or other relevant considerations, to support the increase in cost.

Hierarchy of claims for investigative technologies

In some circumstances, investigative technologies may provide information that does not markedly affect health in a way that can be quantified, although they may affect personal wellbeing in ways that cannot be attributed to changes in the provision of health resources (Technical Guidance 28).

When deciding on the appropriate claim, it is important to consider a hierarchy of the claims that may be considered by MSAC. Where several claims are possible, claims that are higher in the hierarchy will be more informative for decision making. The exception to this may be if a technology is likely superior in terms of health outcomes, but is no more costly than the comparator. In this case, it may be pragmatic to make a claim of non-inferiority and pursue a more simplified economic approach.

In all circumstances, the **clinical claim** must relate to health outcomes. An assessment report may state that the clinical claim is accompanied by additional relevant considerations or the consideration of other or personal utility (see Section TG 1.4).

Table 2 Evidence of value scale proposed by PASC and MSAC

Rank	Investigative technologies	Claim
1	Direct from test to health outcomes evidence which shows an improvement in health outcomes. Change in clinical practice for at least a proportion of investigated patients and linked evidence of an improvement in health outcomes. (A linked approach is more strongly supported if there is a clear co-dependency with a targeted therapy, but may also be mediated by other means such as a stratification according to risk or prognosis, such as staging of cancer, with consequent changes in subsequent clinical management for which the specific therapies might be less clear or more varied, but for which health outcome improvements could still be shown.)	Superior health outcomes (clinical utility)
1a	Direct from test to health outcomes evidence which shows no change in health outcomes. Linked evidence of no change in health outcomes.	Non-inferior health outcomes or clinical utility
2	Change in clinical practice for at least a proportion of investigative patients, without clear evidence of an improvement in health, but clear evidence or rationale that health is not diminished.	Non-inferior health outcomes or clinical utility
3	Change family planning options (This would apply for testing of heritable mutations only.)	Superior health outcomes or clinical utility (although may be restricted to intermediate outcomes)
4	Be more compelling (definitive, accurate, conclusive) than current investigations, and thus lead to a reduction in current testing (eg diminish the “diagnostic odyssey”).	Non-inferior health outcomes or clinical utility (suitable if test cost is offset by avoiding subsequent tests) Other utility may be appropriate if there is value in knowing a test result earlier. A claim of other utility is required if the proposed test is more costly than the comparator test strategy).
5	Provide a basis for determining a clinical classification and thus informing variation in prognosis or risk, but without changing clinical practice.	Non-inferior health outcomes Other utility (value of knowing)
6	Provide reassurance in a diagnosis or confirming the conclusions of other investigations. This may also include ensuring a more complete diagnostic work-up in the event that a future therapy becomes available which would elevate the test to the above most preferred type of clinical utility.	Non-inferior health outcomes Other utility (value of knowing)

TG 1.3 Comparing healthcare costs

The choice of a clinical claim, and the necessary approach to support that claim, is contingent on both the availability of the evidence to substantiate the claim, and the cost of the new service relative to the cost of current practice. The two most common scenarios are:

- A proposed service that results in a greater cost to the healthcare system should substantiate an overall improvement in health. A claim of superior health outcomes would be required, and the approach would seek to establish the magnitude of these benefits.
- A proposed service that is cost neutral or cost saving to the healthcare system should show at least no loss in health. A claim of non-inferior health outcomes is possible, and in some cases, the approach required may be simplified compared with a claim of superior health outcomes. In the circumstance where the new service is expected to improve overall health but does not cost any more than the existing service, a claim of non-inferiority would be sufficient to support the application.

TG 1.4 Making an additional claim

In some circumstances, the health outcomes derived from the proposed health technology (and resulting changes to downstream management) may be insufficient to justify the incremental cost associated with implementing the technology. In general, such health technologies would not be considered cost-effective, and would require additional evidence to support a positive funding recommendation.

For health technologies that would be regarded as cost-effective, *no additional claim should be made (see Figure 2)*. However, additional relevant considerations may remain an important component of the assessment report and should be presented in Section 5.

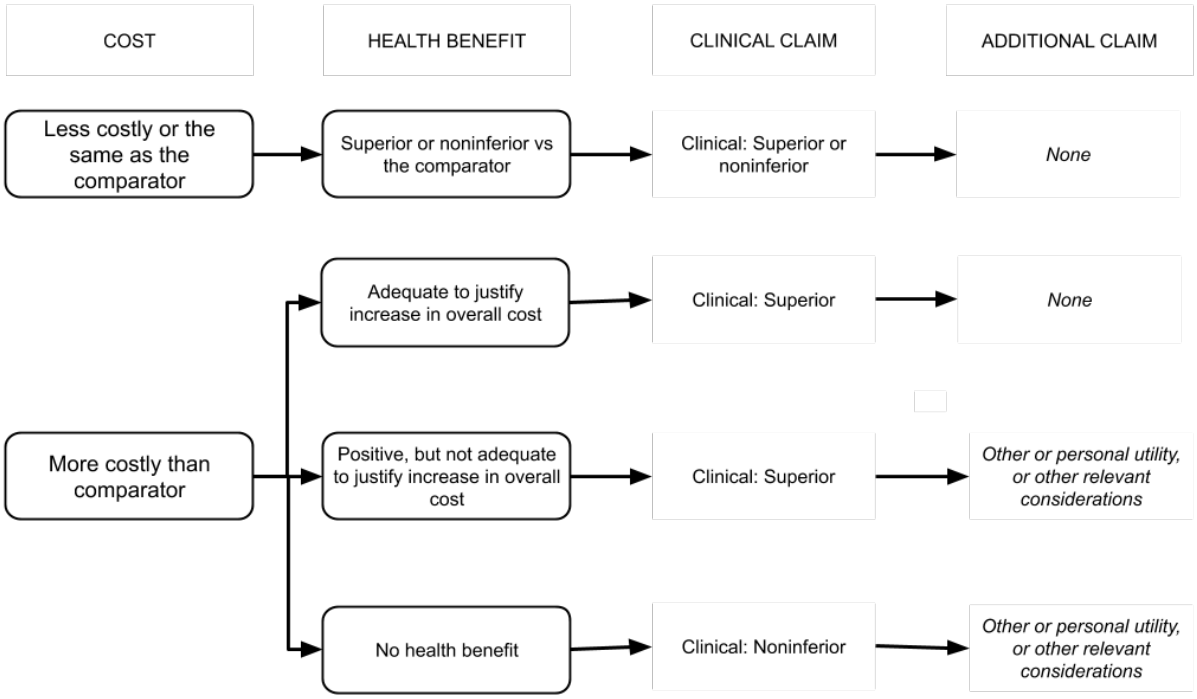


Figure 2 Deciding whether a claim in addition to the clinical claim is required to support the overall cost of the proposed health technology

It is expected that most health technologies that are required to make an additional (non-clinical) claim would be investigative technologies (tests).

If an additional claim is required, first a clinical claim is made (ie, superior, non-inferior or inferior health outcomes) and this is accompanied with a statement that the health technology results in other additional benefits versus the comparator. When presenting the additional claim, briefly state the nature of the additional benefits associated with the use of the (usually investigative) health technology. These benefits (discussed in Technical Guidance 28 and Technical Guidance 29) are presented in Section 5 of the Assessment Report.

Other types of utility

Health technologies (therapeutic or investigative) may result in non-health outcomes, health outcomes that affect others or health outcomes that may be difficult to quantify (such as quality of life related to knowing a diagnosis). These outcomes may be both positive and negative, and may involve a cost component.

Outcomes that are not captured in health outcomes claims may include the following (sometimes overlapping) categories:

- Outcomes that affect others that may have an impact on quality of life, such as spillover effects on carers;
- Non-healthcare sector impacts, such as effects on educational attainment, or attendance at work;
- Other utility outcomes, such as a value (benefits and harms) of knowing or naming a diagnosis.

In most cases, the impact of a health technology that is not captured in health outcomes but that may be influential for MSAC decision-making would be addressed in other relevant considerations (see Technical Guidance 29). Non-health outcomes that have an economic impact (such as a reduced time away from work) are unlikely to be informative in most cases, but may be included as a supplementary economic analysis (see Appendix 10).

Other utility outcomes may impact on the individual (personal utility) or may impact on family members, carers or relatives. Typically, an other or personal utility claim will be informative when a test is used to detect a condition (often a diagnosis or prognosis) for which there is no effective treatment, or that will not result in a change in treatment. The value of the test is then explored in terms of the benefits and harms that arise (for the patient or the family) from the knowledge of the test results. Examples of other utility may include benefits from knowing a prognosis so that patients and families can make appropriate preparations, or harms from a diagnostic label such as insurance and work implications.

Although many of these flow-on effects of the test result could theoretically be quantified in terms of improved quality of life for the patient and their family members, the evidence is unlikely to be generated in this way.

When making a claim of other utility, a parallel claim of health benefits (a clinical claim) *must* be made and tested with evidence. Usually, if a test is relying upon a claim of other utility, the claim for an impact on health related outcomes would be that the test is non-inferior to current practice.

Appropriate claims for tests that affect personal or other utility are:

- The use of the proposed test results in non-inferior health outcomes compared to the comparator / standard practice, but provides additional utility to the patient or their families and carers.
- The use of the proposed test results in superior health outcomes compared to the comparator / standard practice and provides additional utility to the patient or their families and carers.

The PICO required to support claims of other utility must include outcomes that measure both the health related claim (non-inferiority / superiority) and the other utility claim. Outcomes for an other utility claim would include the types of options that become available to, or withdrawn from, an individual, or their family, as a consequence of information provided by a test. These options are not likely to be clinical (otherwise the claim would be for health outcomes), but may include the ability to make preparations, change behaviours, access support etc.

Considering relevant ethical issues

In some circumstances there may be specific ethical issues that will need to be considered that may affect the ability of the assessment report to quantify health outcomes. These typically occur in relation to requests for services related to reproductive planning, requests for services where there is a possibility of incidental findings, or requests that may be impacted by or impact on equity of access to services. When such situations apply they may be discussed in as other relevant considerations in an assessment report (see Section 5).

Technical Guidance 2 PICO

An assessment of the effectiveness, safety and cost-effectiveness of a health technology is informed by the Population, Intervention, Comparator and Outcomes (PICO) that are defined for use of the health technology. The purpose of this TG subsection is to provide an outline of the information required for each part of the PICO.

The PICO guidance will inform the approach for developing a PICO Confirmation, or for defining the PICO during an assessment report where no PICO Confirmation is available. This TG subsection will also be relevant during an assessment report for determining if an amendment to a PICO Confirmation is required.

For the purposes of developing a PICO, scoping searches of the health technology and the medical condition will be required. The identification of existing health technology assessments and systematic reviews may help inform the PICO. Relevant existing HTA reports or Public Summary Documents and high quality systematic reviews or key individual studies discovered during scoping searches should be made available to PASC. If no evidence is identified during these scoping searches PASC should be made aware of the lack of evidence.

TG 2.1 Population

The purpose of describing the relevant population for the technology is to ensure that the information describing the effect of the proposed health technology is restricted to the population of interest. The reported characteristics of the population should include factors that may impact the effect of the proposed health technology should it be adopted in Australia.

Provide an overview of the patient population, disease or condition that is targeted by the proposed health technology. Include relevant details of diagnosis, symptoms, prognosis, demographics and other issues relevant to the population targeted by the technology. State how potentially eligible patients are investigated, managed and referred within the Australian healthcare system prior to the use of the proposed health technology.

If the medical service is proposed for use in a subgroup of a population with a specific condition, describe the characteristics that identify the subgroup and a rationale for targeting the proposed subgroup. Explain which subgroups would be excluded from the target population.

Characterise the Australian population for whom the medical service is intended, such as their age, sex, important comorbidities, and disease- or condition-related characteristics. Summarise the incidence and prevalence of the disease or condition in Australia using data from a reputable source, such as those listed in 'Sources of data for use in generating utilisation estimates'[b]. For investigative technologies, provide the incidence and prevalence of the target population for the test (i.e. those suspected of having the condition being tested for) (see Subsection TG 11.8).

Estimate the size of the population expected to use the proposed health technology and consider whether the proposed public funding will improve equity of access to the service, or whether it will likely be used by those who are already receiving the service (through out-of-pocket expenses or through the states and territories). For further discussion on deriving the prevalence of the disease or analyte, see TG 11.8.

[b] www.pbs.gov.au/info/industry/useful-resources/sources

If the health technology addresses health inequalities (such as resulting from differences in access to care in rural and remote areas, or an area of unmet clinical need etc), these different subgroups should be identified.

If the proposed health technology is intended for use across multiple indications, tabulate these indications. In determining whether an indication is distinct, consider the following characteristics of the indication:

- Differences in the target population
- Differences in the disease, or location of the disease in the body
- Differences in the mechanism of action of the proposed health technology

If the proposed health technology is likely to be used for different purposes, and therefore has distinct indications, evidence to support each indication will usually be required (for example, if it is used for diagnosis as well as monitoring). The exception to this is where an exemplar/facilitated approach is used.

If different populations are to be assessed, using evidence from an 'exemplar population' (i.e. a population where there is a large amount of evidence) and making assumptions regarding how this evidence would apply to 'facilitated populations' (where there is little evidence), a biological rationale is required indicating that the disease/condition in the two populations would behave in a similar manner. For more information on this, see TG 5.2.

It is important to provide detailed information on the natural history of the condition to assist in determining whether the health technology alters that. In describing the prognosis without treatment, classify whether the disease is stable (e.g. portwine stain, lodged foreign body), progressive (e.g. cataracts or many cancers) or spontaneously remitting (e.g. colds, viral rashes), fluctuating (e.g. rheumatoid arthritis, eczema, depression), episodic (e.g. migraine, asthma), or probabilistic (a possible future event, e.g. stroke) (Glasziou, P et al. 2007).

Co-dependent technologies

If an investigative technology being considered by MSAC is co-dependent with a medicine being considered by PBAC, or co-dependent technology with a medical device being considered by PLAC, it is important to distinguish between the population eligible for testing and the population eligible for the treatment with the medicine or medical device. A common error is to assume that the treated population is identical to the tested population, whereas usually the tested population is broader than the treated population.

Genetic testing

If the test is for more than one indication, provide the clinical rationale for the grouping. Provide the Online Mendelian Inheritance in Man (OMIM#) classification of the disease.

Describe the genetic variants associated with the population of interest, and classify which of these are most prevalent, and have the strongest clinical utility and/or cost-effectiveness argument (these are proposed to be the "exemplar genes"). Rarer variants may be "facilitated" (see exemplar/facilitated approach, Technical Guidance 5). MSAC will be most receptive to recommending funding when the pre-test probability of the population tested having a pathological variant is ≥10%. For cancer diseases, eviQ is suggested as a suitable initial source of information.

Gene nomenclatures should be italicised, whereas gene products (proteins) should not. Nomenclature guidelines are available from the Human Genome Variation Society (www.hgvs.org). This includes use of the terms 'variant' rather than 'mutation'. Variants are categorised using 5 classes as per below:

Class	Description
1	Cleary not pathogenic
2	Unlikely to be pathogenic
3	Variant of unknown significance (VUS)
4	Likely to be pathogenic
5	Clearly pathogenic

Describe the nature of the variants (such as deletions and copy number variations). Describe if the variants identified are somatic or germline (hereditary). When a condition is hereditary, testing of an index case may have an impact on the clinical management of family members. Cascade testing is the process of extending genetic testing to biological relatives of an index case with a pathogenic variant. This process is repeated as more affected individuals or pathogenic variant carriers are identified. If cascade testing of first and/or second degree biological family members would occur, or reproductive partners, this should be clearly described. For inherited conditions, describe the pattern of inheritance (i.e. whether dominant, recessive, X-linked etc). Describe the penetrance of the variants (i.e. what proportion of people with the variant express the phenotype associated with it).

If there are differences in the variants found in different ethnic groups, provide information on this, so that the assessment can cover those variants most likely to be relevant to the Australian population, including minority groups.

Populations not receiving the health technology

Health technologies may also have an impact on the broader society, if, for example, an infectious disease is detected, or if harms are caused to clinicians in the process of administering a treatment. Consider discussing these under 'Other relevant considerations'. However, if a key claim of benefit of a technology is the impact it has on family members/carers (i.e. improving their quality of life, allowing them to participate in the workforce instead of being a full time carer etc), then consider including studies reporting on these outcomes (in family members/carers) in the assessment report (clearly distinguished from outcomes for the patient).

TG 2.2 Intervention

Proposed health technologies

Describe the key components of the proposed health technology including whether it is an investigative or therapeutic technology (or both, which may be possible for a co-dependent application and some tests that remove the affected tissue, eg colonoscopy).

Provide details of how the proposed health technology is expected to be used, including frequency of use, mode of delivery, clinical setting, specialist training and provider type. Describe the required infrastructure for use of the technology, and whether the health system is currently able to provide this. State whether the proposed health technology is currently funded (in the public or private setting) in Australia for the same or another clinical indication.

If the request is for a therapeutic technology, describe the mechanism of action and the pathological process(es) the technology is claimed to address. Be explicit about whether the proposed therapeutic technology will be used in addition to existing therapies (add-on), as a replacement for an existing therapy, or will displace an existing therapy to a later line of treatment. If the technology can be used

at different points in the management of the patient (i.e. lines of therapy, or for both diagnosis and monitoring of the patient), then clearly describe when the technology will be used.

If the request is for an investigative technology, describe whether it is a diagnostic test, prognostic test, staging test, predictive test, monitoring test, surveillance test, cascade test, screening test etc (See OHTA Glossary[c] for definitions). Be explicit about whether the proposed test will be used in addition to existing tests (add-on), as a replacement for an existing test, or used prior to existing tests (triage test). Describe any other elements of the test strategy, and how the proposed test would be incorporated into this. Describe how the test is ordered, how the test is performed, how the test results are interpreted (including any cut-off points), how the test is communicated to the patient, and used to inform any clinical decisions.

As the benefit of a test is indirect (i.e. it only influences health indirectly through the information provided by the test results), describe the downstream consequences of the proposed test which support the clinical claim (i.e. if a health benefit is claimed, how this is achieved, e.g what treatments follow positive test results and negative test results). Provide a discussion of the biological plausibility for the impact of the test on patient health outcomes (NB for 'black box' tests this may not be possible[d]). If the downstream consequences of the test are likely to change in the near future due to the availability of other technologies, consider including these in the assessment, even if they are not yet established in the Australian healthcare system.

Consider whether there are any contextual factors that could modify the clinical utility, test accuracy, or the safety of the health technology (such as the 'learning curves' of service providers) that should be assessed. Discuss whether there are likely to be any implementation issues (i.e. a change in the specialty that delivers the technology, sample storage requirements, education and training requirements, changes in access to care, communication between and within organisations etc). Provide details of any Quality Assurance Program or training program in place or required. If the technology is investigative, discuss whether the biomarker is correlated with factors such as ethnicity and sociodemographic status, and whether there are therefore ethical considerations that must be considered.

Co-dependent technologies

Describe any additional elements of the health technology for which funding is also sought (i.e. is it co-dependent with a medicine, such that only patients with a specific genetic variant determined by the proposed test will be eligible for a specific medicine, or co-dependent with a device).

If a co-dependent test-medicine combination is being assessed, ensure that the 'intervention' describes what treatment the biomarker positive patients and the biomarker negative patients would receive. In addition, it is important to describe the treatment that each of these groups would have received in the absence of the proposed test.

Proposed medical services that include a medical device

For medical services that use a medical device (such as surgery to place an implant, or adjustment of a pulse generator), provide a list of the eligible devices that are currently registered on the Australian Register of Therapeutic Goods (ARTG), or are subject to ARTG registration. List any devices suitable for use that are currently included on the Prostheses List. State whether the current application to MSAC is combined with a PLAC submission. The assessment of defining which medical devices are relevant may be difficult due to the rapid pace of innovation and series of incremental changes that

[c] http://www.pbs.gov.au/info/industry/useful-resources/glossary
[d] For more information on 'Black box' algorithms, see TG 15.3

devices may incorporates over time (Fuchs et al. 2017). Consider whether it is reasonable for similar devices to be in or out of scope, noting that many medical devices are approved by the TGA on the basis of evidence collected for a predicate device. However, if the current generation of a device is not substantially equivalent to an older generation of the device care will need to taken to using evidence derived from the older generation device.

Connected medical devices / software as a medical device

"Software as a medical device" (SaMD) is a class of medical software that can act as a medical device. Mobile medical applications are software applications with a therapeutic or investigative purpose and are part of SaMD (such as a medical app for cognitive behavioural therapy or to monitor heart rate or blood pressure). Therapeutic software is currently regulated in Australia as a general medical device, and where applicable as an implantable medical device (Moshi, Tooher & Merlin In submission). Medical applications are subject to pre and post market regulatory oversight by the TGA. When assessing a SaMD, indicate whether it is listed on the ARTG. Any algorithms used should also be described.

Prognostic/predictive tests

If a prognostic or predictive test uses a combination of variables to predict a clinical endpoint, provide a description of how the variables were chosen, and how the algorithm which combines these variables was developed and validated. The biological rationale and potential clinical application of the proposed test should be made clear.

In some circumstances, it may not be possible to provide the description of the algorithm development and biological rationale as the algorithm is either commercial in confidence, or has been developed by machine-learning, and considered a 'black box' (i.e. it is unclear what components are actually used in the algorithm). For a fixed algorithm, the dataset that was used for training the algorithm should be clearly described, including whether it was a convenience sample consisting of some positive and negative cases (diagnostic case control), or whether it was a cohort of patients who represent the characteristics of the target population in real-life practice (Yuste et al. 2017).

If the algorithm is not fixed (i.e. it is a self-learning algorithm), the methods of quality assurance should be described i.e. can it be ensured that the algorithm does not become biased. More description of multi-component algorithms may be found in TG 15.3 on page 99.

Genetic testing

In recent years there has been a move away from the assessment of individual genetic variants and individual genes, towards the assessment of multiple genes for broad disease areas. Describe the type of genetic testing being performed to identify the variants described under the 'Population'.

Describe the scale of gene analysis proposed, using the following MSAC endorsed classifications:

a) Monogenic testing – limited mutation testing or whole gene testing;
b) Small gene panel – assaying 2 to ≤10 genes;
c) Medium gene panel – assaying 11 to ≤200 genes;
d) Large gene panel – assaying >200 genes, but remaining sub-exome; and
e) Non-targeted – whole exome sequencing or whole genome sequencing.

If there are possible alternate testing scenarios (such as alternative timing, or alternative combination of genes/variants), describe these.

Describe the type of samples required (e.g. cheekswabs, tumour tissue, blood). Describe if there is any need for confirmatory testing if a variant is identified, and the methods used for any supplementary

testing. Classify the interpretive complexity (low/medium/high), taking into account things such as qualitative aspects (for example, level of expertise required, complexity of bioinformatics pipelines, software requirements), and quantitative aspects (for example, time component of labour required, cost of software licencing). This information should be sufficient to enable an estimate of the resources required to generate an adequate interpretation of the test results.

For any hereditary variants, describe if cascade testing of family members or testing of prospective reproductive partners is warranted.

TG 2.3 Comparator

Select the comparator(s) in the context of the Australian population with the targeted condition, the current alternative health technologies for that condition in Australia, and the technologies most likely to be replaced (or added to) in clinical practice. A single comparator will be appropriate in most circumstances. The comparator(s) should be selected based on the technology most likely to be replaced or added to in clinical practice, rather than based on the availability of evidence.

For therapeutic technologies, most comparators will involve one of the following:

- A current MBS listed therapeutic technology. If the proposed therapeutic technology is likely to replace an existing MBS listed service, the relevant comparator would be the existing therapeutic technology.
- Standard medical management (with/without placebo or sham treatment). If the proposed therapeutic technology does not replace a current therapeutic technology, or is used in addition to a current therapeutic technology, the comparator would usually be standard medical management. Standard medical management may include the use of medicines, medical services, best supportive-care or conservative management.
- Current PBS listed medicine/s. If the proposed medical service is likely to replace pharmacological management of the target population, the relevant comparator would be the current PBS listed medicine/s.

For investigative technologies, most comparators will be one of the following:

- A current MBS listed test (or multiple existing tests/test strategy).
    - If the proposed test is likely to replace an existing MBS listed test, the relevant comparator would be the existing test.
    - If the proposed test is likely to be used in addition to an existing MBS listed test, the relevant comparator would be the existing test, with *no additional testing,* and the intervention should be the proposed test plus the existing test (or plus or minus the existing test if the proposed test is a triage test).
- No testing and standard medical management. If the proposed test does not replace a current medical service/test, or is used in addition to a current medical service, the comparator would usually be standard medical management / no testing.

The expectation is that the chosen comparator is a health technology with established cost-effectiveness. Where the cost-effectiveness of the comparator is unknown, then the cost-effectiveness of the comparator as well as the intervention will need to be established.

In situations where the health technology proposed for public funding is already established practice (i.e. it has already 'diffused'), the comparator for determining the comparative benefits/harms and cost-effectiveness of the health technology should be what was used prior to the introduction of the health technology. If other healthcare changes have occurred in addition to the introduction of the proposed health technology, the comparator may be a hypothetical one, and reflect what would be

expected to occur in the absence of the proposed health technology. The comparator for the budget impact analysis (Section 4) should always be current practice, regardless of whether a historical or hypothetical comparator is used to determine the safety, effectiveness and cost-effectiveness of the health technology.

Justify the selection of the comparator. The comparator should be clearly identifiable in the clinical management algorithm. Identify factors that may affect the main comparator in the future, such as the introduction of other near-market health technologies. If there is a reasonable expectation that another health technology will enter the Australian market for the same targeted population, it may be appropriate to include it as a supplementary comparator.

If multiple comparators are identified, describe whether different comparators are used for different subpopulations of the overall target population. Include details of the subpopulations in the Population section (TG 2.1). Multiple comparators may also be required for proposed health technologies intended for more than one target population.

In circumstances where multiple comparators have been identified in the PICO Confirmation, a comparison of the proposed health technology with each comparator must be presented. In the absence of a PICO Confirmation, the usefulness of comparisons against multiple comparators for MSAC decision making may depend on:

- The evidence supporting the choice of each of the comparators
- The risk that the proposed medical service is less effective than the comparator in a subpopulation
- The size of the subpopulation as a proportion of the overall target population (if there are comparators with a small market share, they may be appropriate to mention, but not to focus on in the assessment report).

An assessment report that makes a comparison against a "basket" of comparators, where the effect of the proposed technology against individual comparators cannot be derived, should be avoided. If a "basket" of comparators is required because there is demonstrable ambiguity for the choice of the appropriate comparator (or comparators), and the evidence presented in the literature involves a comparison against the "basket" of comparators, the applicability of the "basket" to the Australian setting is crucial to present.

Outline the funding arrangements for the comparator (i.e. provide details of any MBS items, and other key healthcare resources required to deliver the comparator).

Co-dependent technologies

For co-dependent test-medicine combinations, be explicit about the comparators for the different components of the pairing. For example, if a new medicine requires a test for determining eligibility, but is compared against standard practice that does not require a test, 'no testing' will be the appropriate comparator for the test, whereas 'standard practice' would be the comparator for the medicine.

TG 2.4 Reference standard (relevant for investigative technologies only)

If a linked evidence approach is used, the proposed test strategy, compared to the accuracy of the comparative test strategy will need to be determined. If the concordance between the two test strategies is not exceedingly high, then it should be determined which is the more accurate strategy. The two test strategies should be compared against a reference standard.

The reference standard is a test or series of tests that are used to determine the presence or absence of the target condition or clinical information of interest. Ideally, the reference standard is the best available, clinically accepted, error-free procedure to do so. If there are any disagreements between the reference standard and the proposed test, then it is assumed that the proposed test is incorrect. Thus, the choice of an appropriate reference standard is a very important determinant in establishing the accuracy of a test.

The reference standard need not be a viable substitute for the proposed test. For example, if the purpose of the test is to determine the extent of a cancer in order to plan a surgical resection, the reference standard may involve the examination of the resected tissue.

If the purpose of the test is to predict a future health outcome, then the reference standard is the health outcome (i.e. for prognosis, the reference standard may be the likelihood of cancer recurrence within 5 years; or for a predictive test, the reference standard would be whether biomarker positive and negative patients respond differently to a targeted treatment versus standard care).

There will be some instances where a reference standard does not exist (or where the proposed test is considered to be the reference standard), and the accuracy of the proposed test itself will need to be demonstrated by direct from test to health outcomes evidence showing a health benefit resulting from use of the test, or by a comparison against a suitable clinical utility standard.

Clinical utility standard

A special type of reference standard is the test that was used in the generation of health outcomes evidence (direct from test to health outcomes evidence). This test (including the method of acquiring the sample, testing characteristics, and interpretation of the results) is called the clinical utility standard (i.e. it has had the clinical utility established, so is being used as a reference standard). If direct from test to health outcomes evidence is available, a comparison of the proposed test (or full range of tests expected to be used in Australia) with the clinical utility standard (and all associated testing characteristics) may be informative.

TG 2.5 Outcomes

The outcomes chosen influence the scope of evidence included in the assessment report. The assessment phase should identify whether evidence is available addressing the chosen outcome measures.

Identify the patient-relevant health outcomes for the target population, disease or condition, and decide which are the most critical outcomes to assess in order to address the clinical claim. The outcomes chosen should depend on the clinical claim being made, and (for investigative technologies), the approach chosen. The outcomes which will be most influential for MSAC are those which are patient relevant and demonstrate the clinical utility of the technology (i.e. how safe and effective the technology is compared to the comparator). For example, in cases where an effective treatment or preventative measure is available following testing, the harms and benefits of this treatment/prevention is generally the primary outcome to be used (Botkin et al. 2010).

Ideally, the outcomes chosen should be based on what is important, not on what is measured. The importance of outcomes is informed by patient input, and may be guided by patient values. Therefore, the choice of outcomes is justified by describing the impact that the outcome has upon patients (and sometimes family or carers), and the provision of evidence to support the patient relevance of the chosen outcome. Person (or patient) centred outcome measures (PCOMs) are outcomes identified as being important solely by the patient or their parent(s) and/or carer(s) (Morel & Cano 2017). For some conditions, studies have been conducted to identify a core outcome set (COS). A database of studies that have reported on COS, and more information relating to COS, can be found at the COMET

Initiative[e]. One method of then assessing these relevant outcomes is through the use of Patient Reported Outcome Measures (PROMs).

In the absence of patient input (regarding the most relevant outcome measures), sources for identifying relevant outcomes may be recent studies of the proposed health technology or the comparator, health economic models of the disease or condition and expert opinion.

In some cases, ethical considerations may affect the choice of outcomes, particularly for technologies associated with conception or pregnancy, children, minority groups or vulnerable patients (see Technical Guidance 29). Clinical utility / patient-relevant health outcomes (relevant for both therapeutic and investigative technologies) are:

- Outcomes that are directly patient-relevant that reflect improvements in quality or length of life. Surrogate or intermediate outcomes are acceptable if they have been validated as being able to predict the patient-relevant outcomes (Ciani et al. 2017). If known at the PICO Confirmation stage, provide validated examples of transformation from the surrogate to a patient-relevant outcome.
- Outcomes that relate to the direct safety of the health technology or comparator (e.g. harms from biopsy or radiation) or the indirect safety (e.g. harms caused by learning curve or insufficient training or lack of equipment maintenance, inappropriate patient selection etc).
- Outcomes that relate to the safety of any downstream investigations or interventions.
- Outcomes that relate to the effectiveness of any downstream interventions.
- Outcomes that are expected to change if the proposed medical service is publicly funded.
- Outcomes that are expected to be no different if the proposed medical service is publicly funded, and is important in the assessment of non-inferiority.

If the proposed technology is being used for multiple indications (such as diagnosis in index cases, and predisposition testing in family members), then different outcomes may be appropriate for each population. There will also be situations where people who have not received the technology themselves receive a benefit or harm due to the technology.

Family and societal impact outcomes are:

- Outcomes that relate to benefits or harms of the intervention or comparator beyond the index patient (i.e. psychological or physical benefits or harm to family members, fetus, health care professionals, health care setting, public and the environment).
- Outcomes that reflect broader consequences to the health care system including changes to healthcare resource use (hospital days, patient through-put, impacts on other services or medicines).
- Outcomes that relate to health disparities (e.g. equity of access, areas of unmet clinical need).

Given resource constraints, prioritise which outcomes are most important to include. PICO Confirmation developers should *explore* a wide range of outcomes, and then seek feedback from PASC and other stakeholders regarding which are the most important.

Clinical outcomes should be limited to those that would reasonably impact on MSAC decision making. It may not be meaningful to identify more than seven direct health outcomes, and the outcomes nominated should include the most important outcomes in terms of patient relevance.

[e] http://www.comet-initiative.org/

If a claim of non-inferiority is being made, define and justify the non-inferiority margins of key outcomes. If a claim of superiority is made, define the minimum clinically-important difference for the key outcome measure(s).

Identify any other relevant factors that will affect the implementation of the medical service, such as patient acceptability or compliance, or privacy concerns and biosecurity (relevant to software as a medical device, genetically modified organisms and biologicals). For more information on other relevant considerations, see Technical Guidance 29.

In order to follow the GRADE approach for rating the certainty of the collated evidence, outcomes should be categorised, depending on their importance for decision-making as either:

- Critical;
- Important but not critical; or
- Of limited importance.

Therapeutic technologies

There are learning curves associated with many interventional procedures (such as implanting medical devices), which may result in the safety and effectiveness of therapeutic technologies being superior when performed by experienced clinicians. It may therefore be worthwhile examining if there are moderating factors which influence the final health outcomes (such as size of the trial, or results provided in the 'key trial' versus other settings). Identification of learning curves influencing results may be useful to MSAC, so they may consider whether creating a 'centre of excellence' for particular procedures may be worthwhile.

Investigative technologies

If an investigative technology requires a linked evidence approach in order to demonstrate clinical utility, then test accuracy, change in management, and health outcome gains (the impact of the change in management) may also be required.

Test accuracy (as part of a linked evidence approach):

- Concordance of the range of tests used in Australia with the clinical utility standard (the test used in the key trial demonstrating clinical utility of the test).
- Outcomes related to the diagnostic accuracy of the test (demonstrating the accuracy of biomarker or disease detection compared to a reference standard).
- Outcomes related to the longitudinal accuracy of the test (demonstrating how accurately the test estimates the health outcome of interest)
    - Prognosis (with reference to a health outcome at a later time point)
    - Predictive (with reference to response to treatment)
    - Monitoring of disease (with reference to changes in a health state)
    - Monitoring of treatment response.

Change in management (intermediate outcomes, as part of a linked-evidence approach):

- Outcomes related to changes in diagnostic thinking and subsequent testing.
- Outcomes related to changes in preventive or therapeutic strategies.
- Outcomes related to adherence to therapeutic or preventive strategies.
- Outcomes related to referral patterns and/or the frequency and timing of follow-up.

Health outcomes (as part of a linked evidence approach, resulting from changes in management):

- Outcomes demonstrating the clinical utility of a change in management eg mortality, morbidity, quality of life.
- Outcomes assessing safety of the downstream implications of testing.

Tests may also provide personal or other utility to patients or to their family members and carers, and discussion of these outcomes could supplement an assessment of the clinical utility of the test. In cases where clinical utility is not able to be demonstrated (i.e. where no actual change in patient management occurs as a result of the test information), and an additional claim is made regarding the personal or other utility of the test or other relevant considerations, these outcomes will be key to consider. Examples of personal/other utility include:

- Ending the patient's diagnostic odyssey
- Reproductive planning
- Long-term planning (education, career, housing, finances etc)
- Increased / decreased sense of control
- Psychological (positive or negative) impact on index patient
- Stigmatisation or discrimination
- Access to National Disability Insurance Scheme
- Greater understanding of future health care needs
- The ability to connect with others in the same situation (Wurcel et al. 2019).

For more information on personal/other utility, see Technical Guidance 28.

The importance of different outcomes will differ depending on the approach taken and the type of test being assessed. For example, an adverse psychological impact, and legal and ethical implications may be more important for a screening test than a diagnostic test, as there is a much higher risk of false positive test results from screening. Additionally, testing occurs in non-symptomatic people, so any adverse effects of the testing will produce a different benefit to risk balance than if the population was symptomatic (Segal 2012). For critical outcomes, low level evidence (e.g. case series) may be useful to include in the assessment report. For some outcome measures, quantitative research may not be appropriate, and evidence in the form of patient impact summaries may be relevant to include in 'Personal utility' or 'Other relevant considerations'.

The full range of downstream effects of testing should be considered. For example, with imaging tests where there is a high likelihood of identifying unexpected findings, the test may trigger further investigations, and treatment for conditions that may never have a clinical impact on the patient. These "incidentalomas" may cause psychological distress, and have both health and cost implications (Segal 2012). This also relates to the concept of "overdiagnosis" (Carter et al. 2016; Moynihan et al. 2018).

To increase the likelihood that the assessment is future-proofed, any likely downstream consequences of the proposed test should be considered. For example, include a supplementary analysis if there is a particular targeted treatment that is not yet publicly funded in Australia but which would represent a potential benefit for patients undergoing the proposed test. Although the cost of such treatments will be unknown, inclusion of any publicly available clinical evidence for the treatment would be informative.

When describing the outcomes, it could be beneficial to consider how the outcomes relate to each part of the assessment framework (see Technical Guidance 9 on describing the approach), as well as

the target population for each of the outcomes (i.e. the patient, their family members, broader society).

Co-dependent technologies

For test-medicine co-dependent technologies, specify which outcomes are relevant for assessing each component (i.e. for the test, the relevant outcomes are likely to be test accuracy of biomarker detection and concordance between tests, whereas for the medicine, clinical utility outcomes are more relevant).

TG 2.6 Clinical management flowcharts

Prepare a flowchart that depicts current management or investigations plus management of the disease or condition in the Australian target population in the absence of the proposed health technology (i.e. the comparator). Prepare a second flowchart that depicts the eligible patients and the circumstances of use of the proposed health technology if the MBS listing or other public funding is implemented as requested. The two flowcharts may be captured on a single flowchart, if appropriate.

The flowcharts provide clarity about the target population, intended use of the proposed health technology, the replaced, added to, or displaced comparator technologies, possible changes in patient management due to the proposed medical service, and changes in resource use. They must be consistent with the population, intervention and comparator discussed earlier in this Technical Guidance subsection, and should inform the structure of the economic model (Technical Guidance 18).

Indicate whether the proposed health technology can be used at different points in the flowchart, or present multiple flowcharts for different clinical indications or uses. Justify the positioning of the proposed medical service in the clinical management flowchart.

Use the following sources to inform the flowchart(s) (in order of preference):

- a literature review of relevant published clinical management guidelines. Independently developed, up-to-date evidence-based clinical practice guidelines developed for the Australian setting are preferred. If possible, verify that the identified guidelines are currently in use and adhered to. Aspirational guidelines may not reflect current practice, and guidelines developed in an area that is rapidly evolving may no longer be relevant; or
- existing studies on the management of the condition; or
- an expert panel and/or a well-designed survey (if current guidelines or literature are not available); or
- expert clinical opinion. See Appendix 5 for further advice on expert input.

Identify the following characteristics in the flowchart(s):

- diagnostic criteria and/or prior tests to determine the target population (eligible patients), including tests required to support any proposed continuation criteria;
- important characteristics of eligible patients (such as risk factors, severity of disease and remaining treatment options);
- circumstances of use of the proposed medical service, including who is providing the service, whether special training or specialised facilities are required; provide a justification for these below the flowchart;
- treatments provided, including any required prior medical services or treatments, required co-administered therapies, and consequences for subsequent therapy options; give particular consideration to whether a proposed medical service is likely to replace a

currently available option, or whether it is likely to displace that option to a later line or therapy; and
- health care resource provision, both before and after the point in the flowchart that the proposed medical service is introduced.

Extend the clinical flowcharts to the expected end of the disease or condition process, or until the flowchart for the proposed medical service and the main comparator(s) are expected to be the same.

Comparison of the two flowcharts

Summarise the differences between the current and proposed clinical management, as depicted in the flowchart(s). Ensure that the flowcharts identify all differences in resource provision, both before and after the place(s) in the flowchart at which the proposed medical service is introduced.

Technical Guidance 3 Proposed funding arrangements

Applications should state and justify the funding arrangement being proposed (i.e. whether a new MBS listing is requested, an amendment to an existing MBS listing, funding for a package of care, or whether another form of public funding is sought, and why).

It is also important to outline the expected changes to healthcare resources if a patient receives the intervention, rather than the comparator. This is discussed in more detail in Technical Guidance 22 (healthcare resource use and costs).

TG 3.1 Proposed MBS item descriptor and MBS fee

If public funding is sought through the MBS, the medical service may be funded under an existing MBS item, an amended MBS item, a new MBS item, or it may require a combination of these options. Follow the guidance relevant to the circumstances of the proposed medical service below.

Existing MBS item

Provide the existing MBS item and descriptor that the medical service is proposed to be covered under.

Amended MBS item

Provide the existing MBS item and descriptor, and the draft changes to the MBS item (use highlighting or strikethrough to clearly present proposed changes). Report the nature of proposed change. The nature of the change may include one or more of the following:

- An amendment to the way the service is clinically delivered under the existing item(s)
- An amendment to the patient population under the existing item(s)
- An amendment to the schedule fee of the existing item(s)
- An amendment to the time and complexity of an existing item(s)
- Access to an existing item(s) by a different health practitioner group
- Minor amendments to the item descriptor that does not affect how the service is delivered
- An amendment to an existing specific single consultation item
- An amendment to an existing global consultation item(s)
- Other (describe)

New MBS item

Draft an MBS item descriptor that defines the target population and the medical service that would define eligibility for MBS funding. The objective of the MBS item descriptor is to contain criteria that would reasonably ensure that the performance and the costs of the proposed medical service are consistent with the conclusions in the assessment report. A broad MBS item descriptor may permit access to a population for whom the performance of the medical service is limited or unknown, and may result in a higher cost to the Government.

Discuss the relevance of the included criteria in the proposed MBS item descriptor (such as the health practitioner group that may access the item, the population and the intervention). State whether other items are expected to be used in conjunction with the item (such as anaesthetic items). Identify possible eligibility criteria that have been omitted from the MBS item.

Specify who is able to request the service, and how often the service may be claimed (e.g. times per year, or if there is a once per lifetime limit). Specify where the service will be provided (i.e. will it be as an in-hospital service on an admitted patient, or an in-hospital service on an admitted or non-

admitted patient, or out-of-hospital / outpatient service etc). Where relevant, state the applicable type of procedure: Type A, B or C.

- Type A procedures are overnight procedures;
- Type B procedures are same day hospital procedures;
- Type C procedures are often out-of-hospital procedures which do not normally require admission.

Specify any exclusion criteria (such as items which may not be co-claimed), and any exemptions.

MSAC has a preference for 'technology agnostic' language for item descriptors, allowing for a variety of different health technologies (particularly tests) to use the same item, without the need to amend the MBS item descriptor. Use of trade names should therefore be restricted to cases where it can be demonstrated that generic language may result in dissimilar health technologies being used and claimed under the proposed MBS item.

Multiple MBS items may be required to plan and administer a technology, or to use over the life of a technology; i.e. for implantation and removal of devices. Draft as many items as required.

If the costs associated with using the proposed technology in different populations will differ, separate MBS items will be required to allow for different fees.

If cascade testing of biological relatives or testing of reproductive partners is of relevance, draft additional MBS items to cover these indications.

Proposed MBS fee

Explain how the proposed MBS fee has been derived. In circumstances where the medical service is proposed to be covered by an existing or amended MBS item, explain why its fee is appropriate.

A comparison of MBS fees of similar services may provide some context for the proposed fee. However, greater justification for a fee will usually be required. Typically, the service should be costed with reference to relevant input costs. These may include the costs of the time taken for the provider to perform the service (before, during and after the service) and direct service costs, such as costs of the time taken for identified 'non-rebated employees' to be involved in the provision of the service, costs of identified consumables, and costs of shared use of identified reusable equipment.

If the proposed medical service is already in use in Australia, provide a summary of fees currently charged for this service.

Include any information about any Extended Medicare Safety Net (EMSN) risk, where an out-of-hospital benefit cap may need to be applied. This usually applies to out-of-hospital services where practitioners are charging large amounts. Where MBS fees are higher than the Greatest Permissible Gap (GPG) threshold, include the likely GPG and EMSN rebates (for out-of-hospital services), if applicable.

Table 3 Proposed MBS item descriptor

Category	[proposed category number] – [proposed category description]
Group, Subgroup	[proposed group number] – [proposed subgroup]
Proposed item descriptor	[provide proposed item descriptor]
Proposed Fee	[provide the proposed fee]

Genetic tests

MSAC endorsed gradations for panel sizes listed in the 'Intervention' section correspond to different fee levels, so benchmarking against MBS items which fall into these categories would be appropriate.

Genetic counselling may be recommended for patients or family members undergoing genetic testing and will be relevant to discuss in other relevant factors (Technical Guidance 29) and incorporate into the economic modelling and financial impact, however it is not included as a criterion on the MBS item descriptor.

TG 3.2 Alternative arrangement for funding

If public funding is not sought through the MBS, please provide a description of the proposed funding arrangement. Explain the process of identifying the target population, and methods for restricting use to the intended population.

Explain the proposed fee or the amount to be charged.

Technical Guidance 4
History of MSAC submissions for the health technology

Tabulate the dates of previous committee considerations for the proposed health technology and indication in Table 4.

Table 4 MSAC submission history

Committee	Meeting date(s)
PASC	[add]
ESC	[add]
MSAC	

For re-considerations, present a table with a summary of the issues raised by the MSAC (cross-reference to the MSAC Public Summary Document, where possible), and show how the current assessment report addresses the issues, with cross-referencing to the relevant sections of the current assessment report.

The table summarises key matters from the previous MSAC considerations, and how the current assessment report addressed those concerns. Highlight key points from the MSAC Public Summary Document, citing the relevant paragraph / page. Identify any important matters of concern raised by ESC that were not over-ruled by MSAC. Issues raised by ESC or PASC that are not specifically mentioned by MSAC remain outstanding, and may become more important or have a larger impact in the current assessment report.

Table 4a Summary of key matters of concern

Component	Matter of concern	How the current assessment report addresses it
[Identify the relevant section of the previous assessment report, eg comparator, clinical claim, economic evaluation etc.]	[Cite paragraph of the MSAC PSD (use abbreviated referencing in tables), identify matter]	[{*Addressed/Not adequately addressed/not addressed}* Comment and/or cross-reference to where addressed below in the executive summary or main body.]
Example text is provided below		
Clinical place in therapy	Example text: MSAC suggested the descriptor should reinforce that psychotherapy must have been previously trialled (PSD, p.2)	Example text: *Addressed. Restriction amended to reflect MSAC comments.*
Clinical effectiveness	Example text: MSAC noted there was other available evidence which could be informative on the relative effectiveness that was not presented in the resubmission, including the EUnetHTA 2017 and Ontario Health 2016 Reports (PSD, p.3)	Example text: *Addressed.* *The efficacy results from EUnetHTA 2017 are now applied in the economic modelling, as this is the more recent of the two reviews requested to be reviewed by MSAC.*

Source: {Table/Figure}, {p[]/pp[] of the assessment report}.

Technical Guidance 5 Methods of assessment

Part of the process of going to PASC should be to establish whether all components of the proposed listing should be evaluated with a full health technology assessment, or whether the exemplar/ facilitated approach may be incorporated.

TG 5.1 Full health technology assessment

The majority of health technologies assessed by MSAC will require a full health technology assessment (HTA) to be performed (including a systematic review of the safety, effectiveness and determination of the cost-effectiveness and budget impact of the technology).

For methodology regarding literature searching, assessing the risk of bias, GRADE, and extracting the key characteristics of studies, see Appendix 2 to Appendix 5. The methodology for assessing therapeutic technologies is well established, and therefore quite stable. However, the methodology for assessing investigative technologies is less established, and new guidance has been provided on examples of risk of bias tools for assessing tests with a range of different uses, and how to adapt the GRADE approach for different linked-evidence components.

TG 5.2 Exemplar / facilitated approach

The exemplar/facilitated approach has been introduced as a means to simplify and streamline the assessment of related technologies, and take a pragmatic approach to allow MSAC to make decisions on a broader number of potential listings at once (or inclusion of broader range of genes on panels included in the single listing etc). The exemplar / facilitated approach is intended to be used almost exclusively for investigative technologies.

The purpose is to find strong areas of commonality across genes or indications, or versions of devices etc such that the strength of evidence required is commensurate with the risk to patients or the financial impact of the technology to the healthcare budget. Examples of where this approach has currently been trialled, include for panel testing of a broad range of genes for a single disease area, assessment of staging for PET-CT across tumour types, and for assessment of tumour-agnostic test-medicine co-dependent technologies (Table 5).

Table 5 Summary of the exemplar/facilitated approaches used

Type of facilitated approach	Exemplar	Facilitated
Additional test parameters (ie a gene panel rather than single gene tests) Different intervention same population	One or several genes on a panel that have evidence to support the clinical utility and cost-effectiveness of testing.	Additional genes that may be included in the same panel that would be used in the same population*, but that do not have strong evidence due to the rarity of the gene variant.
Additional indications (ie, staging or imaging for multiple tumour types) Different population but same modality (technology)	One or several tumours that have the evidence to support the clinical utility and cost-effectiveness of staging / imaging.	Additional tumours that may be detected using the same imaging, but that do not have strong evidence due to the rarity of the disease.
Alternative technologies that would be used in clinical practice Substantially equivalent devices (same population, non-inferior technology)	One or several technologies (a clinical utility standard, or a medical device) that have evidence to support the effectiveness, safety and cost-effectiveness of the technology.	An alternative technology (different assay, different device used for the same indication) that does not have evidence of effectiveness but has evidence of non-inferiority on a surrogate (test accuracy, substantial equivalence).

*Increases in population would require additional considerations (see Figure 3).

The underlying principles common across the exemplar/facilitated approaches for investigative technologies are:

a) The exemplar listing has had its comparative safety, effectiveness, cost-effectiveness and budget impact established (normally in quantitative terms through the use of a full HTA).
b) The request for a facilitated listing is for a smaller population than the exemplar listing.
c) The facilitated listing would not differ in purpose from the exemplar listing (i.e. if the clinical utility of the exemplar is demonstrated through its ability to help determine a patient's eligibility for a treatment, the facilitated listing should not be for another purpose such as staging or monitoring; or if the exemplar listing is for germline testing, the facilitated listing should not be somatic testing).
d) The frequency of use of the health technology should not be greater for the facilitated listing than for the exemplar listing(s).
e) The requested fee or unit cost of the facilitated listing *per patient* would be no more than the exemplar listing per patient.
f) There should be a clear (but non-quantified) basis for MSAC to conclude that the overall cost per patient (including downstream management changes) will not be greater for the facilitated listing than the exemplar listing.
g) There should be a clear (but not necessarily quantified) basis for MSAC to conclude that health outcomes will not be inferior on an average per eligible patient basis (including both comparative safety and comparative effectiveness) for the facilitated listing compared to the exemplar listing.
h) The facilitated listing should not encourage changes in clinical management that are not already available in clinical practice, i.e. through
    a. enrolling the patient into a research environment (e.g. a clinical study)[f]; or
    b. starting the patient on a therapy which is beyond its existing subsidy arrangements (e.g an unregistered medicine, or a medicine beyond its PBS restrictions).

Health technologies which use a device

MBS items which refer to use or implantation of devices are non-specific to which individual device should be used (the language is generic to a class of device). Applications for an MBS item involving use of a device are therefore prime candidates to use the exemplar/facilitated approach. If evidence is identified to establish the incremental safety and effectiveness of at least one of the devices (compared against the comparator of what would be done in the absence of the proposed devices), this becomes the exemplar. The additional devices in the same class may therefore be assessed using a facilitated approach. In order to establish that the facilitated devices are non-inferior to the exemplar device, the two (sets of) devices should be tabulated to compare and contrast the following:

- Intended purpose of the devices (in regard to indications and contraindications)
- Technical equivalence (design, specifications, physiochemical properties, energy intensity, deployment method, principles of operation)
- Biological characteristic (biocompatibility of the materials in contact with the body).

This process is similar to methods used by the TGA in establishing that new devices are substantially equivalent to a 'predicate' or 'similar marketed device'. The Global Harmonization Task Force definition of substantial equivalence is "the devices should have the same intended use and will need to be compared with respect to their technical and biological characteristics. These characteristics

[f] The *Health Insurance Act 1973* does not allow research to be considered an eligible service

should be similar to such an extent that there would be no clinically significant difference in the performance and safety of the device".[g]

Note, the exemplar/facilitated approach for devices is not appropriate if the devices are considered to be high risk or highly novel technologies.

Genetic tests

Genetic testing has developed such that broad panels of genes (see TG 2.2 for classification) may be tested for simultaneously for relatively small incremental cost over testing for particular genes. The method for assessing genetic testing therefore needs to be suitable for testing panels of different genes at once.

In the pre-assessment phase (development of the application form and the PICO Confirmation), it should be determined which genes are most prevalent and have the most evidence of clinical utility for each disease area. A full assessment quantifying the clinical utility and cost-effectiveness of testing the exemplar gene(s) is required.

The concept of the exemplar/facilitated approach is that if the cost-effectiveness of testing for the exemplar gene(s) can be justified, then testing of additional genes (with less evidence), need not be assessed through a full HTA approach. Instead a streamlined approach may be used (see TG 5.2 regarding the streamlined approach).

During the development of the PICO, the gene(s) considered exemplars should be defined, so that it is explicit which the minimum number of genes which are required to be considered through a full HTA approach.

If adding the facilitated genes does not increase the cost of the panel, and no expansion of indications are expected, then the clinical utility of the facilitated genes may be described qualitatively (i.e. the size of the benefit need not be quantified), as the cost-effectiveness of the panel may be justified on the exemplar genes alone. If the cost of panel increases with additional genes, the facilitated approach would not be appropriate, as per the underlying principles.

Likewise, even if the cost of the panel remains the same, if patients with the additional genes are able to access higher cost treatment (which is assumed to be acceptably cost-effective), there will be an incremental cost associated with the downstream implications of testing. This incremental cost would need to be justified by a quantified clinical benefit, and would therefore need a full HTA rather than the streamlined approach.

If there are additional costs of testing due to an increase in the size of the population being tested, then the financial implications of adding the additional genes will need to be assessed. The additional costs should be incorporated into the exemplar model, but the benefits of testing the additional genes or patients need not be quantified unless the cost-effectiveness is substantially reduced. If there is a clear rationale why the benefits accrued by testing the additional genes would be similar to the exemplar genes, then the clinical utility of the facilitated genes may be assumed to be the same as the exemplar genes. If they cannot, then the decision regarding expansion of the panel (to include the facilitated genes) is likely to be deferred by MSAC, until clinical utility of the facilitated genes can be quantified.

[g] Powerpoint presentation by Simon Singer (2017), Clinical evidence Guidelines, Beyond the CSR – Demystifying Clinical Evidence Requirements for Medical Devices, TGA, Department of Health available from https://www.tga.gov.au/presentation-clinical-evidence-guidelines-0

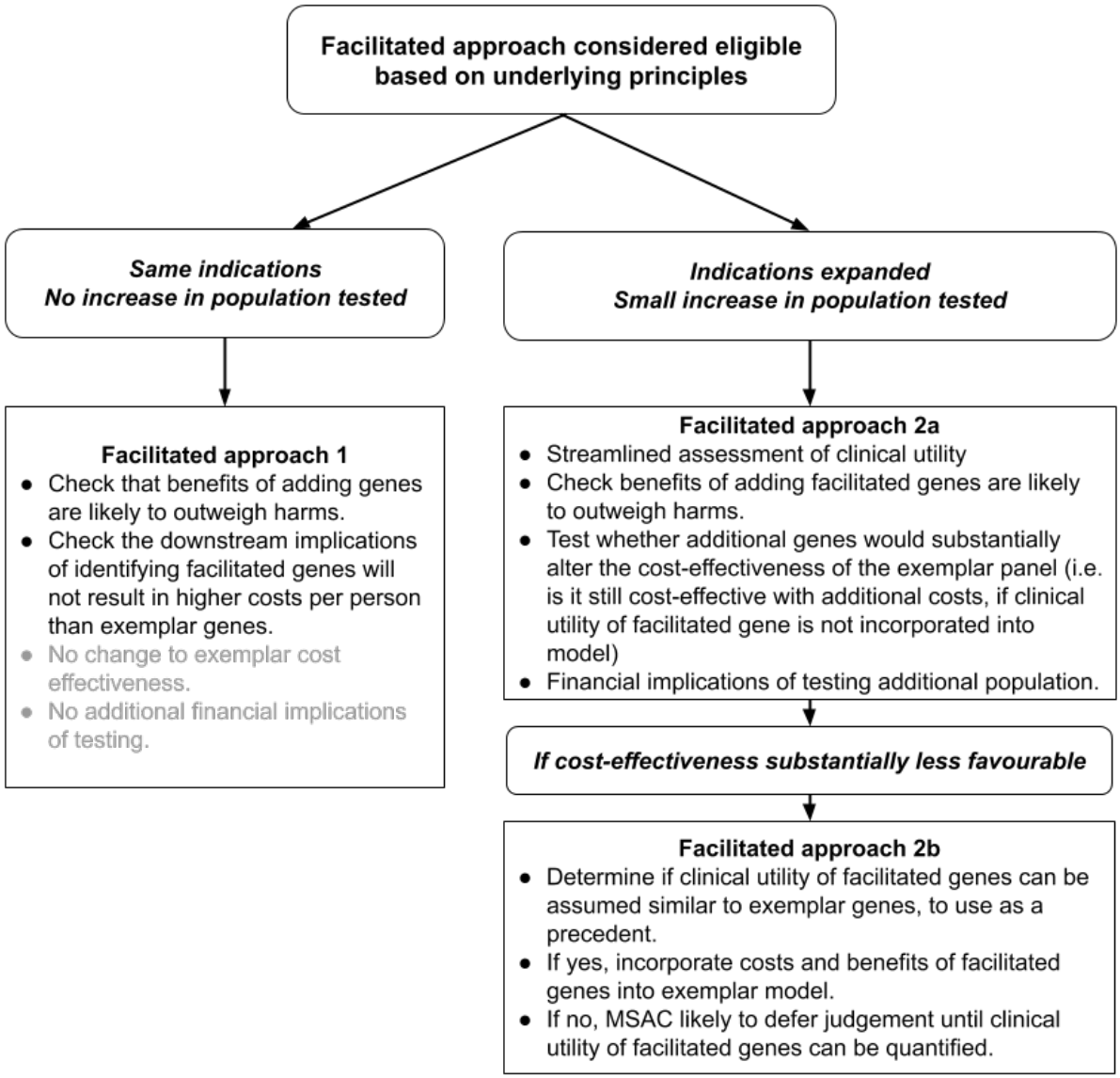


Figure 3 Example of different approaches for facilitated listings, based on whether population is identical to exemplar listing or slightly different

Imaging

MSAC has tested the feasibility of developing a 'frame of reference' model for positron emission tomography (PET) to be used for 'staging' FDG-avid tumours in different parts of the body. The concept is to move away from condition-based assessment, and instead to address the clinical utility of PET in relation to all FDG-avid tumours, regardless of the origin or site of the cancer.

The method developed, uses the exemplar/facilitated approach. The FDG-avid tumour types with the strongest case for clinical utility and cost-effectiveness (i.e. the more common cancer types) were selected to be exemplars, to be used as the basis on which to build assumptions for the inputs of the model. For other cancer types which are rarer, a pragmatic approach has been taken to base the assumption on the exemplar population. The expectation is that the exemplar/facilitated approach will be able to generalised to other indications (e.g. diagnosis or monitoring), and other forms of imaging.

The primary criterion to proceed to facilitated assessment is cancer incidence of equal to or less than 12 cases/100,000 population/year[h]. If the cancer is not uncommon, then a full MSAC assessment is required, as per exemplar cancers. There also needs to be some indicative evidence that the proposed imaging changes management and health outcomes. If no treatments are available for the cancer identified, then the application is classified as unsupported. The logic map developed is shown below.

Table 6 Logic map for the evidence expectations to facilitate the appraisal of PET services

Purpose	Triage level	Test accuracy	Change in management	Clinical utility	Economic	Financial
Staginga	Exemplar	Quantified increment	Quantified change in management	Quantified increment	Quantified increment	Quantified increment
	Facilitated	Qualitative (similar accuracy)	Strong clinical plausibility	Qualitative (at least same direction)	Qualitative (at least same direction)	Quantified increment
	Unsupportable	Not able to meet one or more categories of facilitated level of expected evidence				

a Similar mapping process to be applied to diagnosis, treatment response and recurrence

If an exemplar/facilitated approach is proposed to be used for other types of imaging or for other indications than PET for staging of cancer, the applicant or assessment group should consult with the Department of Health for the outcomes of the pilot process, and latest methods recommended. Integrated co-dependent submissions for tumour agnostic or pan-tumour cancer medicines

Medicines that have pan-tumour capability are able to affect tumours originating from any part of the body, through a common mechanism of action. This means that the medicine is "tissue/site agnostic"[i]. Many of the medicines on the horizon which are pan-tumour require a test to determine eligibility, so will require a co-dependent assessment. Similar to the exemplar/facilitated approach for genes, the concept would be that the evidence for exemplar tumour types would be evaluated, plus the biological rationale for how the facilitated tumours would react similarly to the exemplar tumours. From the point of the view of the MSAC submission, the key issue is considering how the positive predictive value of the test differs for each different tumour, given the differing prevalence rates of the biomarker between different tumours. For more information, see the discussion paper on pan-tumour biomarker testing to determine eligibility for targeted treatment on the MSAC website[j].

Approach to assessment of facilitated listings

A PICO Confirmation discussing both the exemplar and facilitated listing would normally be expected, outlining the assessment questions to be addressed.

The streamlined approach for facilitated listings is to use targeted (non-systematic) searches to identify relevant evidence. Scoping searches should be performed to see whether existing HTAs, systematic reviews or evidence-based clinical practice guidelines relevant to the Australian healthcare system are available. If no secondary research is identified, then primary research may be used.

h "rare cancers" have an incidence of <6 per 100,000 per year; and "uncommon cancers" have an incidence of 6-12 per 100,000 per year.

i https://www.fda.gov/Drugs/InformationOnDrugs/ApprovedDrugs/ucm560040.htm accessed 7th February 2020

j http://www.msac.gov.au/internet/msac/publishing.nsf/Content/assessment-groups accessed 10th February 2020

In order to establish that the underlying principles of using the facilitated approach are met, the size of the population expected to the use technology needs to be identified (and must be smaller than the size of the population expected to use the exemplar technology).

The requested fee or unit cost of the facilitated listing per patient must be reported (including downstream implications) and must be no more per patient than the exemplar technology. The budgetary impact of the facilitated listing must be reported.

The amount of benefit required to be established for the facilitated listing depends on whether it is being used in the same patients as the exemplar listing (i.e. is it simply adding genes to a panel with exemplar genes) or whether it is for a different population.

Possible facilitated approach 1:

If the facilitated listing does not result in any expansion of the population (and has no additional downstream costs), then the size of health benefit does not need to be quantified, as long as there is logic that the benefits are likely to outweigh any harms. In this situation, the cost effectiveness of the exemplar listing is not likely to differ with the addition of the facilitated listing, and the financial implications won't differ. For example, if *BRCA1/2* testing is considered cost-effective in patients with breast or ovarian cancer, adding additional variants in *STK11*, *PTEB*, *CDH1*, *PALB2* and *TP53* genes to the panel can be done without quantifying benefit for these additional genes relevant to the same population.

Possible facilitated approach 2a:

If the population differs between exemplar and facilitated listings, the additional financial impact requires that the cost-effectiveness should be considered.

One of the underlying principles for the exemplar/facilitated approach is that the exemplar listing is cost-effective. If the costs associated with the facilitated listing can be incorporated into the exemplar model (without any benefits attached), and the listing remains cost-effective, then the size of the expected health benefit of the facilitated approach need not be quantified.

Possible facilitated approach 2b:

If incorporating the costs of the facilitated listing into the exemplar listing substantially reduces the cost effectiveness of the exemplar listing, then the benefits of the facilitated listing need to be estimated.

If it is logical (based on evidence, expert clinical opinion or biological plausibility) to estimate that the benefits for the facilitated listing are non-inferior to the exemplar listing, then benefits of the exemplar can be assumed to apply to the facilitated listing. Costs and benefits can therefore be incorporated into an economic model.

Unsupportable scenario

When any of the underlying principles are not met for the facilitated approach, the listing cannot be supported without a full HTA being performed.

Section 2 Clinical evaluation

In Section 2, an evaluation of the clinical evidence for the proposed health technology compared with the main comparator in the context of the requested listing is presented. The derivation of clinical evidence involves a literature search, critical appraisal and synthesis of the best available evidence.

Guidance for the clinical assessment of a health technology is separated into guidance for therapeutic technologies **(Section 2A)** and for investigative technologies **(Section 2B)**. Methodology guidance that is common across both therapeutic and investigative technologies are presented in Appendix 2, Appendix 3, Appendix 4 and Appendix 5.

Section 2A
Assessment of therapeutic technologies

The assessment of a therapeutic technology is well established. The approach includes:

- A systematic literature search for relevant evidence (Appendix 2)
- An assessment of the risk of bias of the included evidence (Appendix 3)
- The presentation of characteristics of the included studies (Appendix 5), and
- An overall assessment of the certainty of the evidence for each outcome (Appendix 4)

The TG subsections presented in Section 2A describe the assessment and presentation of the effectiveness and safety of a therapeutic technology.

In almost all cases, the assessment of a therapeutic technology will require evidence of the health outcomes and safety of that particular technology. Rarely, an assessment may pursue a facilitated approach. An example of a facilitated approach for a therapeutic would include the assessment of other (facilitated) devices that may use the same MBS item as the device for which there is evidence of effectiveness and safety (the exemplar device). Technical Guidance 5 discusses the exemplar/facilitated approach.

Technical Guidance 6
Effectiveness of a therapeutic technology

Therapeutic technologies: A type of technology that is expected to, or claimed to be able to, directly improve the health of people receiving it. Nothing else needs to be rendered to achieve the improvement in health outcomes. Examples of therapeutic technologies are devices, medicines, vaccines, procedures, programs or systems.

The objective of presenting the systematic overview of the results of therapeutic technologies in the assessment report is to efficiently present the most relevant study results and synthesis for decision making. The assessment report should contain key results and a discussion of the results as they relate to the clinical questions identified in the PICO process. Extensive tables are placed in the technical report and referenced in the main body.

The structure used to present the study results can be adjusted to address the quantity and the nature of the evidence that needs to be presented. Typically, subheadings representing each outcome identified in the PICO process in descending order of patient relevance is an appropriate method for organising the results. An example of the structure that may be used as a guide is presented in Appendix 1.

The results of the included studies that are presented in an assessment report includes the following, presented separately by outcome:

- results from individual studies (eg trials, studies or metaanalyses identified in the literature search)
- metaanalysis of results (if appropriate)
- indirect comparisons (if appropriate)
- discussion of the quality and certainty of the included evidence
- summary of supplementary evidence (if appropriate)
- discussion of the overall evidence base in the context of the risk of bias (Appendix 3), quality (Appendix 4), confounding (Appendix 5) and applicability to the proposed target population (Technical Guidance 2 and Technical Guidance 3).

Presenting all information relating to one outcome prior to addressing results for a subsequent outcome will improve readability.

TG 6.1 Presenting results of individual studies

It is important to ensure that the results from individual studies are reported in addition to pooled estimates. Relevant details of individual study results include:

- the number of patients at risk or providing data to the results
- the number of patients experiencing the event (if appropriate)
- the percentage of patients with the event, and means (standard deviation) or medians (interquartile range) within groups as appropriate
- confidence intervals (CIs) of the outcomes within groups
- relative and absolute differences between groups, and CIs

The format of tables for the presentation of results will need to be adapted to the data available and the type of outcome. The assessment report contains example tables for presenting different types of outcomes.

The following information may be relevant to present along with the individual study results, to aid with the interpretation of tabulated data:

- The timing of the outcome assessment (eg EORTC-QLQ C30, change from baseline at six weeks).
- Whether studies measured the same outcome at more than one time point (eg 6 weeks, 12 weeks, 3 months). If only one of these time points is presented in the assessment report, justify the choice. Discuss whether the treatment effect differs across other time points. Present the results for other time points in the technical appendix, or clearly reference the source.
- Discuss the appropriateness of thresholds used to translate continuous outcomes into dichotomous or categorical outcomes.
- State the statistical method used for any analysis in a footnote to the table. Report any covariates used in the statistical analysis. If a statistical method adjusts for covariates, present the results of an unadjusted analysis in the table footnote. If required, discuss the appropriateness of the statistical analysis, or the impact of different studies applying different methods.
- State whether any assumptions are required to support the statistical method (eg assumption of proportional hazards for a Cox proportional hazards model), and whether the assumptions have been tested. State whether there is a risk that the assumptions supporting a statistical approach are invalid.

Dichotomous data

Dichotomous data are presented as numbers with the event in each arm as a proportion of the total numbers of subjects in the arm (ie n/N). For comparative studies, this permits the generation of a relative risk and a risk difference. If studies present a relative or absolute treatment effect adjusted for covariates, this is the appropriate treatment effect to present.

Continuous data

Many studies measure a continuous variable at baseline and again at a prespecified time point. The treatment effect from such studies can be reported in several ways including mean difference, weighted mean difference and analysis of covariance. Consider the appropriateness of pooling continuous data across studies if different covariates have been applied, or time points differ.

Time-to-event data

Time to event data can be presented by studies in several formats. As well as presenting the individual study estimates for a time to event outcome, it is important to describe the method for analysing the time-to-event data. Many methods rely upon underlying assumptions, which are necessary to present alongside results to aid in interpretation.

Ordinal or categorical data

Attempt a similar approach as the method described for continuous data if the trial results are available as ordinal or categorical data (eg a Likert scale for patient-reported outcome measures). Expert biostatistical advice will be helpful in such circumstances, particularly if meta-analysis is applied.

Multiattribute utility instrument data

MAUI results are commonly reported across several time points. Report detailed MAUI results in the technical appendix. When reporting MAUI results, provide them for each time point and each arm within the study. The number of patients eligible to respond, and those who actually responded, is

important in the interpretation of the results. It should be made clear if compliance rates are not reported as this undermines the confidence in the results.

In the main body of the report, provide the difference between the arms (with 95% CI) as the integrals between the mean utility weights obtained over time up to the median (or other relevant time point) follow-up in the study. If an alternative approach for comparing MAUIs was used, explain how this was done.

State which scoring algorithm has been used to map the MAUI to utilities. Discuss the applicability to the Australian setting.

When providing an interpretation of the MAUI results, discuss the consistency or inconsistency with any concomitantly used patient reported outcome measure in the same study.

Subgroup analysis

If only some of the participants from the whole study population are relevant to the target population, present a subgroup analysis to show the treatment effect of the proposed therapeutic technology in the relevant population.

Ensure that the participant characteristics and treatment details have been extracted (as per Appendix 5) for the whole study population and each of the relevant subgroups.

Provide the following information to support a subgroup analysis:

- Clarify why the proposed medical service should not be available to the patients in the complement of the subgroup and why the study enrolled a broader population
- The plausibility of a variation in treatment effect for the subgroup, as it relates to the biological or clinical rationale for using the therapeutic technology. An unexplained variation is difficult to interpret in the absence of such plausibility (cross-reference to any discussion of biological plausibility that has been provided in the context section of the assessment report).
- Whether the subgroup analysis was prespecified and whether randomisation was stratified by the subgroup. If the subgroup is defined using a threshold (such as a level of marker in the blood, or a severity score), justify the choice of the threshold. Discuss the impact of varying the threshold on the outcomes.
- The number of subgroup analyses originally conducted and any statistical adjustment for multiple comparisons.

Present the analysis of the treatment effect for the subgroup and compare this with the complement of the subgroup. Tabulating results side-by-side improves readability. Test for interaction between the subgroup and its complement to support and quantify the association between the treatment effect and the covariate defining the subgroup. In assessment reports that rely upon a meta-analysis, individual study subgroup analyses may be presented in the technical appendix (and referenced), however the information required to support a subgroup analysis should be provided in the main body.

If a subgroup must be extracted for only some of the included studies, present a subgroup analysis prior to performing a meta-analysis. Use a random effects meta-analysis for pooling data, if feasible.

TG 6.2 Synthesis of the results

The synthesis of results for an MSAC assessment report requires the consideration of the available evidence.

Meta-analysis

If appropriate, present a meta-analysis of the aggregated data. If a meta-analysis is performed, state the software used, and describe the methods. Select a method that is appropriate for the format of the outcome data (eg dichotomous, continuous, time-to-event). A DerSimonian-Laird random effects model is preferred - justify an alternative approach. Document and reference the methods used so that they are reproducible and verifiable.

Present a forest plot that includes the estimate of the individual study treatment effects, and the pooled treatment effect. In a table note, report whether any studies that reported results for the relevant outcome were excluded from the meta-analysis and explain why any studies were excluded.

Discuss the methods used for pooling time-to-event data, or outcome measures that are derived using statistical approaches that control for covariates.

Report results for statistical heterogeneity (Cochran *Q* with a chi-square test for heterogeneity and the I^2 statistic). Discuss any heterogeneity identified in the meta-analysis with reference to the study characteristics (Appendix 5), outcome definitions (Appendix 5) and study design (Appendix 3). Use an appropriate method to test for the risk of publication bias and comment on the findings.

If multiple studies report on the same or similar outcome, but it is inappropriate to perform a meta-analysis, explain why. Describe the results narratively and nominate the most relevant studies on the basis of study quality and applicability.

The discussion of the synthesis should contain the following elements:

- A statement of the direction of the treatment effect
- An estimate of the magnitude of the treatment effect
- The consistency of the treatment effect across studies
- A discussion of the strength (certainty) of the evidence base

The discussion for each outcome should reference the concerns identified relating to the search, risk of bias and study characteristics. In addition, discuss the applicability of the study populations or their characteristics to the target Australian population.

Magnitude and direction of treatment effect

Discuss the estimate of the magnitude of the treatment effect (or pooled treatment effect, if appropriate) in the context of the clinical relevance of the magnitude (MCID). If relevant, discuss the comparative estimate of effectiveness in the context of a nominated non-inferiority margin.

If available, comment on the consistency of the treatment effect across key subpopulations (eg by patient or disease characteristics).

Strength of the evidence

Use a GRADE approach to present an overall assessment of the quality of the evidence for each outcome. GRADE requires an assessment of the following domains to rate the *quality* or *certainty* of the body of evidence.

- Risk of bias or study limitations
- Imprecision
- Inconsistency
- Indirectness
- Publication bias

Following the presentation of individual study results and meta-analysis (if appropriate), present a discussion of the imprecision, inconsistency, indirectness and risk of publication bias across the evidence base (per outcome). The use of GRADE tables is preferred. For each outcome, discuss the overall strength of the evidence base. Do not present extensive GRADE tables in the assessment report.

TG 6.3 Other approaches

Indirect comparison

An indirect comparison may be appropriate if no direct randomised controlled trials are identified in the systematic literature search. Relevant studies are identified following the guidance provided in Appendix 2.

Describe the method(s) used for the indirect comparison, such as the Bucher single pairwise method,[24] matching-adjusted indirect comparison,[25] simulated treatment comparison,[26] network meta-analysis or mixed treatment comparison. Where there is more than one randomised trial with the same intervention and common reference, separately pool the treatment effect prior to performing the indirect comparison.

Where there are multiple common comparators in the network, perform pairwise comparisons for each possible pathway in the network. The Bucher method[24] is widely used; it describes how to indirectly compare the odds ratios from randomised trials that share a common reference arm. This method has been extended to include other treatment effect measures, such as relative risk, absolute risk and hazard ratio.[27]

More complex methods, such as network meta-analyses, may be presented as supplementary analyses. For network meta-analyses, present the results of pairwise comparisons for each link in the network. Although some methods consider nonrandomised studies in a network, avoid including nonrandomised studies. Where nonrandomised studies must be included, present the results of the network meta-analysis both with and without the nonrandomised studies.

Unadjusted indirect comparisons (such as a naive comparison between single arms), or indirect comparisons where differences in trial characteristics may affect the transitivity of the trials in the comparison, are difficult to interpret and reduce the confidence of MSAC in decision making. Where patient-level data are available for at least one study in the comparison, use matching-adjusted indirect comparisons or simulated treatment comparisons to correct for trial differences to improve the transitivity of the comparison.

When considering complex approaches (eg matching-adjusted indirect comparisons, simulated treatment comparisons, network meta-analyses, mixed treatment comparisons), balance the additional information requests and challenges these approaches may present with any reduction in uncertainty they may deliver. In the technical report, provide sufficient detail to repeat the analysis, including programming code for statistical software such as Stata, R, SAS or WinBUGS. For methods that require individual patient data (matching-adjusted indirect comparison or simulated treatment comparison), attach the individual patient dataset in a spreadsheet. Justify where this is not possible.

When presenting the results of an indirect comparison, include the following:

- The number of studies included in the indirect comparison, and whether any studies identified in the systematic literature search have been excluded (and why)
- An assessment of the balance of potential confounders across arms in individual trials
- An assessment of the heterogeneity of any meta-analysis

- A comparison of the event rates across the common reference arms of pairwise comparisons. Discuss the implications of differences in the event rates. If event rates indicate a difference in baseline risk across trials, discuss whether the relative treatment effects are consistent across baseline risk
- Nominate and justify the choice of outcome measure (eg odds ratio, relative risk, absolute risk difference) – the choice of outcome measure should minimise the variation in the comparative treatment effect within each and all sets of included randomised trials.
- If a relative outcome measure is nominated and the desired outcome is an absolute risk difference, convert the indirect estimate of relative treatment effects to an absolute risk difference
- Present the indirect estimate of effect as relative risk and/or odds ratio (or the ratio of hazard ratios) with its 95% CI (or if previously justified, the absolute risk difference)
- If trials have been excluded, include sensitivity analyses in which these trials are included (if possible).

Justify the use of methods more complex than simple pairwise comparisons (Bucher method). In the technical appendix, provide an explanation of the method, the statistical code and the assumptions required for each approach (and how the assumptions were validated). If individual patient data are required by the statistical approach, provide these data, or justify their omission. In the main body, compare the results with those derived from a simple indirect comparison method, and explain any difference.

Adjustment for treatment switching

Adjustments to correct for the influence of treatment switching on the treatment effect may rely upon assumptions that are difficult to validate. Evidence without treatment switching is preferred.

In circumstances where participants in the control arm of the included study 'switch' and receive the proposed therapeutic technology, it may be reasonable to statistically adjust the treatment effect to remove the effect of the proposed therapeutic technology on subsequent endpoints. If switching to the proposed therapeutic technology reflects clinical practice, then the appropriate comparator would include subsequent treatment with the proposed therapeutic technology, and no adjustment is necessary.

If an adjustment for treatment switching is necessary, describe the mechanism for switching. If switching (or the extent of switching) does not reflect clinical practice, provide the following:

- Baseline characteristics of switchers vs nonswitchers (and discuss differences)
- Reasons for switching

Several methods are available for adjusting survival estimate for treatment switching (Latimer et al. 2014). Using simple methods may be acceptable when the adjusted estimate of the comparative treatment effect is clearly toward the null. If complex methods are used, provide details on the approach, assumptions (and how they have been tested), and a comparison across more than one method.

Provide a discussion and interpretation of the results.

Where there is a largely uncontaminated estimate of an outcome that occurred before switching, discuss whether the outcome is a valid surrogate, and translate the surrogate to the final outcome.

Combining an adjustment for treatment switching with the use of subgroups or indirect comparisons will result in a high degree of uncertainty and should be avoided. If this is necessary, ensure that the results of any analyses are unlikely to overstate the benefit of the proposed therapeutic technology.

Technical Guidance 7
Safety of therapeutic technologies

An assessment of the impact of health outcomes from the use of a health technology includes an assessment of relative safety versus the main comparator. The assessment of safety has two key parts for therapeutic technologies:

- The assessment of the direct and more immediate impacts of the use of the health technology (often captured to a varying degree in the included clinical studies);
- The assessment of longer term or rarer safety events unlikely to be captured in clinical studies.

In some cases, safety outcomes, or harms, can be difficult to distinguish from effectiveness outcomes. For example, the advantages of laparoscopic surgery compared with open surgery may be reduced blood loss, and as blood loss is an established complication of surgery, it may be regarded in some studies as an effectiveness outcome. Where such outcomes represent the key outcomes from clinical studies, they may be presented alongside effectiveness outcomes. Guidance for presenting effectiveness outcomes is provided in Technical Guidance 6.

TG 7.1 Adverse events

Identify the key safety events in studies, and determine whether any important safety events have been omitted. If the omission is related to poor reporting, additional study evidence will be required. If the omission is related to the rarity of the safety outcome, or the insufficient follow up in the clinical study, refer to the guidance below on the extended assessment of safety.

When reporting safety from a clinical study, as a minimum, the following categories of adverse events should be considered:

- any adverse event
- any adverse event resulting in discontinuation of the randomised treatment
- any serious adverse event[18]
- any adverse event resulting in death
- each and every other type of adverse event where the frequency or severity differs substantially across groups.

Where additional adverse events are to be reported (e.g. treatment-emergent adverse events, adverse events of special interest), explain the importance of the adverse event and interpret the result.

Report adverse event data as both the number of patients reporting an adverse event in each category and the absolute number of adverse events in each category. The absolute number of events in each category may be a more appropriate estimate for costing adverse events in an economic or financial analysis, rather than the number of patients who experience an adverse event, because the latter will not capture patients who experience two events in the same category.

For each important adverse event, present these results as for dichotomous data, and include relative risks and risk differences with their 95% CIs across the groups for each study, separately. Where appropriate, meta-analyse the results using a random effects model and provide an interpretation.

Analyse the relative adverse event rates (events per period at risk), if the average period at risk per participant varies substantially between treatment groups (eg using a straight Poisson regression or a negative binomial approach). Present the assumptions associated with statistical analyses and how they were tested.

TG 7.2 Safety unlikely to be captured in clinical studies

The assessment of safety beyond clinical studies is necessary for new therapeutic technologies, or therapeutic technologies used in a new indication. This assessment may therefore be relevant for assessment reports of investigative technologies if the change in management involves a therapeutic technology described above.

An extended assessment of the direct safety of a test may also be useful if a novel mode of testing is used, or there is uncertainty for the longer term or rare effects of testing.

Ideally, the estimate of the relative safety of a health technology is derived from high quality comparative studies. Clinical trials are often inadequate for providing data on comparative harms for a few reasons:

- Trials tend to enrol patients who are healthier, have fewer comorbidities or concomitant medications, and have more stringent monitoring than the target population.
- Trials are usually underpowered and of insufficient duration to detect important adverse events.
- Adverse events in clinical trials designed to emphasise efficacy results are often underreported (Pitrou et al. 2009)

Discuss whether the included evidence base is adequate for identifying:

- Less common adverse events or safety concerns
- Adverse events that may occur in the longer term
- Harms that may occur due to differences in the target population and the more selected population that may be enrolled in a clinical trial

If the included evidence is not sufficient to capture long term or rare adverse events, or adverse events in patients with comorbidities or receiving concomitant treatments, present additional evidence. Describe the search strategy for identifying nonrandomised studies of the proposed health technology, or registry data. Include evidence of safety involving the proposed health technology in other indications, if appropriate. Where the proposed medical service is delivered in combination with an implantable device, provide an assessment of the safety of that device. Sources of safety information may include device registries, regulatory databases, complaints registries and post-market surveillance studies.

Technical Guidance 8
Interpretation of the therapeutic evidence

The objective of summarising the overall evidence base is to describe the results of the assessment report as they apply to the clinical claim in the specific context of the Australian setting.

TG 8.1 Therapeutic evidence interpretation

Provide a summary of the overall evidence base (without repeating evidence from other sections). Consider:

- the level of the evidence, taking account of the directness of the comparison
- the quality of the evidence
- the clinical importance and patient relevance of the effectiveness and safety outcomes
- the statistical precision of the evidence
- the size of the effect
- the consistency of the results across the clinical studies and across subgroups
- the strength or certainty of the evidence
- the applicability of the evidence to the Australian setting
- any other uncertainties in the evidence, including missing outcomes or populations
- other relevant factors that may have an influence on decision making, particularly implementation and ethical factors.

TG 8.2 Conclusion of the clinical claim

The interpretation of the clinical data presented in Section 2 is crucial in determining the success of the submission. It is important to classify the health outcomes of the proposed health technology in relation to its main comparator (ie whether it is superior, inferior or non-inferior to the comparator).

The conclusion of the clinical claim of the therapeutic technology should be a simple and unequivocal statement that is supported by evidence provided in the submission.

Example:

The use of [proposed health technology] results in superior/non-inferior/inferior effectiveness compared with [comparator].

The use of [proposed health technology] results in superior/non-inferior/inferior safety compared with [comparator].

Section 2B
Assessment of investigative technologies

Investigative technologies: A type of health technology that is expected to, or claimed to be able to, generate clinically relevant information about the individual to whom the service is rendered. To achieve an improvement in health outcomes, this information must result in a change in the clinical management of an intermediate intervention. In this sense, investigative procedures can only indirectly improve health outcomes. Examples of investigative technologies are imaging, pathology, genetic testing, and clinical assessments for diagnosis, prognosis, staging, monitoring, prediction of treatment response, surveillance and cascade screening. For ease of reading, the word 'test' is used throughout the document as an alternative term for 'investigative technology', and is intended to reflect the broad range of investigative technologies available.

The clinical component of an assessment of a health technology determines whether the technology is inferior, non-inferior or superior in terms of health outcomes. In some circumstances, the assessment will incorporate other types of benefits (personal utility) or other considerations. The method for assessing the clinical claim for tests will vary according to the type of evidence that is available. Likewise, the relevant TG subsections in these guidelines will vary according to the evidence that is available.

Figure 4 describes the TG subsections that would be relevant for different approaches taken in a test assessment. Technical Guidance 9 explains the types of evidence that may be available or will be required.

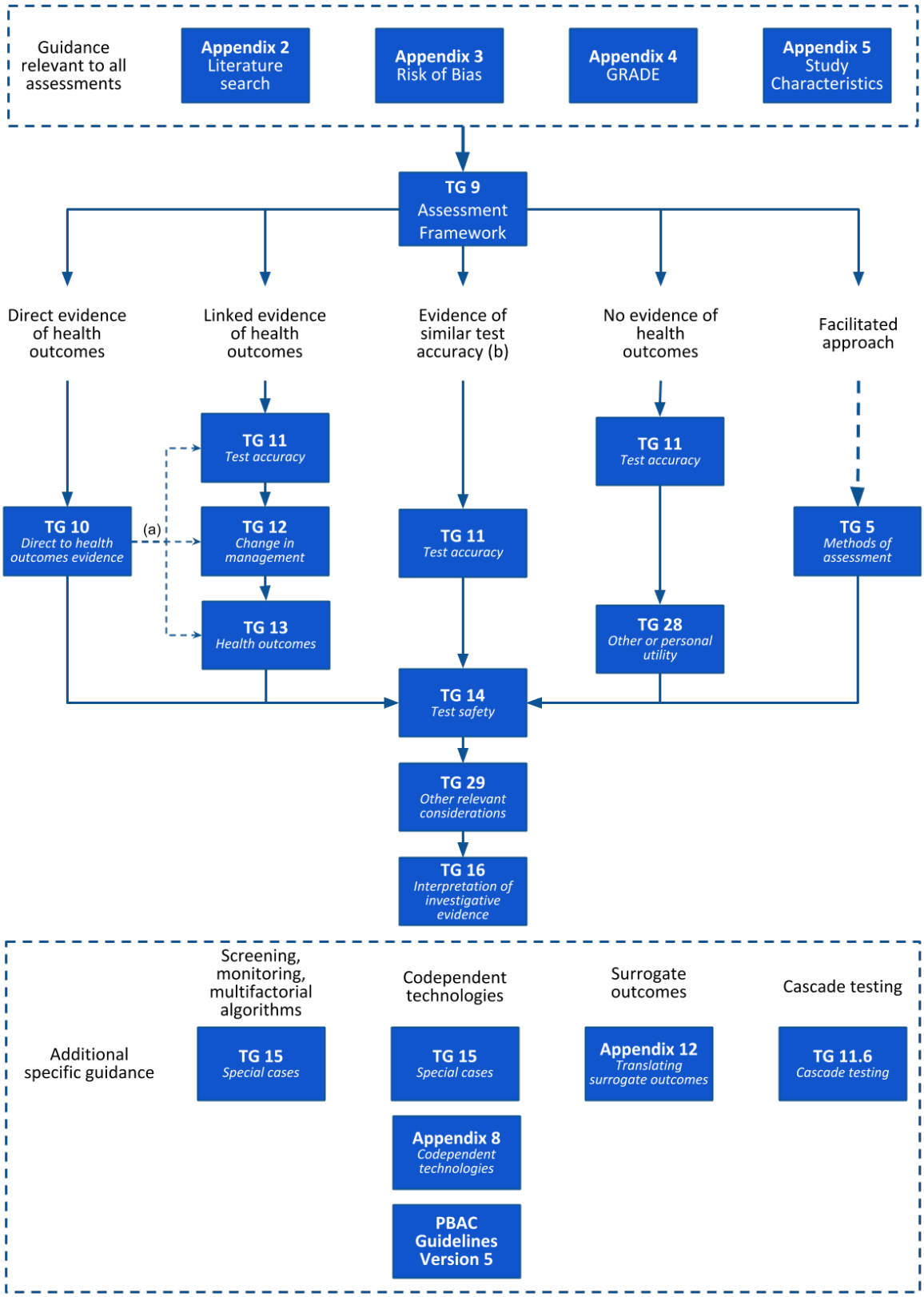


Figure 4 Navigation of the clinical components of the guidelines for investigative technologies based on the type of evidence required (informed by Technical Guidance 9) to assess the clinical claim

a Linked evidence guidance may be relevant to technologies with direct from test to health outcomes evidence if the applicability of the test, management decisions or treatment outcomes is in question.

b An assessment of a technology that relies upon only a comparison of test accuracy is permissible only if this comparison is against a currently reimbursed comparator.

Technical Guidance 9 Assessment framework

An assessment framework is required for the evaluation of investigative technologies.

In general, an assessment framework is not required for therapeutic technologies. However, if evidence for therapeutic technologies does not provide a direct link between the intervention and health outcomes, an assessment framework may prove useful to describe the steps required to make this link. Examples of when an assessment framework may assist in the evaluation of a therapeutic technology may include if the technology results in a change of behaviour or management. This may include health programs (where subsequent behavioural change is required prior to an impact on health outcomes, or prophylactic interventions (where interventions may lower the risk of experiencing an event, which may reduce ongoing monitoring).

The link between a test and health outcomes is rarely clear. The introduction of tests may have limited or no impact on health outcomes, and there may not be a reliable link between the accuracy of a test and the magnitude of health outcomes (Siontis et al. 2014). For a test to have an impact on health outcomes, it must inform a sequence of actions. The method adopted in these guidelines to describe each step from testing to health outcomes is the assessment framework.

This TG subsection describes the method for performing an assessment of the health impacts of a test. The actual assessment of a test is described in subsequent TG subsections. However, there are additional components required for a health technology assessment that are not included in the assessment framework, including other relevant factors (ethics, social, organisational, patient or consumer input), health economics and financial implications.

If the technology can be used for different purposes (ie diagnosis and monitoring), more than one assessment framework may be required.

TG 9.1 Constructing the assessment framework for a health technology

The assessment framework describes the evidence required to verify the clinical claim, and to support the economic approach. The purpose of creating an assessment framework is to help ensure the assessment provides all useful information to MSAC, without providing unnecessary information. The assessment framework is not a management algorithm, although for simple testing strategies, it may resemble one. The structure of the framework describes conceptually the steps between the target testing population and the final health outcomes.

The concept of an assessment framework is based on the process used by the US Preventative Services Task Force to guide their assessments of preventative services and health promotion (US Preventive Services Task Force. "USPSTF procedure manual." (2015)). Many of the USPSTF services have been tests or screening programs, however the frameworks are flexible and can be adapted to represent the steps between any intervention and the consequent health outcomes. The use by MSAC of the “linked evidence approach” is consistent with these assessment frameworks.

The general structure of the frameworks is a diagram that includes populations, outcomes (or information derived at each step) and actions or inferences that link the boxes. The diagram is accompanied by annotated questions corresponding to items in the diagram.

The number of steps in a framework reflect the number of actions or inferences that need to be taken between the decision to test and the final health outcomes. The components of the framework are consistent with the PICO elements described in the PICO Confirmation or as amended in Section 1 of the assessment report.

As a minimum, the initial framework should contain (letters refer to Figure 5):

A. Test population – a brief description of the test population
B. Test – the name of the test
C. A link between the test and the final health outcomes (direct from test to health outcomes evidence)
D. The information provided by the test
E. Change in management / further testing or treatment options
F. Outcomes (if surrogate outcomes are included in the PICO, provide these as a step prior to final outcomes)
G. Adverse events associated with actions (the test or subsequent management decisions)

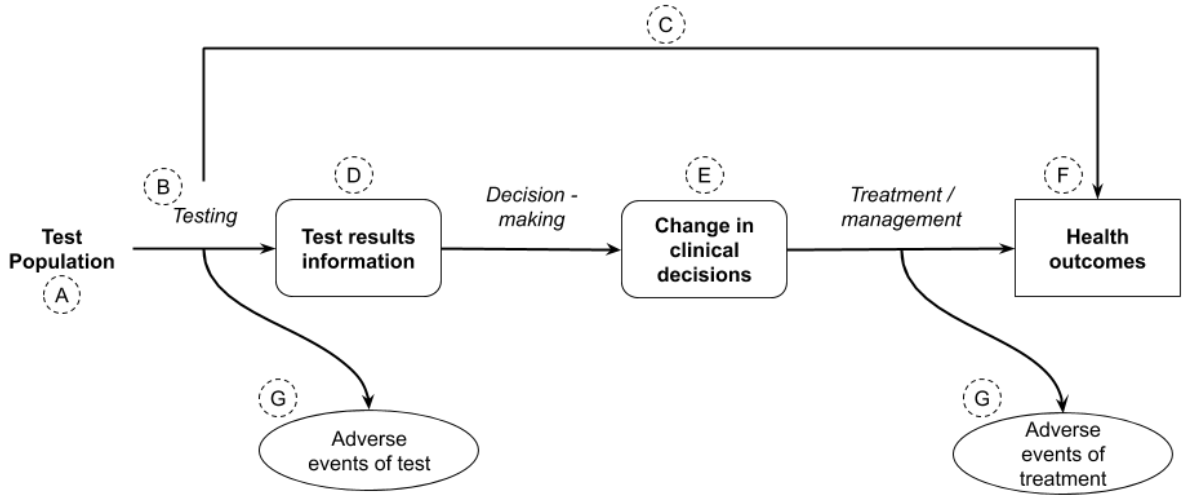


Figure 5 Components of an assessment framework (adapted from the USPSTF procedure manual 2015).

Conventions for the format of the framework:

- Outcomes are depicted using a rectangle, with intermediate outcomes depicted using rounded corners and final outcomes depicted using square corners.
- Harms relating to tests or to subsequent management decisions are often final health outcomes, but are separated from the flow of the framework as they require separate assessment questions (and may commonly require different sources of evidence compared with the clinical effectiveness studies). These are represented as ellipses.
- Actions such as testing, change in diagnostic thinking, change in management / treatment, are represented by arrows connecting the different steps.
- If outcomes are connected by association or inference (such as may occur between an intermediate or surrogate health outcome and the final health outcome), it is connected by a dashed line (see Figure 6).

Once the initial framework has been constructed, add in any additional steps that occur between the point of testing and final outcomes. Common inclusions are subsequent tests, particularly if the initial test is a screening test. Lines are drawn between actions and outcomes, with lines that omit steps representing more direct linkages.

Annotated questions may be derived from any part of the assessment framework, however are necessary for each of the arrows that link the different steps. Assessment questions explicitly

request evidence that compares the proposed test with the comparator. Each link is numbered to enable easy reference to the assessment questions.

Additional components that may be relevant may include non-disease outcomes (such as ethical implications, efficiency outcomes or personal utility outcomes).

TG 9.2 Complete assessment framework required for a claim of superior health outcomes

An example of a generic assessment framework is given in Figure 6 and may be adapted as required.

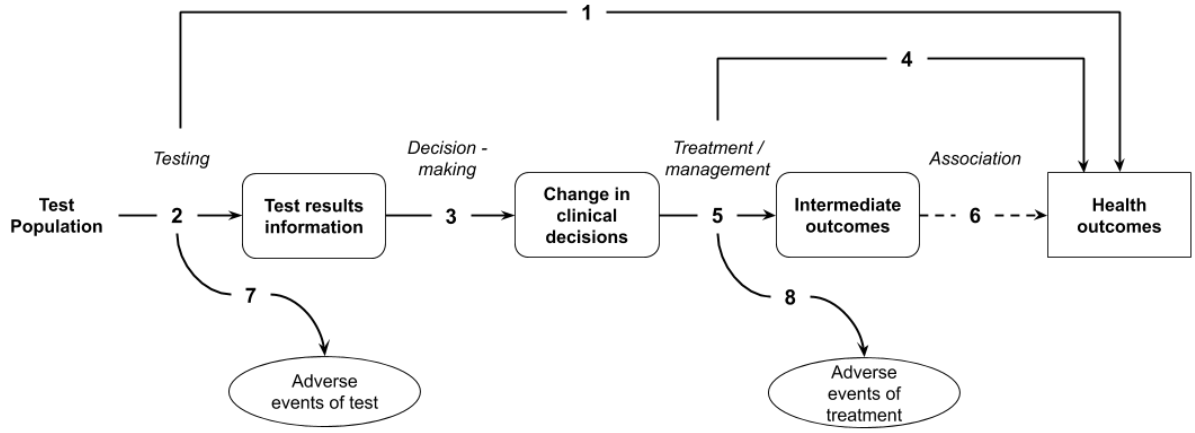


Figure 6 Generic assessment framework showing the links from the test population to health outcomes

Assessment questions for a claim of superiority (Figure 6)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of the test strategy in place of the current test strategy (comparator) result in the claimed superior health outcomes?
    a. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or clinical utility standard?

LINKED EVIDENCE

2. How does the information from the proposed test differ from that of the comparator? What is the concordance of the findings from the proposed test relative to the comparator? What is the accuracy of the proposed test (against a relevant reference standard) compared with the comparator?
    a. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or clinical utility standard?
3. Does the availability of new information from the proposed test lead to a change in management of the patient (compared to the information gained from the comparator)?
4. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention), result in the claimed health outcomes?

5. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention), result in the claimed surrogate outcomes?
    a. Has the treatment/management been provided to a population with the same spectrum of disease that the proposed test identifies? Is it biologically plausible that the treatment/management will be as effective in the population with this spectrum of disease? *There is some concern when the proposed test detects more patients than the comparator as the treatment effect evidence may be based on a more narrowly defined (usually higher risk) positive population where expected benefits may be greater.*
6. Is the observed change in surrogate outcomes associated with a concomitant change in the claimed health outcomes, and how strong is the association?
7. What are the adverse events associated with the proposed test strategy and the comparative test strategy?
8. What are the adverse events associated with the treatments / interventions that lead from the management decisions informed by the test and by the comparator?

Not all steps in the assessment framework need to be presented. For example, if there is adequate direct from test to health outcomes evidence to support the assessment, then a presentation of linked evidence may not be necessary. If there is evidence of the effect of treatment on final health outcomes, then the step including surrogate outcomes may be removed.

Even though there may be uncertainty regarding the availability of evidence for any step of the framework, these steps should not be removed from the framework. In the assessment it should just be made clear that evidence was not available to address the assessment question. However, not all steps will need to be explored if shorter paths to health outcomes are possible.

Additional relevant questions could relate to subgroups in the population, the prevalence of a disease, concordance of the tests used in Australian practice against the clinical utility standard, feasibility/efficiency of the testing procedure, personal utility benefits, ethical issues etc.

TG 9.3 Truncating the framework for claims of non-inferiority

The initial framework identifies the steps between the test and the final, patient relevant health outcomes. The full linked evidence approach cannot be shortened if:

- The clinical claim for the proposed test is superior health outcomes;
- The clinical claim for the proposed test is inferior health outcomes;
- There is insufficient evidence to support a claim of non-inferiority based on earlier steps of the framework (e.g. test accuracy, change in management / decision making). This is particularly relevant for triage tests where there may be a lack of information on the health consequences of being ruled out from subsequent confirmatory testing (Merlin et al. 2013).

In these circumstances, evidence of the impact of the test on health outcomes is required to substantiate the clinical claim, the magnitude of the difference in health outcomes, and to inform the economic analysis.

If the clinical claim is that the test is non-inferior in terms of health outcomes, the framework may be truncated in some circumstances.

Same test accuracy

For claims of non-inferior health outcomes, where the claim is based on the proposed test providing the same information as the comparator, the approach *may be* reduced to a comparison of the information provided by the test and by the comparator.

If the proposed test reports on the same parameter, then the concordance of the proposed test and the main comparator is required. If the tests are concordant, then it may be reasonable to infer that there would be no difference in management, and health outcomes would be non-inferior. The use of concordance in circumstances where the comparator is not the reference standard will not provide an estimate of the performance of either the proposed test or the comparator test. Where the accuracy of the comparator test is uncertain or poor, concordance should be accompanied with test accuracy derived for both tests against an appropriate reference standard.

A more likely scenario is that the proposed test will not be absolutely concordant with the comparator that is currently used in Australia. In these circumstances, it may be reasonable to pursue a claim of non-inferiority and adopt a truncated approach (rather than a full linked evidence approach) if adequate evidence can be provided to support that the proposed test is more accurate than the comparator test. This evidence would include a discussion of the true classification of cases that are discordant between the proposed test and the comparator, and a discussion of the downstream implications of the different test results. The goal of the assessment of a test with a small improvement in accuracy over the comparator will be to establish that the proposed test results in health outcomes that are no worse.

Discordance that results in an increase in false positives or false negatives does not represent an improvement in accuracy and would need to present direct from test to health outcomes evidence or a full linked evidence approach to establish non-inferiority.

When the proposed test is likely more accurate than the comparator (and results in discordant results), the assessment should provide:

- Sensitivity and specificity of the proposed test strategy, and of the current testing strategy, as derived by a comparison with an appropriate reference standard. If no reference standard is available, justify how the discordance is known to represent an improvement in accuracy.
- Quantification of how the tests differ in terms of true / false positives and true / false negatives.
- An explanation of why the tests differ in their detection of the parameter.
- A discussion of the basis upon which evidence based management decisions were established.

Most importantly, the assessment should present the implications of downstream management decisions for the discordant cases. That is, discuss the impacts of the proposed test detecting fewer false positives or false negatives. If the proposed test identifies more positive patients than the comparator, clearly establish that there is no change in the spectrum of the disease (see Technical Guidance 13 for a discussion on the implications of a change in the spectrum of the disease).

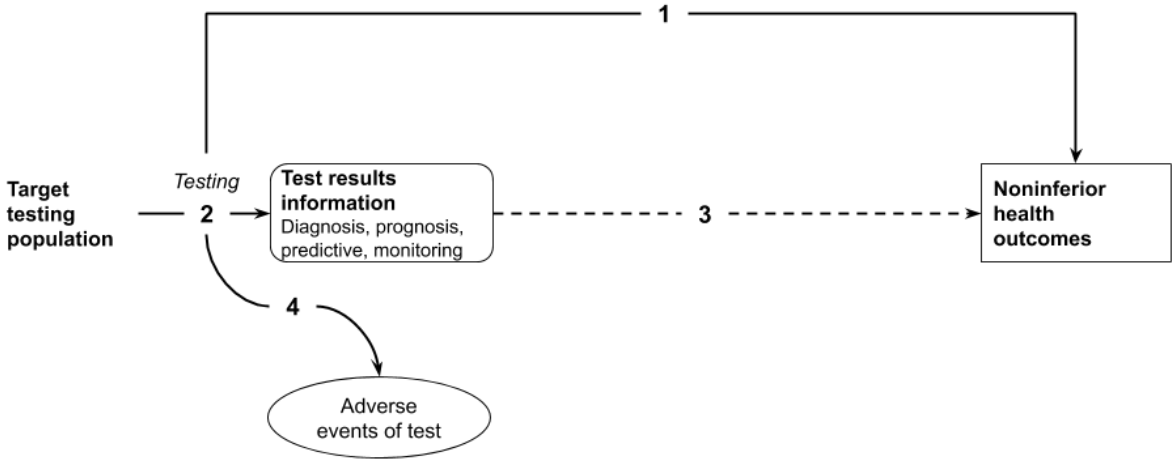


Figure 7 **Assessment framework that has been truncated at test accuracy (concordance, test accuracy) with the inference that identical test accuracy will result in the same health outcomes.**

Assessment questions for a claim of non-inferiority for a test based on comparative accuracy of test results (Figure 7)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of [the proposed test] in the [target testing population] result in [key health outcomes – e.g. survival, quality of life] that are no worse than the [the main comparator]? (If adequate direct from test to health outcomes evidence is available, go to Assessment question 4).

LINKED EVIDENCE

2. Is the [the proposed test] in the [target testing population] concordant with [the main comparator]?
    a. If concordance is unavailable – are the rates of false positive and false negative patients similar in accuracy studies? Is there additional evidence to support that estimates of similar test accuracy indicate that the proposed test and comparator test are categorising patients similarly?
3. *Inference that similar test results from both proposed test and comparator will result in the same management decisions, and non-inferior health outcomes.*
4. What are the harms of [the proposed test] and of [the main comparator]?

Assessment question 2 may require judgement and justification. For tests that are 100% concordant, no discussion of non-concordant cases is required. Where tests are almost 100% concordant, it is important to explain the cause of the non-concordance, for example, are non-concordant cases a consequence of identifying more or fewer patients, and is there evidence (compared with an appropriate reference standard) of their true test result? Non-concordance arising from tests that are less accurate cannot pursue a claim of non-inferiority using a truncated approach. Non-concordance arising from tests that are more accurate should consider a full linked evidence approach to explore the impact of the threshold at which non-concordance would invalidate the claim of non-inferior health outcomes using a truncated approach is not set, and is contingent upon the causes of non-concordance.

Additional frameworks are provided in Appendix 1.

TG 9.4 Adapting the framework for other or personal utility

All tests must make a clinical claim relating to health outcomes. For tests that rely upon an additional claim relating to personal or other utility, add an additional link relating to personal utility in the framework. This can be done for a test that is claiming non-inferior or superior health outcomes.

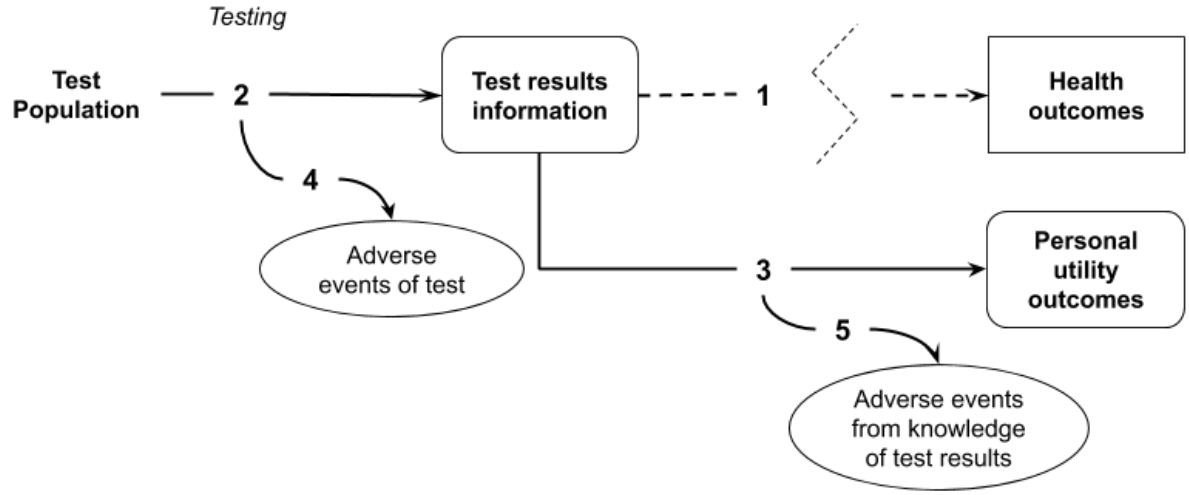


Figure 8 Assessment framework representing a claim of personal utility outcomes.

Assessment questions for establishing a claim relating to personal utility

1. *Address the assessment questions relevant to establishing the claim relating to health outcomes (either non-inferior or superior health outcomes).*
2. How does the information that the test provides differ from that of the comparator?
3. What impacts does the availability of proposed test information compared with the comparator have on personal utility outcomes? What behavioural changes or actions (taken by the individual, families, carers or others) are associated with knowledge of the test results, compared with the information provided by the comparator?
4. What are the adverse effects of the test?
5. What are the adverse impacts associated with knowledge of test results?

TG 9.5 Performing the assessment

Generate the assessment framework

Step 1: Prepare and report the PICO.

An assessment of the proposed test includes defining the relevant population, intervention (including prior tests – may be separated out), comparator (and reference standard) and outcomes. The process for defining these and preparing a clinical management algorithm is discussed in Technical Guidance 2.

Step 2: Construct the assessment framework

Once the PICO is defined, construct the assessment framework based on the PICO. Where there are multiple comparators or multiple populations, the evidence required may still be captured by a single assessment framework (although it would inform multiple sets of assessment questions), or more than one assessment framework may be required.

If permitted by a clinical claim of non-inferiority, truncate the preliminary framework and present the final framework and the arguments necessary to support the truncation.

Step 3: Generate the assessment questions from the assessment framework

Capture the assessment questions related to the assessment framework and record these in the PICO summary tables.

Address the assessment questions

Step 1: Address the direct from test to health outcomes evidence assessment questions (Technical Guidance 10)

For frameworks based on a claim of superior health outcomes, or that cannot be truncated to an earlier step, report on the available evidence for direct outcomes. Also report if no direct from test to health outcomes evidence (or inadequate direct from test to health outcomes evidence) is identified. Adequate direct from test to health outcomes evidence will not require the support of subsequent steps from the linked evidence approach.

Step 2: Address the test accuracy (Technical Guidance 11)

Compare the test accuracy against an appropriate reference standard. If there is a key trial providing direct from test to health outcomes evidence for a particular test (the clinical utility standard), assess how the range of tests used Australia perform compared with the clinical utility standard.

Step 3: Address change in management (Technical Guidance 12)

For frameworks based on a claim of superior health outcomes, or that cannot be truncated to steps earlier than a change in management, report on the evidence for a change in management. Where the clinical claim relies upon a change in management, and evidence for this step cannot be identified, the assessment will not be able to establish the magnitude of the benefit of the test in terms of health outcomes and so a claim of superiority is not appropriate. The assessment report should seek advice from MSAC, and/or incorporate an assumption relating to change in management and highlight this as a major area of uncertainty. The method of derivation of the change in management used in place of an evidence based estimate is required.

If there is evidence of no change in management, a claim of superiority in terms of health outcomes is unlikely to be appropriate (unless there are marked safety benefits). A claim of non-inferiority may still be possible.

Step 4: Address the impact of change in management on outcomes (Technical Guidance 13)

Report on the expected health related outcomes associated with the management decisions. If evidence for final health outcomes is not available, present evidence for validated surrogate outcomes, and translate the surrogate outcomes to final outcomes in a subsequent step.

Step 5: Discuss the consistency and transitivity of the evidence across the framework

Direct from test to health outcomes evidence is preferred to linked evidence. There is greater uncertainty as the number of steps between the decision to test and the final health outcomes increases(Merlin et al. 2013). As with any linked evidence approach, one source of uncertainty is the transitivity of the evidence across each step.

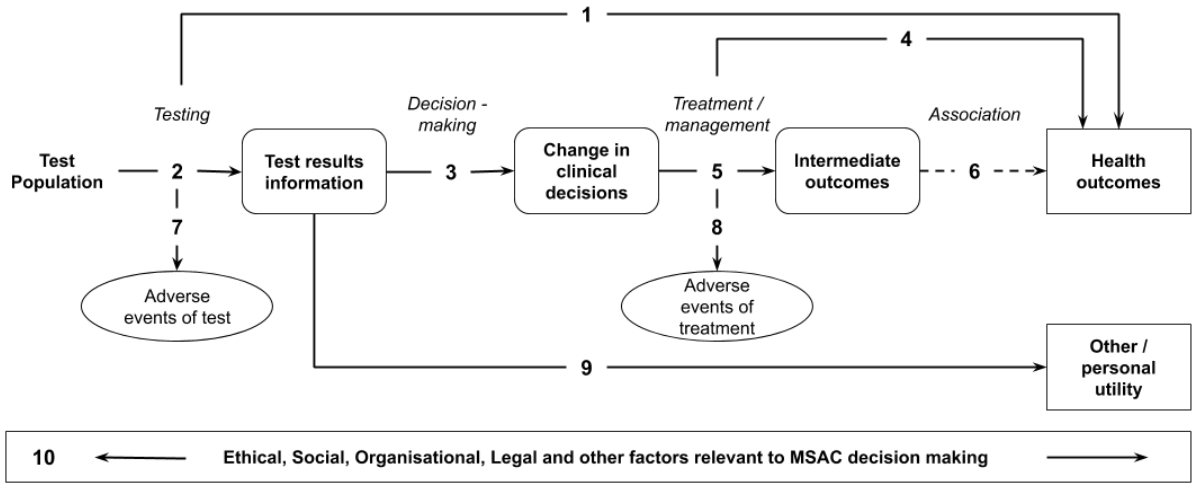
Step 6: Discuss the applicability of the evidence across the framework to the target population and setting of use of the technology

Step 7: Discuss social, ethical, legal and organisational issues associated with implementation of the test

Step 8: Summarise the results

Provide an overarching summary of each of the steps. Discuss uncertainties in each of the steps, and the implications of these uncertainties on the final estimate of health outcomes.

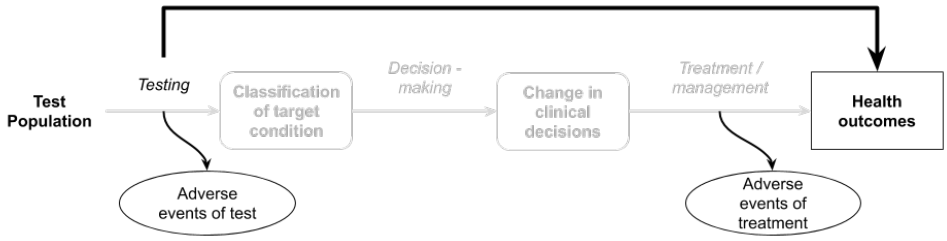
TG 9.6 Location of guidance provided for assessment questions



These guidelines provide advice relating to each of the steps in the framework.

Step in the above framework	Description	TG subsection(s)
1	Presentation of direct evidence of the impact of testing on health outcomes	Technical Guidance 10
2	Test accuracy, sensitivity/specificity, concordance	Technical Guidance 11
3	Evidence of change in management	Technical Guidance 12
4	Evidence of the effect of treatment/management on health outcomes	Technical Guidance 13
5	Evidence of the effect of treatment/management on surrogate outcomes	Technical Guidance 13 and Appendix 12
6	Evidence of the association between a change in the surrogate outcome and the target final outcome	Appendix 12
7	Evidence for the safety of the test	Technical Guidance 14
8	Evidence for the safety of downstream management decisions	Technical Guidance 14
9	Evidence for the impact of the test on other or personal utility	Technical Guidance 28
10	Evidence for social, ethical, legal and organisational impacts associated with implementation of the test	Technical Guidance 29

Technical Guidance 10
Direct from test to health outcomes evidence



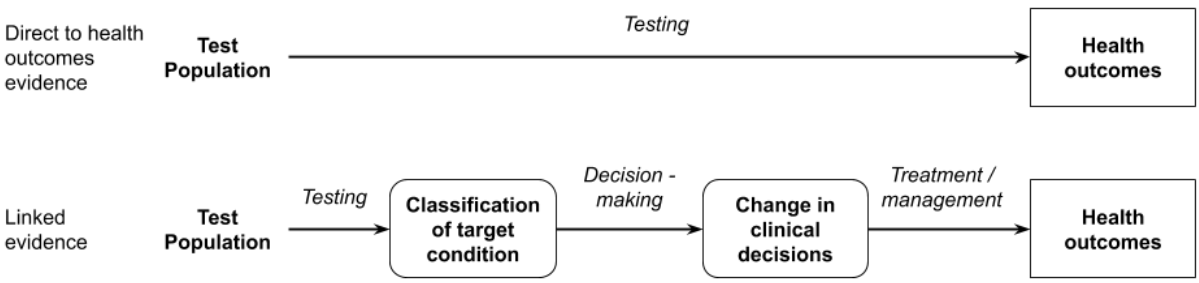
TG 10.1 Purpose of guidance

The assessment of investigative technologies involves identifying the impact of a test on health related outcomes and in some cases on personal utility outcomes. The evidence required to establish an impact on health outcomes can involve either a direct from test to health outcomes evidence approach, or a linked evidence approach. This TG subsection relates to a direct from test to health outcomes evidence approach and discusses:

- A definition of direct from test to health outcomes evidence
- Study designs that can be used to inform a direct from test to health outcomes evidence approach
- Considerations relating to this type of approach including limitations and gaps that may arise from particular study types
- A suggested approach to presenting direct from test to health outcomes evidence for an investigative medical service

TG 10.2 Direct from test to health outcomes evidence

Direct from test to health outcomes evidence refers to evidence from trials or studies specifically designed to inform the effect of a test on a therapeutic outcome. Direct from test to health outcomes evidence may be called clinical utility evidence for simplicity. Clinical utility evidence is characterised by the measurement of key patient relevant health outcomes in the same study in which patients receive a test that informs treatment decisions. This is distinct from a linked evidence approach in which either the effect of test results on change in management, or the effect of change in management on health outcomes, are not captured in a single study.



TG 10.3 Direct from test to health outcomes study designs

Direct from test to health outcomes evidence can encompass a variety of study designs, both comparative and non-comparative. However, given MSAC are informed by a comparison of the proposed test with current practice (to establish incremental benefits and harms), non-comparative studies of the proposed test will require additional evidence of the main comparator, in order to perform the relevant comparison, as well as supportive evidence to establish transitivity of the studies.

There are several study designs that may provide clinical utility evidence, however comparative studies that randomise patients to the proposed test vs comparator test are preferable.

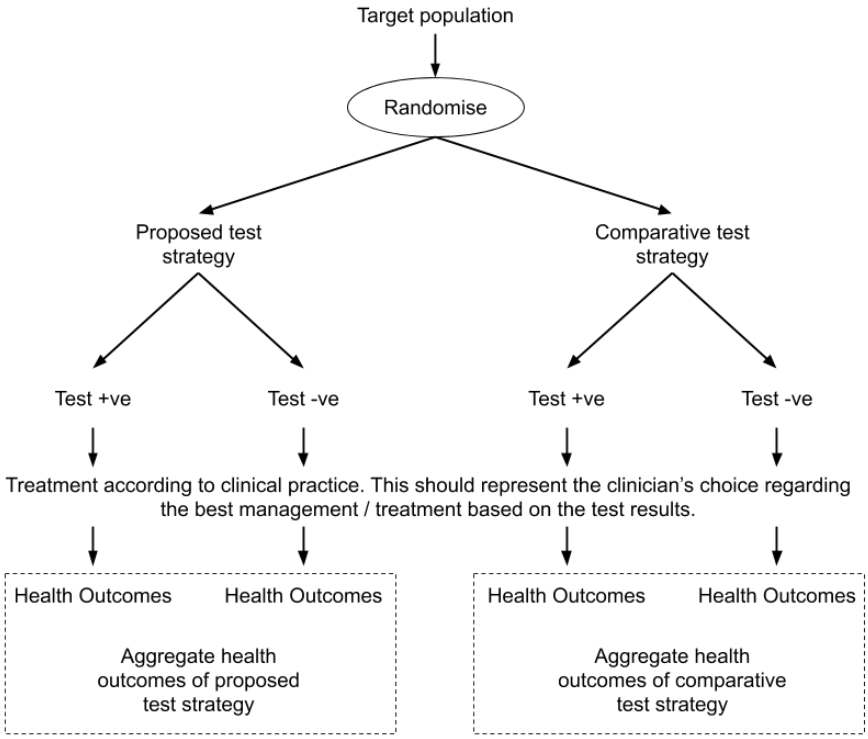


Figure 9 Single-randomised direct from test to health outcomes trial evidence comparing health outcomes from new test strategy and existing test strategy (test information used to derive treatment condition) (Preferred option 1 for assessment of any claim regarding clinical utility of a test)

A study that randomises subjects to receiving the proposed test vs the comparator test (or standard practice) will provide evidence of the overall comparative health benefit of the test. There are many study designs that will provide part of the information derived by this design. Studies that provide incomplete estimates of the comparative impact of a test on health outcomes should be used with caution (see below).

TG 10.4 Considerations relevant to a direct from test to health outcomes evidence approach

Direct trials of tests may not provide results which are transferable or generalizable to the target population, and are often underpowered to detect a difference in health outcomes (Doust, J 2010). In these circumstances, additional components of linked evidence may be beneficial to assess.

Observational studies

Observational (nonrandomised) comparisons of patients who receive a test vs those who do not may represent selection due to adherence or compliance, and other potential confounders, rather than random chance. Consequently, those who adhere to testing may differ systematically to those that refuse or do not seek the test (Pletcher & Pignone 2011). This confounding may have a large impact on the subsequent health outcomes and should be treated with caution.

Incomplete direct from test to health outcomes evidence

Ideally, the health outcomes following the use of the proposed test are compared with the health outcomes following the use of the comparative test strategy (standard practice) in the same study. Some studies will provide only health outcomes associated with the use of the proposed test and not the comparative test strategy.

If more than one study is required to describe the outcomes associated with the proposed test with that of the comparative test strategy, the transitivity of these studies must be adequately described. Indirect comparisons of test to health outcome studies have a high risk of bias. Unlike indirect comparisons of treatments, where often the largest uncertainty relates to the baseline characteristics of the populations, an indirect comparison of tests may have transitivity issues with the populations and prevalence of a biomarker, the thresholds of the tests, the clinical practice decisions and the treatments. As such, indirect comparisons of test to health outcome studies should be avoided, where possible. If they are required, the transitivity of each of the components of the studies must be rigorously assessed (see Appendix 2 for sources of heterogeneity).

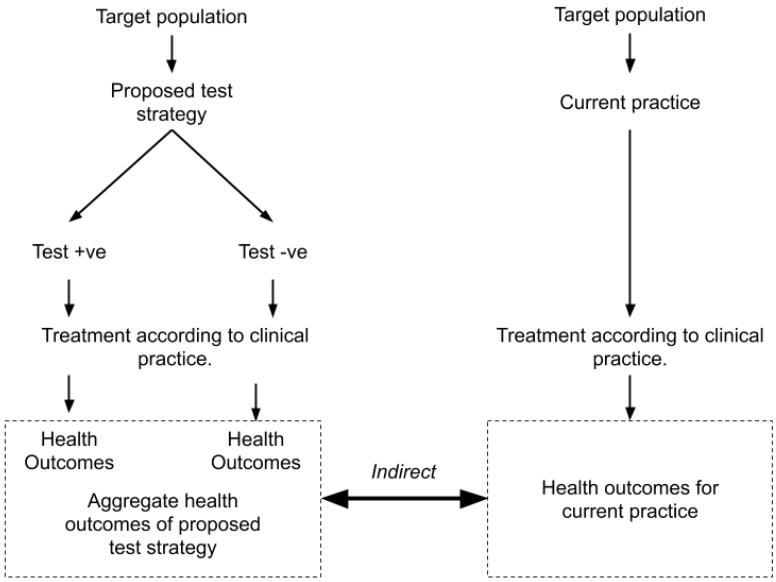


Figure 10 Single arm study of the test reporting on health outcomes with an indirect comparison of current health outcomes in the absence of the test (or using an alternative test)

TG 10.5 Assessment of the applicability of direct from test to health outcomes evidence

Direct from test to health outcomes evidence is characterised by a study that measures the impact of health outcomes following the use of a test. While this type of evidence provides more robust internal validity than a linked evidence approach, the applicability of the evidence must be assessed to ensure that the results from the study will be replicated in clinical practice.

Concerns relating to applicability arise when the population or interventions (and/or circumstances of use) in a study differ from the target setting. For therapeutics, an assessment of applicability primarily consists of comparing the populations and the interventions across the study and target setting, and identifying differences that may have an impact on the expected outcomes.

The applicability of direct from test to health outcomes evidence depends on the applicability of not only the population and the test, but also the actions or subsequent steps taken in the study. Where there are concerns with the applicability of any component of the direct from test to health outcomes evidence, additional supportive evidence may be required.

An assessment of applicability includes a comparison of the study and the Australian setting for the following characteristics:

Table 7 Applicability assessment of direct from test to health outcomes evidence

Applicability domain	Sources of applicability concerns	Additional supportive evidence that may be required
Applicability of the testing population	Baseline patient characteristics *Prevalence of the biomarker / disease*[a]	Comparison of the prevalence of the biomarker in the trial with that of the target testing population
Applicability of the test	Is the test used in the study (the clinical utility standard) the same as the proposed test?	Comparison of patient classification through use of the clinical utility standard with that of the tests developed in Australia.
Applicability of the clinical decision making.	Will clinical practice in Australia reflect the clinical practice in the study? Was the choice of treatments at clinician discretion in the study, or was it protocol driven?	Additional evidence that the test will impact on clinical decision making. Evidence that the treatment options are the same in Australia as in the study.
Applicability of the treatment / management.	Was the treatment delivered in the study as it would be delivered in Australia (in terms of clinical setting, dose, duration, concomitant therapies etc).	Evidence that treatments are delivered similarly in the study as in the Australian setting.

a The prevalence of the biomarker is fundamental to the assessment of the applicability of direct from test to health outcomes evidence. Prevalence of the biomarker in the study may be influenced by methods for identifying the test population or may vary by race (particularly if the biomarker is a genetic marker). Eligibility criteria (such as requiring particular symptoms or including only high risk populations) may be used to enrich the biomarker positive population. If the prevalence in the study differs from the prevalence in the Australian target population, the aggregate health outcomes from the study will not be valid.

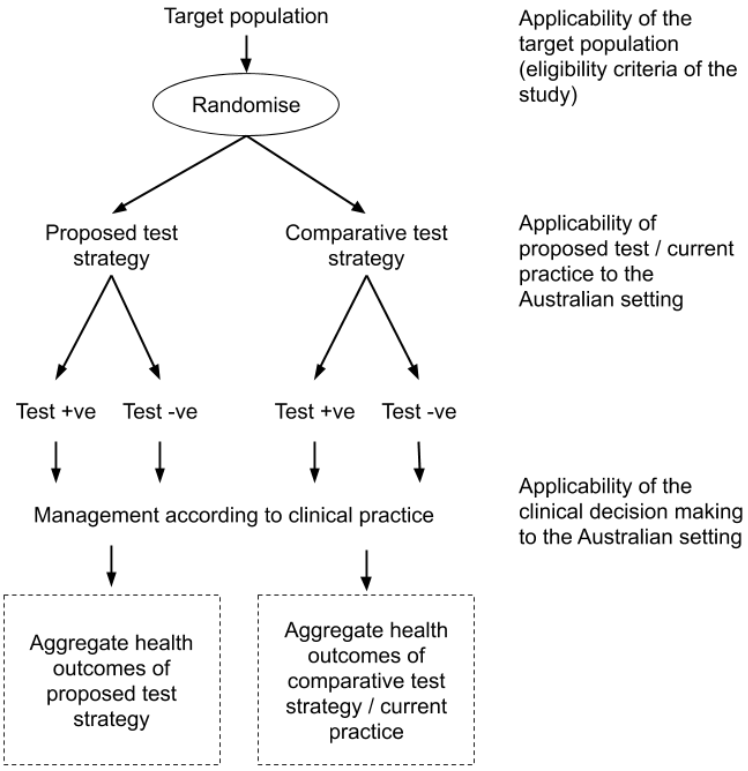


Figure 11 Assessment of applicability of direct from test to health outcomes evidence

Applicability of the study population

Where the proportion of patients with the biomarker (prevalence) differs in the study to the target population, the aggregate health outcomes associated with the test may also differ. Additional evidence comparing the testing population and the trial population is required. The population in the study may differ from that of the target population in several ways, however a key concern is if the trial population has been enriched. Enrichment can occur by using eligibility criteria that narrow the testing population to subjects with a higher risk of having the biomarker (this would result in a difference in the proportions of patients who receive a particular test result (ie, prevalence of a biomarker) in the trial compared with the target population. If prevalence differs, the impact on health outcomes may be re-estimated by weighting the outcomes by test result in the trial (if reported separately) before aggregation. This weighting may not be possible depending on the outcome measure, but may be possible to achieve in the economic analysis.

Applicability of the study test (clinical utility standard)

The health outcomes observed in a direct from test to health outcomes study are relevant to the characteristics of the test used in the study (the clinical utility standard).

Where the proposed test is not the same as the clinical utility standard, or there are multiple tests that may be available in clinical practice (that may be eligible for the same funding arrangement), health outcomes may differ if test characteristics differ from the clinical utility standard. This direct from test to health outcomes evidence must be augmented with evidence that compares the test characteristics (sensitivity, specificity and/or concordance) of the tests that will be available in clinical practice with the clinical utility standard (see Technical Guidance 11).

Applicability of decision making

For the health outcomes associated with the proposed test or comparative test strategy to be valid, the change in management observed in the study must mimic the change in clinical practice following the availability of the proposed test.

Where the management decisions associated with test results (such as treatment with a particular medicine or a surgical procedure) are written into the protocol of the study, the study cannot provide evidence of change in management. Evidence with protocol enforced management must be augmented with evidence for change in management (see Technical Guidance 12).

Applicability of treatment / management

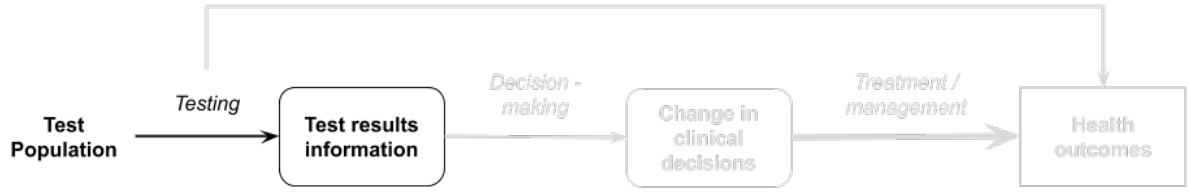
The final step for assessing applicability relates to the treatment or management that is provided in the study. This is similar to assessing the applicability of a therapeutic to the target setting. It is important to compare the treatments in terms of dose, duration and frequency and concomitant treatments. It is also important to compare the setting and other circumstances of use.

TG 10.6 Presentation of direct from test to health outcomes evidence

The principles for presenting direct clinical trial evidence of the effect of an investigative medical service on patient health outcomes are similar to that for presenting clinical trial evidence relating to a therapeutic medical service. However, there are additional components, described below, required to present this information clearly and comprehensively:

1. Describe how the direct from test to health outcomes evidence has been constructed. If multiple sources of information have been used, describe why they are necessary.
2. Present the direct from test to health outcomes evidence in the same way as presenting a therapeutic medical service (ie, describe the literature search, risk of bias, trial characteristics, present the results and meta-analyse, if appropriate).
3. Describe applicability concerns of the direct from test to health outcomes evidence. Explain how additional evidence has been used to address issues with the applicability.
4. Present evidence of the harms (adverse events) experienced by patients who receive the proposed test vs the comparative test (standard practice). These harms should include direct test related harms and harms that are associated with subsequent management decisions (see Technical Guidance 14).

Technical Guidance 11 Linked evidence - test accuracy



TG 11.1 Purpose of guidance

In the absence of high quality direct from test to health outcomes evidence, an assessment will take a linked evidence approach. One key uncertainty is whether patients are appropriately categorised by the test (e.g. test accuracy). This information is needed so that the flow-on effects of test categorisation on subsequent evidence linkages, exploring how the proposed test would change patient management and its likely impact on patient health outcomes, can be determined.

This TG subsection will discuss the methods required to determine the proportion of patients who will be appropriately classified by the test and the proportion who will not. The results from this TG subsection will provide guidance on:

- Estimating the sensitivity and specificity of the proposed test. These measures of test accuracy are compared with those of the main comparator by way of a reference standard test
- Evaluating concordance in the results of different tests that may be available in Australia
- Estimating the prevalence of the disease and/or biomarker in the target population
- Assessing the downstream implications for patients with false-positive and false-negative test results, or in the absence of a reference standard, those with discordant results

The assessment of test accuracy describes the proportion of patients who are identified as true positive, true negative, false positive and false negative. The assessment report must clearly define what is meant by each of these categories, particularly when a clinical utility standard or an imperfect reference standard is used as the benchmark, rather than a known and accepted reference standard. The assessment report must also comment on the possible implications of changes in the number of test positives and test negatives compared to the current testing situation. It must be reported whether patients classified as true positive by the proposed test likely reflect the spectrum of disease that were historically classified as true positive by the reference standard, clinical utility standard or another comparator test.

Ultimately, the assessment of test accuracy seeks to establish how to classify patients for subsequent steps in the management algorithm. By doing so, it also estimates the likely treatment effect or clinical benefit that may be ensue if the test becomes widely available. Treatment effects may differ across true and false positives, for example, and so current evidence of treatment benefit - generated in the absence of the proposed test - may not be applicable.

Tests can be performed for a variety of purposes including:

- Diagnosis of disease in symptomatic patients
- Determining suitability for a targeted treatment in patents with disease
- Monitoring of disease status/progression in affected patients
- Treatment outcome assessment of affected patients
- Prognosis or prediction of future disease outcomes in symptomatic and/or high risk patients
- Risk assessment in asymptomatic patients at increased risk
    - Monitoring of disease occurrence/recurrence

- Cascade testing of relatives at risk of having an inheritable condition
- Screening or carrier testing of the general population

The approach taken to assess test accuracy depends on whether the test determines a current health state (eg, diagnostic) or a future health state (eg, prognostic or predictive) (Figure 12). These two categories of tests are often accompanied by different evidence:

- If the test determines a current health outcome, then cross-sectional studies will provide the evidence base for test accuracy measures. The accuracy of the proposed test should be assessed as outlined in TG 11.3. This would include tests conducted for diagnostic purposes, as well as those used for triaging, monitoring, screening and staging.
- If the test determines a future health outcome, then test accuracy should be assessed using longitudinal data as outlined in TG 11.4. Should the evidence base for the prognostic and/or predictive test include only cross-sectional studies the test should be assessed in the same way as outlined in TG 11.3.

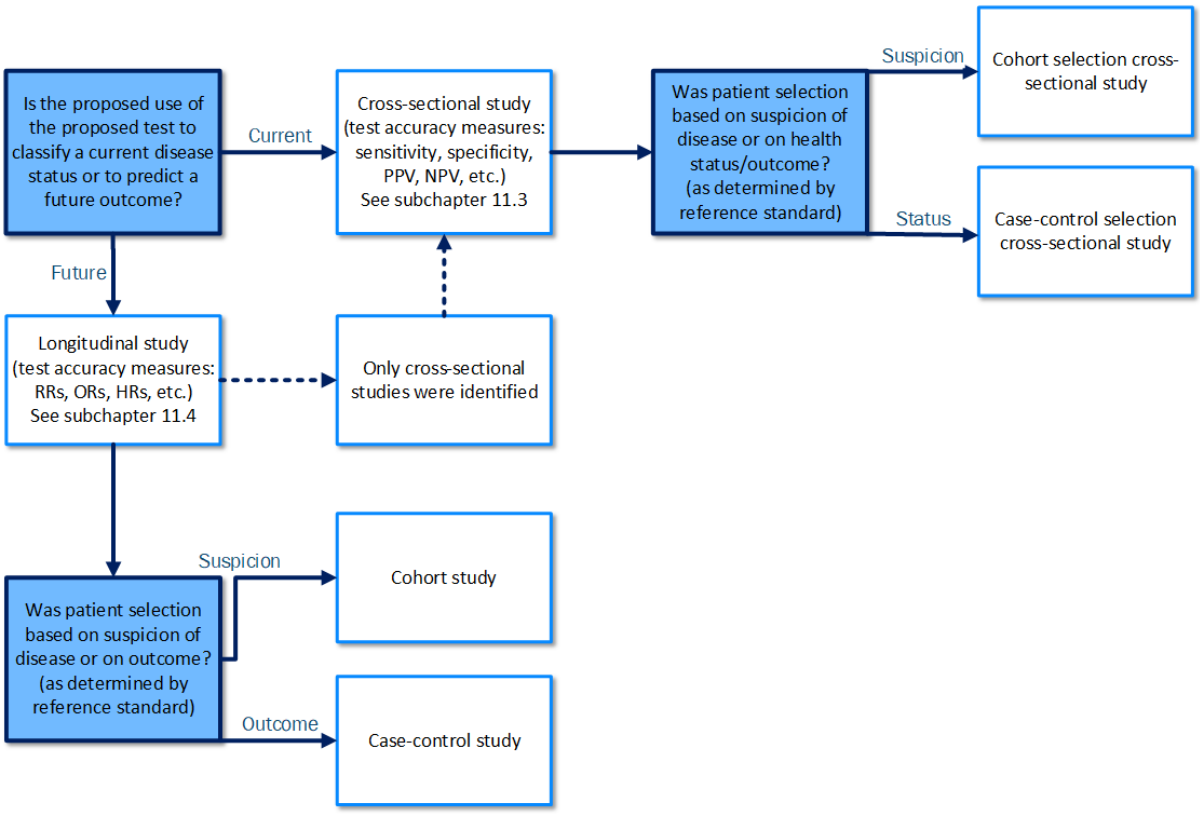


Figure 12 Flowchart for determining the type of evidence and the approach to the assessment of test accuracy (adapted from (Mathes & Pieper 2019))

TG 11.2 Key concepts

Comparison between terminology used in the previous and the new MSAC guidelines

The 2017 MSAC Technical Guidelines (for preparing assessment reports for the Medical Services Advisory Committee – Service Type: Investigative) adapted the terminology used by the ACCE framework for evaluating genetic tests (Haddow & Palomaki 2004). ACCE is an acronym of Analytic validity, Clinical validity, Clinical utility and Ethical, legal and social implications.

However, use of the term 'analytical validity' can cause confusion. Analytical sensitivity and specificity of a test has a broad definition that includes technical aspects of test accuracy, such as the limits of

detection and quantitation, the measuring range, linearity of the test (sensitivity), and factors that may cause interference or cross reactivity and affect test results (specificity).

In the ACCE model, analytical validity refers to the sensitivity and specificity of the test measured against a non-clinical reference standard that measures the biomarker, as well as reliability and reproducibility measures. Clinical validity refers to the test's ability to identify or measure the target condition, such as threshold value, test sensitivity and test specificity, positive and negative predictive value.

To avoid further confusion in the current guidelines, the terms, analytical validity and clinical validity, have been replaced by 'test accuracy'. This term is used whether the reference standard is clinical (measuring the disease or condition) or non-clinical (measuring the biomarker). Figure 13 shows how the ACCE model was incorporated into the previous and the new MSAC guidelines.

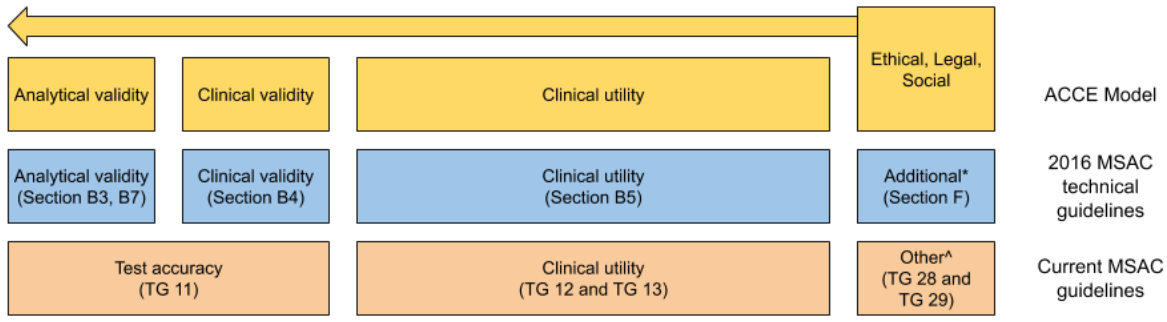


Figure 13 Comparison of terminology used in the previous MSAC technical guidelines and the new MSAC guidelines, compared with the ACCE model

*Additional = options to present additional relevant information
^Other = other relevant considerations including organisational, social, legal and environmental issues.

Comparisons between proposed tests, comparator tests, and reference standards

The comparisons that are required in an assessment report will differ across applications. It is important to provide an explanation for why each of the comparisons is necessary for the interpretation of subsequent steps in the linked evidence approach. Providing comparisons against all possible tests without justifying the need for the comparison, or interpreting the result, is unhelpful.

Figure 14 describes the comparisons that may be relevant in an assessment of test accuracy, depending on whether or not a reference and/or clinical utility standard have been identified. The various reference standards that may be available are discussed below.

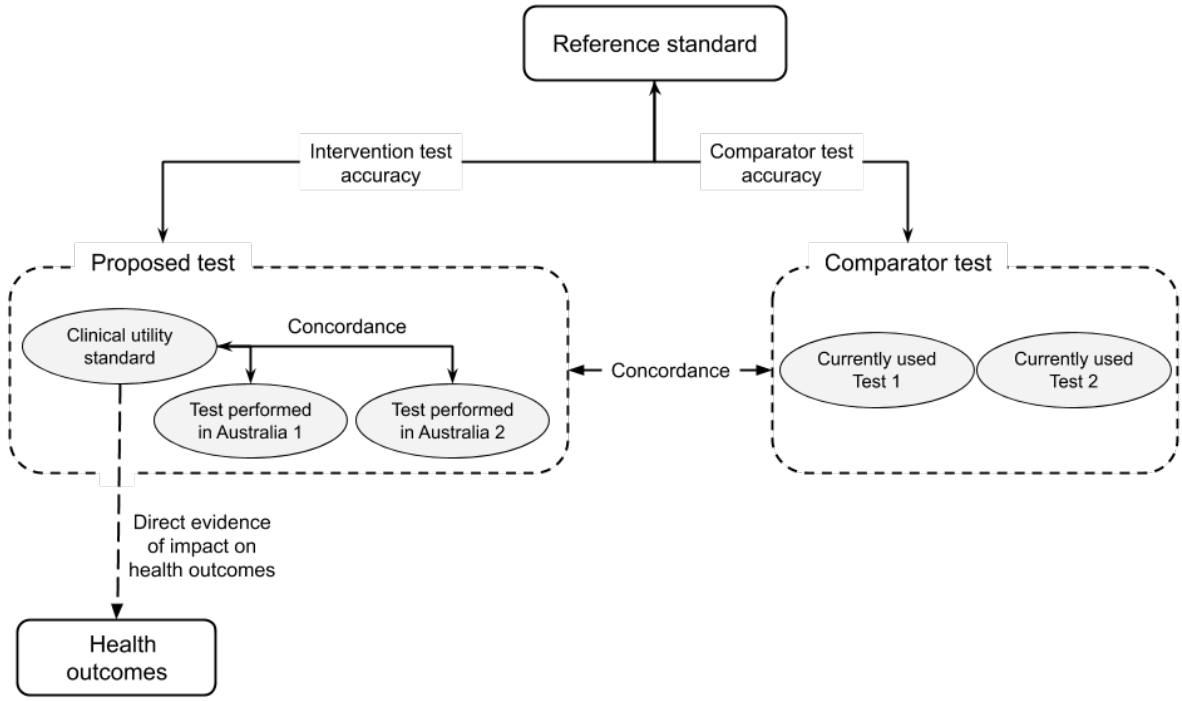


Figure 14 Comparisons of interest to determine the accuracy of the proposed test compared with other available tests

The assessment of the accuracy of the proposed test may include one or more of the following comparisons:

- If a reference standard is available, the accuracy of the proposed test(s), the clinical utility standard and any comparator tests that may be used in Australia should be measured against the reference standard.
  Accuracy measures could include sensitivity and specificity, PPV and NPV, etc. The health outcomes for false positive and false negative patients should be discussed.
- If a clinical utility standard is only available, the accuracy of the proposed test(s) should be compared to the clinical utility standard.
  Concordance (positive and negative percent agreement) is the most appropriate test accuracy measure but a discussion of discordant results being considered as false positive or false negative with respect to the clinical utility standard is required.
- In the absence of both a reference and a clinical utility standard, the test concordance between the proposed and comparator tests is required.
  In this scenario, patients with discordant outcomes cannot be identified as having either true or false test results with respect to either test.

In many cases, the comparator test will be a currently used test, or it may be no testing. There will also be circumstances where the comparator test could be the reference standard, an imperfect reference standard or the clinical utility standard.

The reference standard

The term ‘reference’ or ‘gold’ standard refers to a benchmark that is the best available test under reasonable conditions that has a standard with known results. It is not likely to be a perfect test, but merely the best test currently in use. The identification of the reference standard is discussed in TG 2.4.

The availability of an appropriate reference standard creates more certainty around the evidence presented. This allows quantitative assessment of sensitivity and specificity and informs whether the proposed test is superior or non-inferior to the main comparator in terms of accuracy and reliability.

The reference standard may be either non-clinical (comparing the ability to detect a biomarker) or clinical (comparing the ability to detect a disease or symptom of disease). If a clinical reference standard is available, then the accuracy of the proposed test against this clinical reference standard would be preferred over the non-clinical reference standard in most cases.

A clinical reference standard may often be a composite standard, involving multiple tests and clinical assessments to diagnose the disease, and may at times, include the results from a comparator test or even the proposed test itself. Note that if the reference standard incorporates information from the proposed test, the results will be subject to incorporation bias (Whiting et al. 2011).

Clinical utility standard

If a reference standard is not available, a clinical utility standard may be used. A clinical utility standard is the test that was used to generate direct clinical outcomes in patients with and without a biomarker. If the clinical outcome is response to a targeted treatment, the clinical utility standard may be registered with the TGA as a 'companion diagnostic'.

Any false-positive and false-negative results due to the use of the clinical utility standard are accounted for in the clinical outcomes. A comparison of the concordance of the proposed test compared to the clinical utility standard will identify if either additional or fewer patients would receive targeted treatment if the proposed test is used. Health outcomes for patients with discordant test results should be discussed, conservatively, as if they were false positive or false negative with respect to the proposed test. In the absence of a reference standard it is not known which of the tests being compared is the most accurate.

Imperfect reference standard

Reference standard tests may be imperfect, and incorrectly identify a proportion of the population as test positive or test negative. Often the imperfect reference standard will be well established in diagnostic laboratories for routine diagnosis of the biomarker or condition (Glasziou, P., Irwig & Deeks 2008). This is most likely due to a more accurate test being unavailable or unnecessarily invasive.

When comparing the accuracy of the proposed test to an imperfect reference standard, care should be taken when interpreting the false-positive and false-negative rate. If the proposed test is more accurate, these "false" test results may actually be true positives and/or negatives that are misclassified by the imperfect reference standard. In cases where the imperfect reference standard is clearly inferior in terms of accuracy (both sensitivity and specificity) to the proposed test, and it is not used to direct treatment, a comparison against an imperfect reference standard is of limited value, and an alternative (such as a clinical reference standard) may be more relevant.

Trikalinos and Balion (2012) indicate that test accuracy measured against an imperfect reference standard can be assessed in four different ways:

- Assess the proposed test compared with a clinical reference standard instead of the imperfect reference standard
- Assess the concordance or agreement between the two tests
- Calculate "naïve" estimates of the index test's sensitivity and specificity compared with the imperfect reference standard but qualify study findings and discuss in which direction they are biased

- Adjust the "naïve" estimates of sensitivity and specificity of the index test to account for the imperfect reference standard. The "adjusted" approach generally requires patient level data to be available.

When a comparison against an imperfect reference standard is required the approach taken should be justified. In many cases, additional supplementary evidence may be available to support the conclusion of improved sensitivity and/or specificity of the proposed test compared with the imperfect reference standard.

Partial verification and differential verification

In some instances, test accuracy studies may use the reference standard only as a confirmatory test. In these studies, only those samples/patients with a positive test result are tested with the reference standard. These studies should only be included if no studies comparing both tests for all samples/patients are available.

It should be noted that if not everyone receives the reference standard (i.e. only positive screening tests have further testing), the results will be subject to partial verification bias (Whiting et al. 2011).

One common example of the incomplete use of the reference standard occurs with next generation sequencing (NGS)-based tests. The reference standard applied is typically older more established tests:

- Sanger sequencing to identify single nucleotide variants and small insertions/deletions
- Multiplex ligation-dependent probe amplification and/or quantitative polymerase chain reaction to identify copy number variants, large insertions/deletions and gene rearrangements

In other cases, a reference standard may not be able to be applied. For example, in screening mammography, observed lesions may be biopsied to determine the presence of cancer with histopathology (reference standard). Mammograms without an observed lesion cannot be biopsied. To determine whether a negative mammogram was accurate, the patient will have to be followed up to see if cancer is detected later (differential verification).

Use of a clinical versus a non-clinical reference standard when evaluating test accuracy

In some circumstances (such as biochemical, cytogenetic and molecular genetics), it is important to distinguish between how accurate the test is at detecting a biomarker, versus how accurate it is at detecting the clinical disorder or outcome of interest.

Note that if good quality data applicable to the Australian setting against a valid clinical reference standard are available, then test accuracy against a non-clinical reference standard may not needed.

Examples of clinical and non-clinical reference standards and a clinical utility standard for some example tests are shown in Table 8.

Table 8 Clinical and non-clinical reference standards and/or clinical utility standards for some example tests

Test	NAAT for tuberculosis	*BRCA1/2* variant test	*CFTR* variant carrier testing	Digital breast tomosynthesis
Purpose of test	Diagnosis of TB	Determine presence of a *BRCA1/2* class 4 or 5	Determine *CFTR* carrier status of	Diagnose breast cancer

		variant as a surrogate measure of likely response to a PARP inhibitor	family members of someone with cystic fibrosis	
Non-clinical reference standard	Ability of the test to accurately detect *mycobacterium tuberculosis* (RS: microbial culture of suitable specimens) Accuracy measures: sensitivity, specificity	Ability of the test to accurately detect *BRCA1/2* class 4 or 5 variants (RS: Sanger sequencing ± MLPA to detect *BRCA1/2* variants) Accuracy measures: sensitivity, specificity, likelihood ratios, PPV, NPV, post-test probability of having the biomarker	Ability of the test to accurately detect the familial *CFTR* variant (RS: Sanger sequencing) Accuracy measures: diagnostic yield	N/A
Clinical reference standard	Ability of the test to detect a case of tuberculosis (RS: composite reference standard including clinical findings, microscopy, histology. cytology, chest radiographic findings, site-specific CT scan/ MRI results, culture results and response to anti-TB drug therapy) Accuracy measures: sensitivity, specificity, PPV, NPV, likelihood ratios	N/A due to heterogeneity of the tumour genomes (pathogenic variants in other genes may influence response to PARP inhibitors)	N/A as no clinical RS for family members	Ability to detect architectural distortions, focal asymmetries and micro-calcifications in benign versus malignant cancers (RS: histological examination of biopsy samples) Accuracy measures: sensitivity, specificity, PPV, NPV, likelihood ratios
Clinical utility standard	N/A	Ability of the test to predict response to treatment in biomarker positive patients (CUS: *BRCA1/2* variant test used in RCT with direct health outcomes)	N/A	N/A

BRCA1/2 = breast cancer gene 1 and 2; CUS = clinical utility standard; CT = computed tomography; HbA_{1c} = glycated haemoglobin A_{1c}; MLPA = Multiplex ligation-dependent probe amplification; MRI = magnetic resonance imaging; NAAT = nucleic acid amplification testing; N/A = not applicable; PARP = poly ADP ribose polymerase; RS = reference standard; TB = tuberculosis

In the absence of a clinical reference standard, the clinical accuracy of a test depends on both the ability of the test to detect the biomarker compared to the non-clinical reference standard, as well as the strength of the biological plausibility linking the surrogate measure with the clinical condition of interest. For example, the link between HbA_{1c} levels and blood glucose levels in diabetes have been well established and thus the surrogate measures provide a solid basis (or strong biological plausibility) for the test being able to identify patients with diabetes. However, the link between *BRCA1/2* pathogenic variants and response to PARP inhibitors is not absolute, as other genes with pathogenic variants also influence the likelihood of response, and provides a weaker basis (or biological plausibility) for determining clinical test accuracy, and subsequently the clinical utility of the test.

Information to support the comparisons

The comparison of tests, particularly comparisons involving imperfect reference standards, or incomplete use of reference standards, may benefit from supplementary information. The additional information should seek to improve the understanding of the derived sensitivity / specificity or concordance. Additional information may be needed to explain why the proposed test results in

greater or fewer positive and/or negative test results. This information may be critical to determining that a reduction in sensitivity or specificity against an imperfect reference standard may be due to the proposed test having greater accuracy.

The following questions may be relevant to explore:

- How do the compared tests differ in terms of test parameters? For example, lower limits of detection or resolution.
- Is there a difference in method of classification of test results across tests? For example, this may occur if the tests use different thresholds for positivity, or access different databases for variant calling.
- Are there differences in what the tests can detect? For example, is the test designed to detect copy number variants in addition to single nucleotide variants, or non-FDG avid tumours as well as those that are FDG-avid?

If no additional information is provided to clearly indicate why the proposed test results differ from a reference standard or the comparator, the discordant test results should be regarded as false test results.

If an improvement in accuracy is expected, and can be supported by additional information, it is important to discuss whether newly positive patients (ie, the increase in true positives due to the proposed test) have the same spectrum of disease as the positives previously detected and if they will receive the same benefit from treatment. Discuss the possibility of overdiagnosis, which is the identification of a pathological lesion or state leading to a diagnosis in a patient, and there is no evidence that this state leads to reduced health outcomes, or no evidence that management decisions will benefit the patient. Note that the best evidence for determining whether a test is more accurate than a comparator test is direct evidence of test effectiveness ie showing the impact of the proposed test on patient health outcomes.

TG 11.3 Cross-sectional accuracy

Study designs

Cross-sectional cohort studies with consecutive or non-consecutive patients that meet the test population defined by the PICO and receive both the proposed test and any comparative test, measured against the reference standard provide a higher level of evidence than studies with a case-control study design (Merlin, Weston & Tooher 2009; NHMRC 2009). If the evidence base is large, there may be grounds for not including case-control studies in the analysis.

Individual study results

The presentation of individual study results is an important step prior to synthesis. Test statistics include sensitivity, specificity, positive and negative likelihood ratios, and the diagnostic odds ratio. These statistics can be derived from 2-by-2 tables populated during data extraction (Appendix 7).

While the positive or negative predictive value of a test can be derived from a study with a 2-by-2 table, the estimate will only be accurate for the population included in that study ie it will be affected by the prevalence of the condition in that study population and the test samples derived from that population. The test sample prevalence may not be applicable to the target population. It is preferable to derive positive and negative predictive values from the pooled sensitivity and specificity values (if pooling is appropriate, after meta-analysis), and an applicable estimate (or range of estimates) for prevalence.

Briefly tabulate the sensitivity and specificity, as well as the sample prevalence, for each included study. The methods for calculating these test statistics (as well as other test accuracy measures) are provided in Appendix 7.

Assessing the transitivity and applicability of the included studies

Studies included in a meta-analysis of test accuracy should be transitive, although it is recognised that evidence on tests is often heterogeneous. The included studies should be assessed to determine whether there are any key differences between them that may affect test accuracy. These differences may include assay characteristics, sample handling, differences in interpreting the results, thresholds for determining positive results, and biological characteristics of the test population.

These characteristics should be assessed for their applicability to the Australian diagnostic setting. If certain population subgroups are not relevant for either inclusion in the testing population, or for laboratory diagnosis, these subgroups should be omitted from the analysis.

Where any characteristics across studies are expected to affect test accuracy, present separate meta-analyses or subgroups within meta-analyses. Where the effect on test accuracy is uncertain, or a threshold effect (see meta-analysis section below) is predicted, a subgroup analysis should be undertaken. Threshold effects should be further analysed by including covariates in bivariate models or using a hierarchical summary receiver operating characteristic (HSROC) curve. Further explanation of meta-analysis methodology is given below. Where a meta-analysis cannot be undertaken, consider the heterogeneity in the evidence base – and likely impact on test accuracy - in the narrative synthesis.

In addition to identifying differences across studies, identify characteristics of the included studies that are different to the Australian setting. Any concerns relating to applicability should be discussed during the interpretation of the results. Studies that are clearly not applicable should be identified as such.

Meta-analysis of test accuracy studies

While most measures of test accuracy can be pooled, the preferred approach for an assessment is to pool only sensitivity and specificity, and to derive the other measures of test accuracy from the pooled estimates of sensitivity and specificity. This is for two key reasons:

- Test accuracy measures that vary across increasing/decreasing prevalence rates should not be pooled; and,
- There are established bivariate meta-analysis methods for correcting for the correlation between sensitivity and specificity.

A bivariate model accounts for the correlation between sensitivity and specificity and is preferred when summary point estimates are sought (see Appendix 7). However, a minimum of four studies are required for this type of meta-analysis. In some cases the bivariate models do not converge, especially if there are few studies or several zero cells in the 2-by-2 table (Takwoingi et al. 2017). If this occurs, separate univariate binomial meta-analyses for sensitivity and specificity can be used with justification and a discussion of the uncertainties in the approach. In some cases, it may be appropriate to use univariate models to pool the diagnostic odds ratio, which is an estimate that incorporates both sensitivity and specificity.

For some tests, there is no universally agreed threshold (or cutpoint) for determining a positive result and some studies may use several different thresholds. If there are a mixture of thresholds used across and/or within studies, and there is no clear reason to limit the analysis to a single threshold, it may be appropriate to present a HSROC. HSROC curves characterise the relationship between sensitivity and

specificity across the included thresholds and this graphical representation of the included studies provides an easy way to examine both the threshold effect and between-study heterogeneity.

Heterogeneity between the studies included in a meta-analysis should be explored as described in Appendix 6. The results between studies would be expected to differ due to chance alone, a consequence of differences in the included samples taken from the entire theoretical population. Statistical heterogeneity is identified when there is more variability than expected, and is a frequent occurrence with test accuracy studies, partly due to the impact of the test thresholds. Thus, it is useful to determine the proportion of the variability that could be attributed to the threshold effect and to chance.

Following a meta-analysis, it is common to present an assessment of publication bias. Publication bias occurs when the outcome of an experiment or research study influences the decision whether to publish. An assessment of publication bias is relevant for the GRADE approach for assessing the certainty of the evidence (Appendix 4).

For a more detailed discussion on the methods used for meta-analysis of test accuracy studies see Appendix 7.

Presentation of test accuracy evidence

The key results presented in this section will depend upon the types of comparisons that are required to best describe the true positives, true negatives, false positives and false negatives from the proposed test. The assessment of test accuracy evidence should consider the following points:

1. Describe the literature search for test accuracy studies. Assess for a risk of bias and extract study characteristics.
2. Present the results for individual studies in a table. Provide relevant test statistics, including sensitivity, specificity and the prevalence of the disease or biomarker in the study (if estimable).
3. Discuss the transitivity of the included studies and justify the separate presentation of any studies based on transitivity concerns.
4. Describe the approach for meta-analysis or narrative synthesis of the data and discuss possible reasons for heterogeneity of results.
5. Indicate whether there is a specific test threshold which should be used to determine test positivity/negativity (if applicable) ie where the test sensitivity and/or specificity is highest for achieving the purpose of the test.
6. Apply the prevalence rate of the disease in the PICO population, as determined in TG 11.8, to the pooled estimate of sensitivity and specificity to generate other test statistics (PPV, NPV etc) (see Appendix 7).
7. Interpret the results. The interpretation should include a description of the comparison, and an explanation of why the comparison is important for the interpretation of subsequent steps in a linked evidence approach. If the proposed test is identified to have a different accuracy to that of the comparator, reference standard or clinical utility standard, discuss whether this represents an improvement in accuracy or a loss of accuracy and provide supplementary data (e.g. test characteristics such as scoring algorithms, thresholds, read depth) to justify the judgement.
8. Explain the implications of changes in test accuracy on the management of patients (change in management evidence), and the likely impact on patient outcomes.
9. Present a summary of the quality of the body of evidence (GRADE).
10. Repeat this approach for all necessary comparisons (proposed test vs reference standard, proposed test vs clinical utility standard, proposed test vs comparator), and provide a justification for why each comparison is required.

In a separate section, describe the search for sources of prevalence estimates, and present a range of estimates. Nominate the most applicable estimate of prevalence and provide justification.

Applicability of results to subsequent linked evidence

A new test that detects additional cases of apparent disease can create uncertainty about whether these additional cases should be classified and treated in the same way as current practice (Glasziou, P., Irwig & Deeks 2008). For example, a new test for a suspected disease may widen the spectrum of patients considered to have the diseases compared with the reference standard test, and the correlation of the findings of the test with the eventual clinical course may be poor. This may indicate that the additional diagnosed cases are either at lower risk, with the treatment having a smaller beneficial effect, or that some patients have been incorrectly diagnosed (false positive, overdiagnosis) with the new test and may have received unnecessary treatment. Therefore, care must be taken when assessing the health outcomes for these newly diagnosed patients.

TG 11.4 Longitudinal accuracy

'Longitudinal accuracy' is when a test is performed for the purposes of determining a future health state. The accuracy of this prediction is measured against a "reference standard", which is the health outcome of interest at a later time point (e.g. length of survival, or response to treatment, etc.).

This section describes how to approach the assessment of the longitudinal accuracy of a test. This is required when the test is being performed to:

- Establish a predisposition for a disease
- Estimate a prognosis to predict a patient's clinical course
- Predict a response to treatment to identify suitability for that treatment
- Measure an early effect on a surrogate outcome to predict a later effect on more clinically relevant outcomes

For establishing this form of test accuracy, longitudinal data must be used. The hierarchy of informative study types (in the absence of evidence of direct test impact on health outcomes) is shown in the NHMRC levels of evidence for prognosis, available on the NHMRC website. MSAC will be most influenced by the results of rigorous prospective data.

Key measures of the effect generated out of the cited literature may be similar to cross-sectional accuracy (i.e. if the outcome of interest is a dichotomous variable, results may be presented in terms of sensitivity/specificity; see TG 11.3). However, longitudinal accuracy data may also be presented as relative risks, odds ratios, etiologic risk (population attributable risk), logistic regression measures, interaction terms, or, they may incorporate data over time through the use of Kaplan Meier curves and hazard ratios. It is important to ensure that some measure of the *incremental* value of the proposed prognostic/predictive test is provided (i.e. what additional value is derived from the proposed test, over and above the information that would be derived in the absence of the test).

The reference standard, or clinical endpoint of interest, must be clearly defined, including the time period of follow-up. This is a key issue as the time interval and intervening variables, such as treatments, can impact on the accuracy of the predictions. Clarify what thresholds are used to determine risk classifications, and whether they reflect the thresholds that would be used in Australian clinical practice.

Predisposition testing

Predisposition testing provides information on the likelihood of an asymptomatic person developing disease in the future. An example of predisposition testing is cascade testing of family members of

someone with breast cancer and a *BRCA1/2* pathogenic variant to determine their risk of developing breast or ovarian cancer (this is a form of targeted screening, targeted to those at high risk).

Most predisposition tests assessed by MSAC, as of the time of writing, have been for specific conditions; however, advancements in genetic testing have resulted in the possibility of people being screened for a wide variety of conditions at once, and the assessment of these panel tests will require use of the exemplar/facilitated approach (See TG 5.2).

If the data are presented without a time-to-event element (i.e. a comparison of test results and clinical outcome without time specified, or at one time point), then present the positive predictive value and negative predictive value of the test (as per TG 11.3).

When evaluating a predisposition test, it is important to make sure that the evidence is derived from a population with the appropriate prevalence of disease, or that the applicability of evidence from another population is considered. The positive predictive value of the test (the probability that a person with a positive test result will develop disease) depends on the prevalence of the disease in the population and the penetrance of the biomarker. For example, data derived from universal screening will not be applicable to targeted screening, as the proportion of false positives to true positives may vary widely. For a more in-depth discussion on the influence of prevalence on the positive and negative predictive values, see TG 11.3.

If data on the accuracy of the test for determining the biomarker is identified, without consideration of how well it predicts disease, then a separate step of considering the penetrance of the pathogenic variant will be required.

When data are presented in a time-to-event format, then Kaplan Meier curves and hazard ratios should be presented.

The hazard ratio is calculated as follows:

$$Hazard\ ratio = \frac{Hazard\ of\ disease\ occurring\ in\ those\ who\ test\ positive}{Hazard\ of\ disease\ occurring\ in\ those\ who\ test\ negative}$$

A hazard ratio of 1 equals a lack of association (i.e. there is no relationship between the test result and likelihood of disease). A hazard ratio of greater than 1 suggests an increased risk, while a hazard ratio below 1 suggests a decreased risk. The term “hazard” refers to the probability that an individual will have a particular event by a particular time. The hazard may be mapped as a Kaplan-Meier curve, showing the proportion of participants remaining event-free over time. The hazard ratio is a measure of the separation between two Kaplan-Meier curves. For example, a hazard ratio of 0.70 means that those who are test positive have a 30% risk reduction of having an event compared to those who are test negative. Precision and uncertainty around the result should be indicated (e.g. through confidence intervals).

For many conditions, the appropriate comparator to the predictive test will be existing risk assessment approaches. For example, risk factors for cardiovascular disease include age, sex, smoking, hypertension, diabetes and lipid levels, and risk stratification may be based on these. If a new predisposition test for the polymorphism on chromosome 9p21 was to be proposed, then the incremental benefit of this new test should be considered. This would determine if there is any benefit of the new test over using the existing risk stratification (Jonas et al. 2012), as it cannot be assumed that the new information that a test provides is of value in the overall risk assessment.

Screening

Screening is similar to predisposition testing but aims to detect pre-clinical signs of disease (such as breast cancer or colorectal cancer). Universal screening is discussed in TG 15.1. Surveillance of pre-clinical signs of disease in someone at high risk (such as those with a defined predisposition) could be assessed as per a diagnostic test, or monitoring test.

Testing to determine prognosis

A prognostic test provides information about a patient's prognosis, without specific consideration of downstream therapies chosen (i.e. it characterises the natural history of the disease).

Many different patient characteristics may provide useful information for determining their prognosis. Prognostic information may be considered to have value inherently for the sake of the knowledge itself (See Technical Guidance 28 on other utility), as well for the way that it influences the downstream healthcare that people receive. It is important that this information is accurate, so that personal and clinical decisions made based on it, are informed correctly. Prognostic tests are developed to assist (not replace) clinical judgement regarding the likely future health outcomes of the patient, and enhance patient decision making (Steyerberg et al. 2013).

Many prognostic tests combine multiple variables in order to predict the risk of experiencing a specific endpoint within a specific time period. This formal combination of multiple factors is called a prognostic algorithm. If the way in which factors are chosen and combined is unclear, the prognostic algorithm could be considered a 'black box algorithm'. For additional considerations for black box algorithms, see TG 15.3.

In order to show the relationship between the proposed test result and the endpoint, present the univariate analyses with the estimated effect (e.g. hazard ratio and survival probability), to demonstrate the prognostic strength, before allowance is made for other variables. If available, the same information should be provided for the comparator. For the effect of a marker on time-to-event outcomes, a Kaplan-Meier plot is recommended (Figure 15), showing the curve for each category (Altman, D. G. et al. 2012).

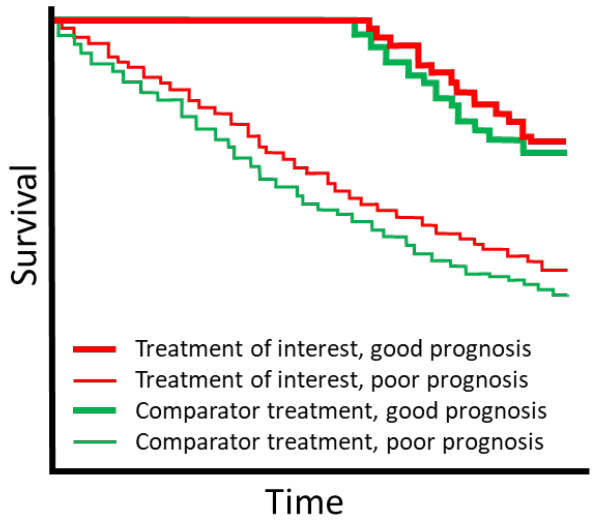


Figure 15 Kaplan-Meier curve showing time to event outcomes for patients with good versus poor prognosis for the treatment of interest and the comparator

For cases where there are two categories (i.e. patients can be classified as test positive or negative), the hazard ratio is calculated as presented in Predisposition Testing, above.

In order to estimate the incremental value of the prognostic marker, multivariate analyses are required, demonstrating the additional information gained by the proposed test over and above existing markers that are likely to be the comparator.

If measures such as odds ratios or relative risks are used, consideration should be given that odds ratios or relative risks that are traditionally considered to be large for association studies, may not be adequate for discriminating between people that are likely to develop the outcome of interest, and those who do not (Pepe et al. 2004). For example, if the absolute risk of developing a disease of interest is only 3 in 1000 people, a relative risk of 3.0 (considered large in epidemiological research), would only mean that people with the particular marker had a 9 in 1000 risk of developing the disease. Information on the absolute risk of the disease should therefore be obtained to put the relative measures of association in context. Relative measures should be used with caution as a means of risk classifying individuals (Pepe et al. 2004).

Testing to predict treatment effect

A predictive test provides information on the expected effect of a therapeutic technology (e.g. a test for the *HER2* gene to predict response to breast cancer treatment). This may result in 'personalised medicine', allowing the therapeutic technology to be restricted to those who are most likely to benefit, and avoiding the harms associated with the intervention in those unlikely to benefit. If the predictive test is co-dependent with a drug being submitted to the PBAC see Appendix 8 on using a co-dependent technology assessment approach.

As with any investigative technology, the utility of a predictive test is best proven through the use of direct from test to health outcomes evidence, comparing health outcomes of patients whose management is guided by the predictive test, as compared to those who management is guided by the comparative test strategy (which could include treatment without a test) (see Technical Guidance 10). However, the evidence on predictive tests is rarely generated in this manner. More commonly, evidence is generated to determine the effectiveness of a treatment, and biomarker status is retrospectively determined (and thus, not used to determine treatment condition).

For a dichotomous outcome (or a continuous measure dichotomised by choosing a cutpoint), the evidence required in order to distinguish whether the test is predictive of treatment response, or prognostic, it is important for the comparison to have all four arms of evidence (Figure 16). If a test is prognostic, then the two arms that receive the control treatment will differ. If the test is predictive, then the relative difference (OR, RR, HR) in health outcomes between the test positive and test negative in the treatment arm will vary from the relative difference for health outcomes between the test positives and test negatives in the control arm.

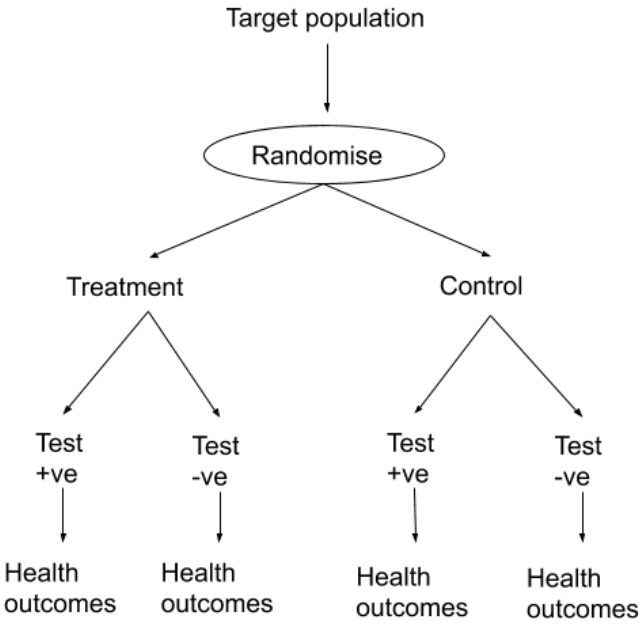


Figure 16 Predictive evidence trial (test information not used to derive treatment condition) (Preferred option 2 for assessing predictive tests)

In Figure 17, the starting populations may have been similar, but the evidence is uninformative in distinguishing between whether the test is predictive or prognostic. Rather, it demonstrates whether targeting treatment according to a test result is better than providing treatment without guidance from a test.

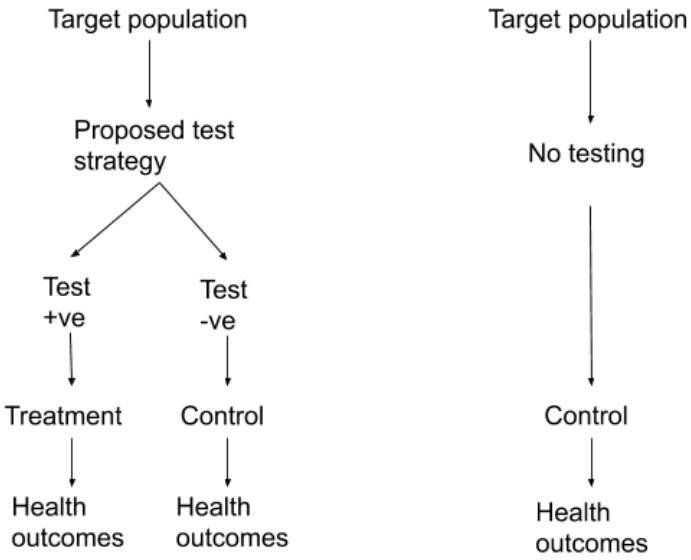


Figure 17 Less informative predictive evidence comparison

For continuous markers (for which statistical methods are limited), it is suggested that marker-by-treatment curves be presented if available, to illustrate how particular test results correspond to different health outcomes, depending on what treatment is chosen. See Janes et al (2011) for details (Janes et al. 2011). This approach may be informative for determining a clinically meaningful threshold for test positivity (Figure 18).

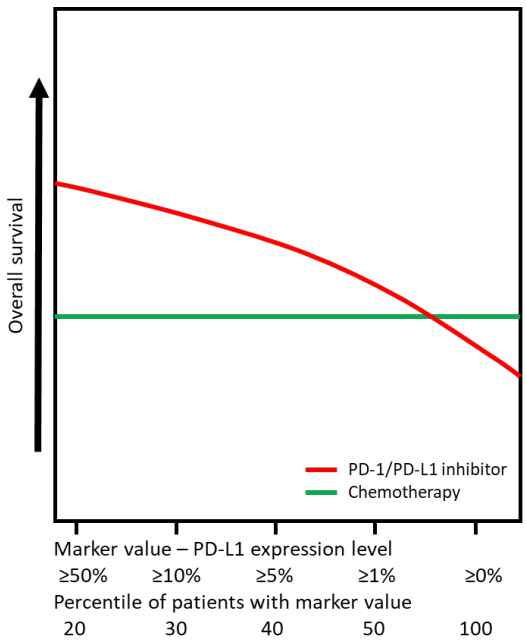


Figure 18 A marker-by-treatment curve showing response to second-line PD-1/PD-L1 inhibitors and chemotherapy in patients with NSCLC according to PD-L1 expression level

Patients with NSCLC who have higher levels of PD-L1 expression are likely to have a greater benefit from treatment with PD-1/PD-L1 inhibitors than from chemotherapy. PD-L1 = programmed death-ligand 1, with its receptor, PD-1: programmed cell death protein.

The interpretation of the evidence depends largely on whether the subgroup analyses are performed on subgroups determined *a priori*, or whether they are conducted *post hoc* (Altman, D. G. et al. 2012). MSACs strong preference is for subgroups determined *a priori* based on pre-specified classifications, to reduce the likelihood that the finding occurred by chance. If the analyses are performed *post hoc* then they are considered to be hypothesis generating, and therefore require validation using an independent sample.

One method often used to determine whether there is an interaction between treatment effect and subgroup, is to present results separately for subgroups, and erroneously conclude if there is a significant treatment effect in one subgroup, and a non-significant treatment effect in the other, that the effect differs by subgroup (Altman, D. G. et al. 2012). A test of interaction should ideally have been performed to rigorously assess whether the effects are different by subgroup, or whether the difference in significance level is due to one subgroup being underpowered (too small to allow detection of a significant difference). The magnitude of the interaction does not describe how useful a marker is for patients (Janes et al. 2011). Therefore, if the test of interaction is significant, then further evaluation may be required to determine the nature of the interaction, i.e. whether the effects are in the opposite direction; or if the effects are in the same direction, but a different magnitude (Altman, D. G. et al. 2012).

For continuous variables, categorisation is a common approach, but is highly dependent on the thresholds used between categories. The checklist by Altman et al (2012) suggests that a preferred approach is to use the multivariable fractional polynomial interaction approach, which avoids specifying the thresholds (Altman, D. G. et al. 2012). It allows interaction terms to be investigated between a binary and continuous variable, with or without adjustments for other variables (Altman,

D. G. et al. 2012). Alternatively, the subpopulation treatment effect pattern plot, as described by Bonetti et al (Bonetti & Gelber 2000), may be used (Altman, D. G. et al. 2012).

In order to determine whether there is a treatment effect modifier, it must be determined whether response to treatment (vs control) varies by the test result (i.e. will the test select a subgroup who do or do not respond to treatment more than those who are test negative).

If it is established that a test is not prognostic (no difference between the new test and standard care in patients' health outcomes), then in order to determine whether a test can predict response to treatment, a comparison of the risk of an event happening for those receiving the treatment of interest in those test positive and negative is of relevance. This is commonly represented by a hazard ratio for time to event outcomes, but may also be presented as a relative risk, odds ratio, or another relative outcome measure.

Distinguishing between whether a biomarker determines prognosis or is a treatment effect modifier may not be possible unless a study provides health outcomes for those who are biomarker positive and negative, and those who have the treatment versus control (i.e. all four arms) (Figure 19).

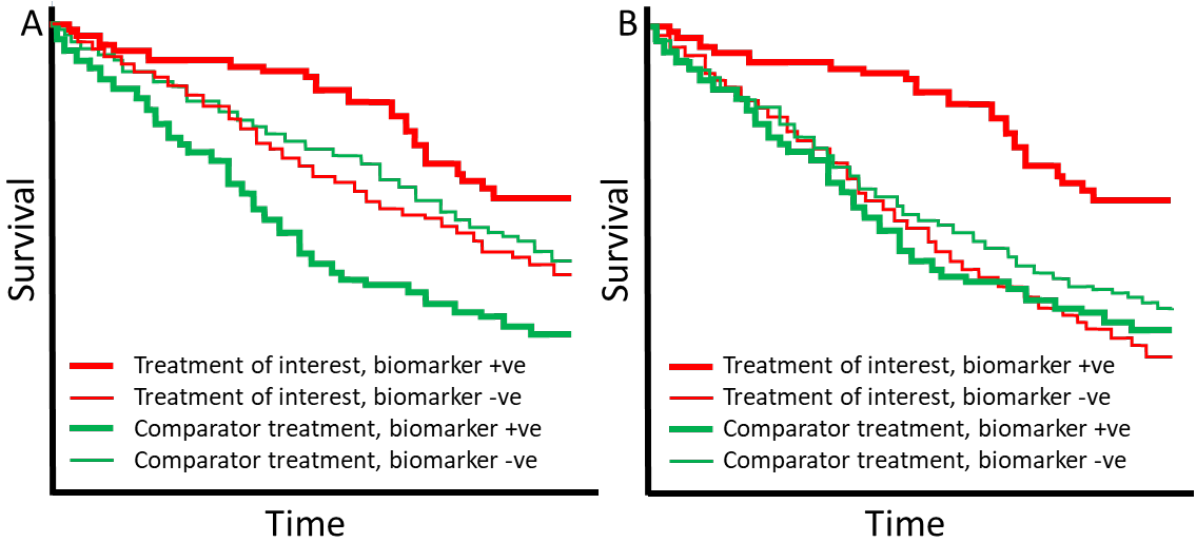


Figure 19 Kaplan-Meier curve showing time to event outcomes for patients with and without a biomarker

(A) The biomarker is both is predictive of response to the treatment of interest and prognostic. (B) The biomarker is not prognostic but is predictive or response to the treatment of interest.

Surrogate outcome purpose:

Surrogate outcomes may sometimes be considered as predicting what final outcome patients are likely to have. In this manner, the surrogate outcome can be assessed for how accurately it predicts the final outcome. There needs to be a clear association between the two outcomes, and a biological rationale for how the two relate (Moons 2010).

If a test is performed to determine a surrogate time to event outcome, the hazard ratio for the accuracy of the prediction of the final outcome may be calculated as follows:

$$Hazard\ Ratio = \frac{Hazard\ of\ final\ outcome\ occurring\ in\ those\ with\ surrogate\ outcome}{Hazard\ of final\ outcome\ occurring\ in\ those\ without\ surrogate\ outcome}$$

Relative outcome measures may also include relative risks and odds ratios, particularly if the time to event outcome is presented as numbers with or without an event at a particular time point.

Assessing the risk of bias

It is particularly important for the risk of bias to be considered for the findings in the studies included on predictive testing. Publication bias is likely to be a major concern for prognostic or predictive studies, as these studies are often performed using retrospective analyses of existing clinical databases, or as *post hoc* analyses of trials. As such, there will not be any indication that the study has been performed until it has been published (Altman, D. G. et al. 2012). Selective publication of prognostic studies is likely to result in larger effects seen in smaller studies and many 'false-positive' studies which have occurred through chance (Riley, Sauerbrei & Altman 2009).

Studies should be assessed to determine whether selective outcome reporting has occurred. Trials will often assess the principle outcomes of time to death (overall survival) and time to recurrence (disease-free survival), however, articles reporting on the studies will often only present one of the outcomes (Altman, D. G. et al. 2012). Trial registries could be checked to see whether additional results are available. Another area of concern is if the studies only report unadjusted results, as these are generally larger in magnitude than adjusted results and confounded by covariates. There are also concerns regarding the risk of selective reporting in some subgroups, so it should be made clear whether the subgroups were pre-planned or not (Altman, D. G. et al. 2012). Likewise, the thresholds used in the validation studies must be consistent with what are likely to be used in the target Australian setting, as studies which select the 'optimal cutpoint' retrospectively introduce considerable bias (Riley, Sauerbrei & Altman 2009).

The time-lag between the prognostic testing and clinically important events should be assessed to determine whether it is long enough, or whether participants in the studies are beyond the age that clinical expression is likely (Jonas et al. 2012).

Generalisability of the evidence

It is important to determine if the findings from a proposed test are generalisable to a different set of patients.

If the proposed test is accurate in patients who were not part of the development cohort (which was used to generate the prognostic algorithm) but are from an identical population (validation set) as the development cohort, then this would mean the test is reproducible. Another important concept is whether the test is accurate in a population which differs from the development cohort, or where the methods used for the test differ from those used in development (i.e. whether the test is transportable) (Justice, Covinsky & Berlin 1999). The transportability of the proposed test may be threatened if there are differences in time period, geography, methods used, spectrum of patients included, and follow-up interval (Justice, Covinsky & Berlin 1999).

TG 11.5 Concordance

Concordance analyses are useful to determine the characteristics of the patients who receive different test results from the proposed test strategy, versus the comparative test strategy, or versus the clinical utility standard. A clinical utility standard is a test that is used to detect the condition (biomarker or disease) in a clinical trial that generates evidence of the effect of testing on health outcomes.

However, if the accuracy of the test has not been proven, it is not considered to be a reference standard.

When the proposed test is evaluated by comparison to a clinical utility standard, sensitivity and specificity are not appropriate measures to describe the comparative results. Instead, measures of test concordance are calculated (Table 30). The positive percent agreement (PPA) and negative percent agreement (NPA), which are calculated using the same numerical calculations used to estimate sensitivity and specificity (see Appendix 7), should be reported. When comparing the proposed test to a clinical utility standard, a discussion of the discordant results as false positive or false negative compared to the clinical utility standard should be included. The exception would be if there was compelling evidence that the proposed test is more accurate than the clinical utility standard and the discordant results are more likely to be true with respect to the proposed test (although this would be difficult to justify without direct test to health outcomes evidence for the proposed test).

Concordance of a proposed test and a comparator test is usually measured using overall percent agreement (OPA) and/or Cohen's Kappa coefficient (κ). It is important to note that "agreement" does not mean "correctness." Thus, the two tests could agree and both be wrong, and if the two tests disagree it is unknown as to which test is right.

The overall percent agreement should not be reported in isolation, as it can be misleading. The overall percent agreement can be good, even if either the PPA or NPA is very low. Thus, the PPA and NPA should also be reported.

Cohen's kappa coefficient is generally thought to be a more robust measure than simple percent agreement calculation, as it takes into account the possibility of the agreement occurring by chance. However, kappa coefficients are difficult to interpret. Altman (1991) provided an interpretation of the kappa coefficient shown in Table 9.

Table 9 Interpretation of Cohen's kappa coefficient

Cohen's kappa coefficient	Interpretation by Altman (1991)
0.01–0.20	Poor agreement
0.21– 0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Good agreement
0.81–0.99	Very good agreement

Altman DG (1991) Practical statistics for medical research. London: Chapman and Hall.

Companion testing

Concordance is likely to be important for determining equivalence between two or more tests that could be used interchangeably in the Australian clinical or diagnostic setting to assess the same disease outcome or biomarker.

One example would be the different commercially available PD-L1 tests used as companion diagnostics for determining eligibility for targeted immunotherapy. In this case, the tests were originally designed to measure the biomarker under different circumstances. They were initially optimised to measure PD-L1 expression on different cell types (tumour-infiltrating immune cells versus tumour cells), using different scoring algorithms (total proportional score versus combined positivity score versus inflammatory score) and different cell locations (cytoplasmic versus cell membrane). Thus, the level of concordance between these tests for a specific application, and the

downstream health outcomes from treatment with the targeted therapy would be important in order to determine the clinical utility of these companion tests.

TG 11.6 Cascade testing for inheritable diseases

Cascade tests are usually a modified use of the broader genetic test used to identify the pathogenic variant causing disease in the index case. Thus, they are used to identify one specific genetic variant in high-risk first- and/or second-degree relatives. The tests are usually highly accurate and have no comparator or reference standard. Studies reporting the accuracy of a cascade test usually only report the diagnostic yield of the proposed test; that is, the number or proportion of patients who had a positive test result out of the number who were tested.

The extent of cascade testing required is typically dependent upon the Mendelian inheritance of the pathogenic variant in question. For example, for autosomal recessive conditions, the inheritance rules suggest that approximately 50% of first-degree relatives and 25% of second-degree relatives are expected to inherit the pathogenic variant of interest. Inheritance may be affected by variable penetrance and expressivity of pathogenic variants.

Other outcomes, such as the proportion who refuse testing, should also be reported as they provide useful information for determining the clinical utility of the cascade test.

TG 11.7 Test reliability

In some cases, determining the reliability of the test may be important when determining test accuracy.

The term reliability (which is analogous to the concept of precision) refers to the agreement between different operators or instruments applying the same investigative medical service. Reliability is sometimes referred to as reproducibility or repeatability.

The reliability of a test result depends on the variability of the same person or instrument making the test score on two different occasions (intra-observer or intra-instrument variability/agreement) and the variability between different observers or instruments (inter-observer or inter-instrument variability/agreement). Reliability might be further affected by factors such as the number of observers, tissue storage and processing, and so on. An investigative medical service that has poor reliability cannot have good test accuracy. On the other hand, good reliability does not assure good test accuracy.

Inter-laboratory variability/agreement should also be considered. However, any variability between laboratories should be mitigated (or controlled) by an appropriate National Association of Testing Authority-approved quality assurance program.

Kappa statistics are the method of choice in an extended assessment of reliability. The kappa value for the intra-observer and inter-observer agreement corrected by chance can be interpreted as shown in Table 9.

Other reliability measures would include the rate or proportion of failed and equivocal test results across the included studies. In addition, other reliability measures may apply to specific tests. For example, next generation sequencing (NGS) tests should include:

- Minimum read depth of included genes/regions and how this affects identification of variants
- Test limits of detection for different types of sequence variants

- Identification of sequence regions or variant types that the test cannot detect with the intended accuracy and precision

TG 11.8 Prevalence of the disease or biomarker in the PICO population

Prevalence refers to the proportion of individuals in a population who have a disease or condition, and includes both new and old cases. Thus, prevalence is the product of incidence of new disease and the duration of the disease (Rotily & Roze 2013).

The approach for identifying an appropriate estimate of prevalence depends on the requirements of the decision problem. The prevalence of the biomarker is relevant if test accuracy is measured as the ability of the test to identify the biomarker (using a non-clinical reference standard). Whereas, the prevalence of the disease is relevant if the test accuracy is measured as the ability of the test to identify the disease (using a clinical reference standard). The prevalence of the biomarker or disease in the testing population will always be required for estimating the use of the investigative medical service, and in the economic analysis. Thus, even if the sensitivity and specificity of the test cannot be determined, the diagnostic yield, which is equivalent to the prevalence, is an important input in the economic analysis.

The most applicable sources for prevalence estimates of disease in the Australian setting may be administrative databases, registries or surveillance data. Some examples include:

- Australian Institute of Health and Welfare (AIHW)
- Specific disease registries

However, it may not be possible to convert this data to the prevalence of the disease in the testing population specified in the PICO. This population may consist of high-risk symptomatic patients, where the likelihood of having the disease would be much higher than in the general population.

Carefully assess the inclusion and exclusion criteria to ensure that the study population has not been enriched, and the entry into the study reflects the proposed use of the test in the target testing population. Studies that do not report how subjects were selected, or that adopt a design which is not suited to estimate prevalence, should be excluded.

Most likely, the primary source for the estimate of the prevalence rate of the biomarker or disease would be the studies that provided the test accuracy data. The key considerations in identifying which of these studies are appropriate for inclusion in estimating the prevalence relate to the applicability of the source to the target testing population. The selection of the studies should consider whether the study population is derived from:

- An Australian population with similar disease characteristics.
- A non-Australian population with similar disease characteristics.
- The patients are at the same point in the clinical management algorithm (that is, the same prior tests or assessments have been performed). If triaging of patients to the test differs, then the prevalence of the biomarker/disease may differ from the testing populations.
- The baseline characteristics of the populations are similar. In the case of genetic testing, gender and ethnicity may be particularly relevant.
- Similar risk factors for some somatic pathogenic variants or disease biomarkers. For example, different oncogenes may be more prevalent in populations with different levels of cigarette consumption.
- An adequately sized study population to provide a reasonably robust estimate of prevalence.
- Prevalence of the biomarker or the disease should be measured using the reference standard or another appropriate test that is in common use in clinical practice.

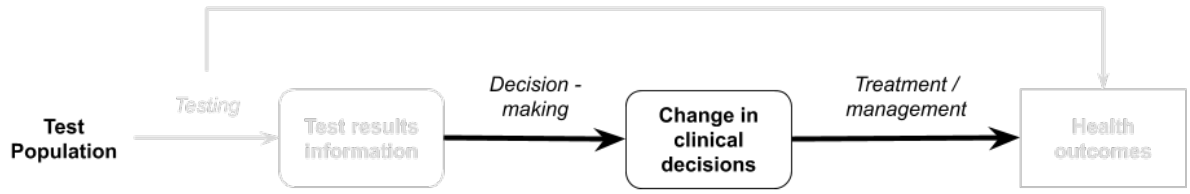
- Tests have become more sensitive over time, and definitions of disease or disease stage have changed over time. Therefore, greater weight should be placed on recent studies.

Due to the inherent heterogeneity or variability between test accuracy studies, determining the median prevalence rate and the range may be more appropriate than determining the mean and standard deviation. If pooling of prevalence studies is required, meta-analysis methods using binomial distributions and/or transformations to approximate normal distributions (e.g. Logit) would be appropriate.

In the absence of studies that completely agree with the target testing population described in the PICO, the most closely aligned studies should be selected for the prevalence rate estimate. It may be necessary to identify any concerns due to the inclusion or exclusion of certain patient subgroups in these studies. This is particularly relevant, if the prevalence of the biomarker/disease is expected to vary from that expected in the targeted testing population.

Generally, there is uncertainty around the most appropriate prevalence rate estimate, and a range of estimates are usually applied in sensitivity analyses of the NPV and PPV estimates and in the economic model.

Technical Guidance 12
Linked evidence - change in management



TG 12.1 Purpose of the guidance

An impact of a test on health outcomes can only be achieved if the interpretation of the test results leads to a change in the management for a test subject. This guidance describes the assessment of change in management following the use of a test. Although the guidance section is labelled "linked evidence of change in management", evidence for change in management may also be relevant to augment direct from test to health outcomes evidence, particularly where the change in management in the direct from test to health outcomes study did not reflect current clinical practice. The TG section discusses:

- A definition of change in management evidence
- How to assess change in management evidence
- Considerations relating to the assessment of change in management evidence
- How to assess the applicability of change in management evidence
- A suggested approach to presenting change in management evidence for an investigative medical service

TG 12.2 Change in management evidence

A biomarker may be used to diagnose a condition, measure disease severity, measure response to treatment, monitor patients over time or predict the prognosis of the patient (Doust, J 2010). The impact that the biomarker has on clinical utility of a test (the net benefit or harm in regard to health outcomes) depends on the series of actions and reactions that happen as a consequence of the test result. The variety of different uses of biomarkers means that the indirect impact on patient-relevant health outcomes needs flexibility in the way it is assessed. The ability for a test to change the clinical management of a patient depends on many factors, such as (but not limited to):

- whether treatments are available for a disease identified, or whether treatments differ for differential diagnoses;
- whether the test provides a different result from the comparator (change in diagnosis/prognosis etc, or whether it changes the spectrum of patients treated);
- whether clinicians trust the test result sufficiently to incorporate it into their 'diagnostic thinking' and treatment recommendations;
- whether patients are willing to receive the treatment recommended; or
- whether it influences patient's adherence to recommendations.

Change in management evidence may be required to determine what subsequent interventions are received following a test, or to satisfy that there is no change in management (if the proposed test is claimed to be non-inferior to the comparator test). Evidence to support change in management must

incorporate the management decisions for each test result (e.g. positive and negative, high or moderate or low risk etc). Change in management may also include a time component. For example, the availability of a new test may result in the same management decisions for patients, but at an earlier time point. In this circumstance, the comparative management strategies would be early versus late intervention.

The nature of the change in management may differ depending on test purpose. For example:

- A diagnostic test may result in the decision to use a treatment for a disorder, rather than an alternative treatment or no intervention;
- A diagnostic test may also result in earlier treatment compared with waiting for a clinical diagnosis;
- A staging test may determine whether radiotherapy is required in addition to surgery;
- A prognostic test may determine whether a patient is likely to have disease recurrence or not, and whether adjuvant chemotherapy should be considered or more intensive monitoring;
- A predictive test may determine whether a treatment is likely to be beneficial for the patient and should be initiated; and
- A predisposition test may influence the rate of lifestyle modifications or adherence to surveillance.

Change in management involves several sequential steps. Evidence may represent how a test result is interpreted (diagnostic thinking), what recommendations are made, and what is adopted by patients (ie, the actual treatment received). It is important to discuss the limitations of evidence based on earlier steps of change in management as changes in diagnostic thinking, or recommended treatment, might differ from the actual treatment received.

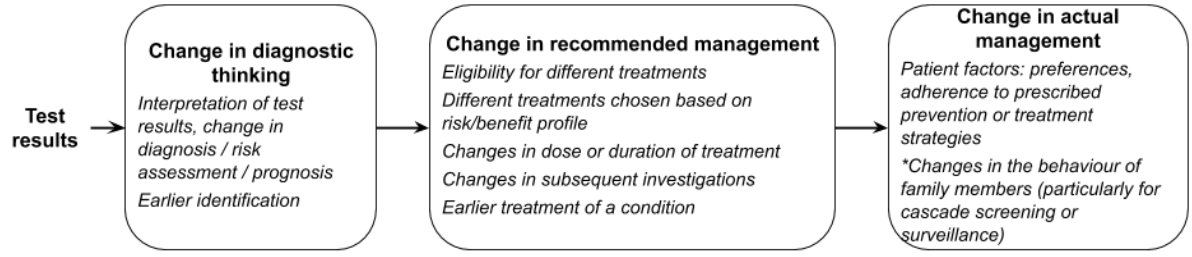


Figure 20 Change in management components from test results to actual management undertaken

TG 12.3 Change in management study designs

Change in management studies may either be experimental or observational. Studies which report actual management provide more directly relevant information than those which report hypothetical planned management. Therefore, a randomised trial, which reports actual management following the use of the comparative test strategy versus the proposed test strategy (Figure 21 – A) contains less risk of bias than a within-patient comparison of pre-test management plan and post-test management received. Randomised trials are suitable for all test types (i.e. replacement tests, add-on tests, triage tests etc), unless there is no longer clinical uncertainty that the test is beneficial (i.e. if the test has become part of "best practice"), and can report on other outcomes such as patient acceptability and safety of the tests (Staub et al. 2012).

Despite the advantages of randomised trials, the most common study design for change in management studies is the observational 'diagnostic before-after' study (Figure 21 – B). These study

designs are useful when the test is an add-on to an existing test strategy, but not if the test is a replacement or triage test. That is, it is suitable if the existing test strategy matches the 'before' component, and the proposed test strategy matches the 'after' component with the addition of the new test, which cannot be the case if the proposed test is to be added before the existing test strategy. These studies are subject to bias, as the clinician may not use the same amount of caution in determining the pre-test management plan if they know they will receive the subsequent proposed test result (Staub et al. 2012). It should be made explicit whether the management plan is made prospectively by the clinician, or retrospectively, based on case notes. Measuring the concordance between the post-test management plan and the actual management received may provide an indication on how planned management is put into practice (Staub et al. 2012).

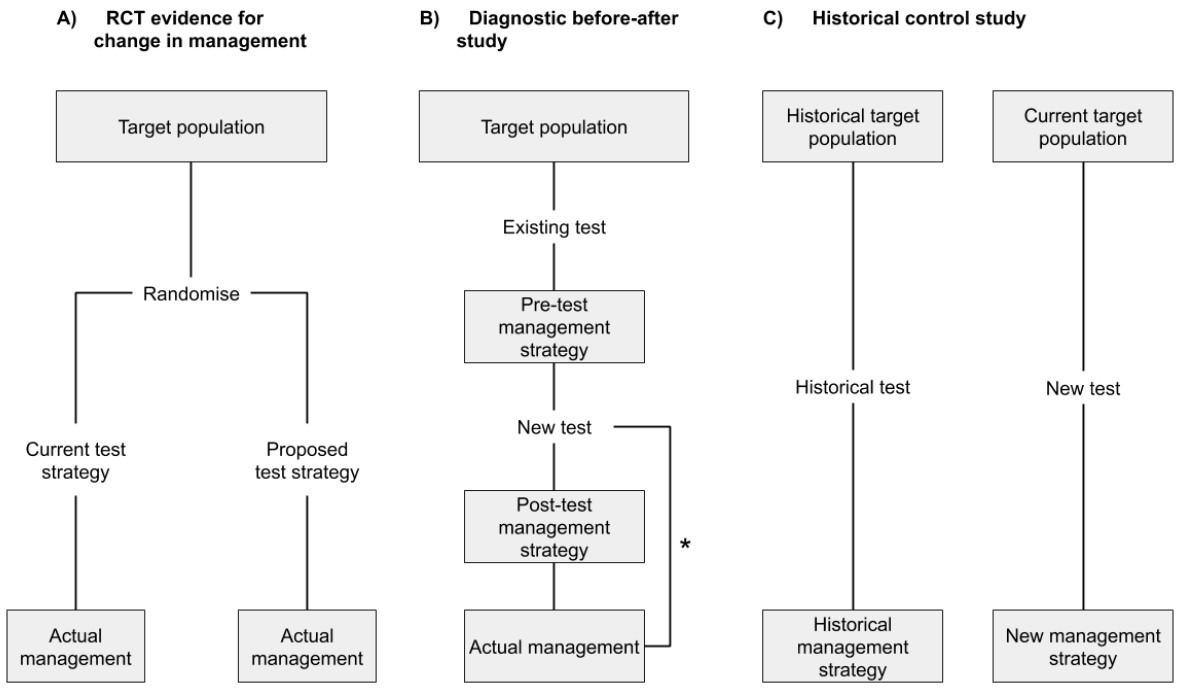


Figure 21 Change in management study designs (adapted from Staub et al. 2012)

***Diagnostic before-after studies may also capture actual management to validate the post-test management strategy. In some cases, post-test management strategy may not be included, with the study reporting on only the pre-test management strategy and the actual management following the test.**

Another study design which may be informative for change in management outcomes are historical-control studies, reporting on practice prior to, and after the introduction of the new test (Figure 21 – C). Historical control studies are at risk of bias due to factors *other* than the test also changing at the same time (i.e. more effective treatments may have become available or recall bias). In a similar manner, cohort studies with a concurrent control are also likely to be biased. If the cohort study compares two settings, which use or do not use the proposed test, there is the risk that the settings will not be similar enough to be able to attribute the changes in management to the test, instead of confounding factors. If an observational study compares individual patients in whom the test is used with those in whom the test is not used, there is the risk that there will be strong selection bias which could influence the results.

If evidence is not available to demonstrate that changes in management occur (i.e. an absence of evidence), expert opinion will be required to supplement the evidence review to justify any

assumptions regarding the impact a test has on the behaviour of clinicians, patients, family members etc.

Change in management studies are assessed as per therapeutic studies, with the NHMRC levels of evidence for interventions providing a relevant hierarchy of ideal study designs, and risk of bias tools relevant for therapeutics being used.

TG 12.4 Considerations relevant to change in management

Risk assessment – If a test is used to classify patients into a high or low risk group, with different treatment indications, the consequences of being reclassified will differ depending on the upwards or downwards reclassification. In some circumstances, there will only be a change in management if the reclassification happens in one direction (for example, current practice is to treat all patients, but the proposed test identifies very low risk patients who may not need treatment). Presenting the results for the different subgroups is therefore helpful, as the overall impact on the whole study sample will poorly reflect the results of the subgroups (Pletcher & Pignone 2011).

Addressing change in management as the first step – in general, when performing a linked evidence approach that requires the complete assessment through to health outcomes, the assessment should focus on evidence to support a change in management in the first instance. A lack of evidence to support change in management for claims of superior health outcomes will require additional justification (for example clinical expertise), and a thorough discussion of assumptions.

Need for empirical evidence – while it is always ideal that good quality evidence for change in management support a linked evidence approach, there are two examples where change in management data should be measured and robustly assessed:

- Tests used for monitoring: A monitoring test is the observation of a parameter over time. In the absence of direct from test to health outcomes evidence, clear evidence of the impact of monitoring on change in management is required. This may include observations that patients start or stop treatment, change the dose or duration of treatment, or takes some other action. Compared with tests that commonly have a clearly defined purpose and threshold for action, monitoring tests may not necessarily result in changes to treatment, or may trigger further investigation(s) which ultimately does not lead to changes in treatment.
- Tests used for outcome monitoring: Outcome monitoring describes a test used to determine response to an intervention. An example may be a CT scan used to determine whether a medicine is having the desired effect on a tumour. While monitoring for response to treatment may commonly be used as part of stopping rules in clinical studies, it is not guaranteed that such stopping rules will be applied in clinical practice. Clinicians may be reluctant to withdraw a treatment if the viable alternatives are limited (e.g. which may be the case for later line therapies), or if the treatment is perceived to provide a prophylactic mechanism (e.g. continuing glucocorticoids following resolution of a COPD exacerbation). As such, the results from clinical studies which employ clear stopping rules cannot be used to inform change in management in clinical practice, and empirical evidence of change in management is required.

In some circumstances, empirical evidence from studies reporting change in management may be less necessary (e.g. where earlier diagnosis of a serious disease is highly likely to result in earlier treatment for that disease; or identification of a biomarker by a co-dependent test leads to the use of a targeted medicine).

Impact of the change in management on the health system – If management of individuals change as a result of the proposed test, this may also have an impact on health care providers for the intervention, the comparator, or downstream investigations/treatments etc. For example, if a triage test reduces the number of patients being referred to a specialist, this may have an impact on the specialist workforce, waiting lists etc. These flow-on impacts should be discussed under 'Other relevant considerations' rather than in 'Change in management'.

TG 12.5 Assessment of the applicability of change in management evidence

Investigative technologies depend on the downstream consequences in order for health outcomes to change. These downstream consequences can vary a large degree based on the setting they are in, as organisational factors may affect their implementation and uptake. It is therefore important to consider how applicable the evidence is to the target Australian population and setting. If the change in management evidence is derived from a different setting to where it is proposed to be used, then the evidence may not be applicable. For example, if a test is proposed to be used in the general practice setting in Australia, but most of the evidence is derived from a specialist setting, this needs to be raised as an uncertainty and the applicability explored with subgroups or supplementary evidence.

Variations in management decisions occur across countries and within Australia[k]. Causes of variations in management decisions may be related to medical opinion[l], clinically driven or be influenced by non-clinical factors (Hajjaj et al. 2010). It may be useful to consider four key categories of factors that, should they differ across settings, may influence management decisions and therefore the applicability of change in management evidence from other settings (Table 10).

Table 10 Applicability issues relating to evidence for change in management

Causes of variation in management	Description of factors resulting in variation in management
Health system factors	Differences in referral systems, payments, remuneration or incentives may influence clinical practice. Geographical barriers or access (e.g. highly specialised care facilities vs rural facilities) or distribution of clinicians may also influence decision making.
Supply related factors	Differences in the availability of technologies across the settings, both in terms of regulatory (market access) availability as well as whether technologies are subsidised differently across settings. Clinicians in Australia will be more inclined to prescribe technologies that are available in their public hospital, or that are subsidised by the Commonwealth government, such as those technologies available on the MBS or PBS. Prescribers in other countries may likewise be compelled to recommend technologies that are subsidised by Government or Insurance.
Demand related factors	Differences in cultures, ethnicity, personal beliefs and values, and patient expectations may influence the management decisions, or

[k] https://www.safetyandquality.gov.au/sites/default/files/migrated/SAQ110_Medical_Practice_variation_V10_WEB.pdf
[l] https://www.dartmouthatlas.org/downloads/reports/agenda_for_change.pdf

	the adherence to management decisions. Education of patients and medical advertising can influence patient expectations.
Need related factors	Differences in population health, indicators of which may include age or demographics, socioeconomic status or environmental factors.

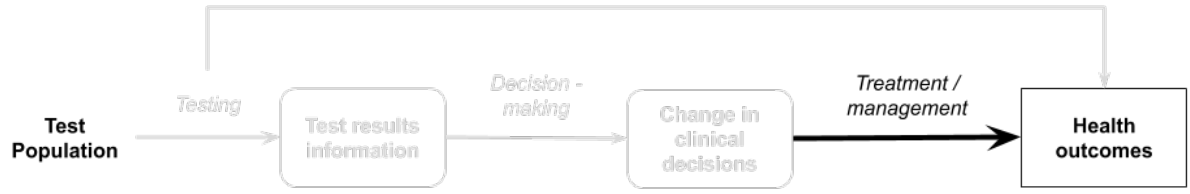
Source: derived from the causes of medical practice variation in https://www.safetyandquality.gov.au/sites/default/files/migrated/SAQ110_Medical_Practice_variation_V10_WEB.pdf

TG 12.6 Presentation of change in management evidence

The principles for presenting change in management evidence are similar to that for presenting clinical study evidence relating to a therapeutic technology. The key result that is sought by the assessment is the extent to which management changes (and the nature of the change) in a circumstance where the proposed test is available compared with when it is not available (and a comparator test or usual practice is applied). In addition to the main results for change in management, the assessment of change in management should consider these additional points.

1. Present the evidence for change in management in the same way as presenting evidence for a therapeutic medical service (i.e. describe the literature search, risk of bias, trial characteristics, present the results and meta-analyse, if appropriate).
2. Discuss reasons for variation in clinical management in patients with similar test results. Discuss whether the change in management may be confounded by other patient factors rather than the test results.
3. Discuss the applicability of the change in management evidence to the Australian setting. Where the evidence for change in management is partially applicable to the Australian setting, explore, where possible, the variation of management across subgroups, or present supplementary evidence to support the generalisability of the study results across settings.
4. Present a summary of the certainty of the body of evidence using GRADE.

Technical Guidance 13 Linked evidence - health outcomes



TG 13.1 Purpose of the guidance

Demonstrating that a test affects health outcomes provides the most confidence for MSAC to support the utility of the test. If direct from test to health outcomes evidence is available demonstrating that a test improves clinical outcomes compared to the comparator, guidance for providing this is presented in Technical Guidance 10. More commonly, evidence of health outcomes is demonstrated through a linked evidence approach, showing that a test changes clinician thinking, management recommendations, and treatments received. The last step of the linked-evidence approach (called therapeutic effectiveness evidence) is to establish the impacts of the change in management on health outcomes.

This TG section will discuss:

- A definition of therapeutic effectiveness evidence
- Therapeutic effectiveness study designs
- Considerations relating to therapeutic effectiveness evidence
- Assessing the applicability of therapeutic effectiveness evidence
- A suggested approach to presenting the therapeutic effectiveness as the final step in a linked evidence approach for a test

TG 13.2 Therapeutic effectiveness evidence

Therapeutic effectiveness evidence, as the final step of the linked evidence approach, includes an estimate of the impact of all the management decisions made as a consequence of using the proposed test in the place of standard practice.

In general, therapeutic effectiveness evidence should attempt to derive the highest quality evidence for the incremental difference in outcomes associated with treatment decisions informed by the proposed test versus treatment decisions informed by the comparator test. If therapeutic effectiveness evidence achieved this goal without concessions, it would resemble direct from test to health outcomes evidence.

In practice, the assessment of therapeutic effectiveness rarely achieves the certainty of direct from test to health outcomes evidence, and relies upon assumptions relating to the generalisability of the evidence across differently selected populations.

The results of therapeutic effectiveness evidence may provide an estimate of the impact on health outcomes for individual test populations, although it may not provide an estimate of the magnitude of the impact on health outcomes for the comparison of the proposed test versus standard practice. This is because there may be multiple sources of evidence for different populations, and aggregating

the overall health outcomes cannot be easily performed without a decision analytic model which links together test accuracy, change in management and treatment effect.

This evidence has several parts:

Assessment question	Description
Is there a treatment available?	Identify whether there are management strategies or treatments available for each of the test populations (this is informed by the change in management link).
Is there evidence that it is effective?	Identify evidence that the treatments are effective for the appropriate indication.
What are the implications for FP and FN?	Discuss the implications of misclassification (false positives and false negatives) on the health outcomes.
Is there evidence or a risk of a change in the spectrum of disease?	Consider whether the evidence for treatment effectiveness can be generalised from an unselected or differently selected population to the new test categories (including a discussion of whether the spectrum of disease following testing has changed).

There are several concerns relating to the applicability of the available evidence to each of the populations identified by test strategies, the transitivity across the evidence, and the subsequent impact on the validity of the economic analysis. Many issues associated with the derivation of the treatment effect for patients in each test category are discussed in greater detail during the construction of the decision analytic, should an economic evaluation be required.

The key clinical uncertainty associated with the final step in a linked evidence approach is that the treatment outcomes are not commonly derived from patients with a known test status. In many cases, treatment outcomes are sourced from unselected populations or differently selected populations, and it may remain unknown if the results are generalisable to the test positive or test negative populations for the proposed test.

The identification of suitable health outcomes evidence is an iterative process:

- Identifying whether there is a treatment or management strategy for the target condition.
- Identifying whether there is evidence that the treatment works in the target condition.
- Identifying whether there is evidence that the treatment works for the test subgroups.
- Assessment of the uncertainty or gaps in the evidence (applicability to the target population, generalisability of the evidence from an unselected to a selected population).
- Assessment of the impact of the uncertainty (direction of the effect of applying the identified evidence).
- Identification of supplementary data to support or reject the use of the identified evidence.

The process is iterative because it may not be apparent that the identified evidence is suitable or not until an assessment of the evidence for applicability and generalisability has been performed. Following this assessment, should the evidence be rejected, further searches may be required. It is not expected that these searches are performed systematically, rather, that targeted searches are performed to try and identify the highest level, or best evidence that address the impact of the change in management. For more information see Appendix 2 on literature searching.

TG 13.3 Therapeutic effectiveness study designs

Study designs required to complete therapeutic effectiveness data vary depending on the results from the change in management evidence and the decision problem. In general, the types of included study

designs are guided by availability of the evidence rather than what might be ideal. Studies may include randomised controlled trials, systematic reviews of randomised controlled trials, observational studies or registry data.

The following general guidelines may be useful in determining the types of studies that may be useful:

1. Comparative studies are useful to explore the impact of changing from one management strategy to another.
2. A relative treatment effect *is not* useful to describe the differences between treatments that are prescribed for different test populations (ie, positive and negative biomarker status).
3. Observational studies are useful for determining the natural history of the disease.
4. Studies comparing the outcomes of the same treatment by biomarker status may also be useful in identifying whether patients' biomarker status has a prognostic effect. Understanding whether the biomarker is prognostic or not may inform whether evidence from unselected populations receiving treatment A can be generalised to test-selected populations using treatment A.
5. Studies reporting on subgroup analyses defined by population characteristics or biomarker status may be useful in determining the applicability of the evidence to the target population.

TG 13.4 Considerations relevant to linked evidence of health outcomes

Generalisability of the evidence

Unless a new test is substantially safer (and avoids adverse events), for a new test to have an impact on health outcomes, it must result in a change in management and alter the allocation of patients across treatments. If the test is relatively new, there is unlikely to be evidence for the outcomes of patients allocated to treatments according to results of the proposed test. Therefore, health outcomes evidence for the treatments identified in the change in management section may not be generalisable to the population receiving that treatment following the use of the proposed test. A simple diagram showing the discordance is presented in Figure 22.

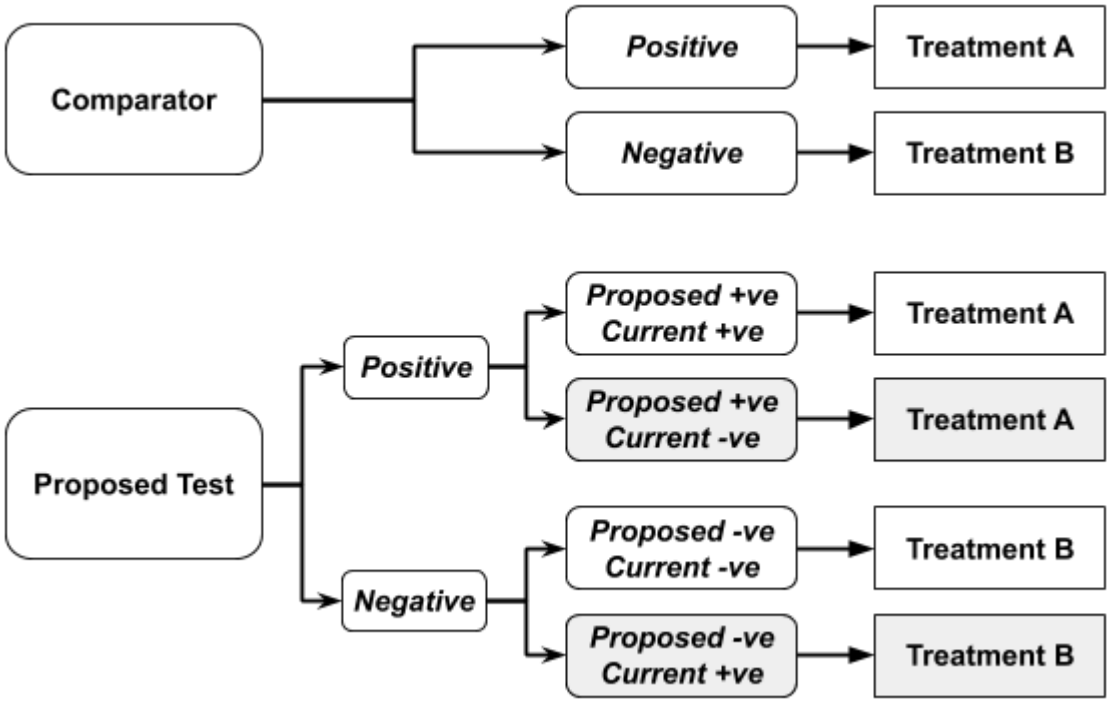


Figure 22 Diagram showing the change in the treatment of patients categorised using the proposed test compared with standard practice

In Figure 22, current practice would allocate patients to established treatments A or B. Current practice may reflect current testing or clinical assessment alone. The treatments provided in the diagram above may be any type of treatment decision (e.g. two different treatments, the decision to give adjuvant therapy, the decision to provide an add on treatment, the decision to withhold treatment).

The figure shows that, using the new test, some patients who previously received treatment B will now receive treatment A, and some patients who previously received treatment A will now receive treatment B.

In general, evidence for the treatment outcomes for patients receiving Treatment A and for patients receiving Treatment B will be available. A key concern is whether the evidence, which contains some patients that would be allocated differently by the proposed test compared with the comparator test, can be used to approximate the health outcomes for the proposed test. Not only may the treatment effect for patients differ depending on their test result, but the prognosis of patients with different test results may also differ. An assessment of the generalisability of the evidence to different populations should include the following:

- A description of the patients who change management (how do these patients differ from those that did not change management?). For example, are the patients who test positive with the proposed test similar to those that test positive with the comparator test? Do any differences indicate a change in the spectrum of the disease being identified?
- Provide evidence whether the test status is related to prognosis.

Spectrum of the disease

A change in the spectrum of the disease may occur when a new test is introduced and may result in either less severe (more common) or more severe disease being identified. For example, a test that

is more sensitive may detect disease earlier, that is less severe, or even inconsequential disease (eg overdiagnosis).

It is important to consider whether the proposed test has resulted in a change in the spectrum of the disease when considering the generalisability of the evidence. A change in the spectrum of the disease may result in:

- Increased lead time (earlier detection of a parameter).
- Change in the efficacy of treatments (either in relative or absolute terms).

If there is a risk that the patients treated following the proposed test differ in the spectrum of their disease compared with the patients treated in the identified therapeutic effectiveness evidence, the evidence may not be generalisable (Merlin et al. 2013). An assessment of the possible impact of a change in the spectrum of the disease should seek to define first how the spectrum of disease has changed.

Proposed test is more sensitive: this will result in more patients being identified as positive. In terms of test accuracy, this reflects the current categorisation of false negative patients moving to a true positive state. Evidence of the treatment effect in these additional patients, who may have less severe or earlier stage disease, may be required. A comparison of early versus late treatment may be informative.

Proposed test is more specific: this will result in fewer patients being identified as positive. In terms of test accuracy, this reflects the current categorisation of false positive patients moving to a true negative state. Evidence of the harms of inappropriate treatment of negative patients may be required.

Although not exclusively related to a change in the spectrum of the disease, one particular issue associated with detecting earlier or less severe disease is the risk of over-diagnosis. Over-diagnosis may occur if a pathological lesion or state is identified, and the patient is therefore identified as having a disease, and there is no evidence that this state leads to a poor health outcome or investigations/treatments that benefit the patient, but there is evidence of potential harm (Carter et al. 2016). Five indicators may be used to identify potential overdiagnosis: is there potential for increased diagnosis? Has diagnosis actually increased? Are additional cases subclinical or low risk? Have some additional cases been treated? Might harms outweigh benefits? (Bell, KJL et al. 2019). Over-diagnosis can result in unnecessary health care, and can result in harm. Therefore, it is important that evidence is presented on the likely future benefits or harms of identifying a condition as an abnormal disease state (Croft et al. 2015).

Risk of bias and transitivity

The assessment of risk of bias is important to determine the internal validity of studies identified for establishing the health outcome gains resulting from management changes. Care should be taken when assessing risk of bias, particularly if the use of the identified study for the purposes of the assessment report differs from the original assessment question of the study. This may arise if:

1. Only part of the identified study is used, such as a single arm of a randomised controlled trial; or,
2. Subgroups are used to address the health outcome gains of the management strategy (this may occur if biomarker subgroups are identified in the trial and are used to determine the treatment outcomes).

The assessment of risk of bias should reflect how the study was applied in the assessment report, rather than the original intent of the study.

A second key concern will arise if more than one source of evidence is required to assess the health outcome gains of all the management options (which is highly likely). Under these circumstances, it is important that the differences in health outcomes across studies is a consequence of the different treatments and that the populations differ only if the evidence is intended to represent different populations (ie, biomarker negative or biomarker positive). Within the economic analysis that will be informed by the health outcomes and treatment effects derived from the health outcome evidence, the different sources of evidence may be applied independently, with outcomes aggregated in the test and comparator arms. This is similar to a naïve indirect comparison, and therefore the transitivity of the evidence presented for the health outcome gains link is important.

Regression to the mean

It may be necessary to source single arm evidence to inform the therapeutic effectiveness link. Single arm evidence can be subject to regression to the mean. If patients are selected for treatment based on the severity of the condition, there is the chance that patients will improve, due to regression to the mean (Morton & Torgerson 2005). Assessing the effectiveness of selecting patients who will benefit from treatment therefore needs to consider whether the patients will have improved anyway. If randomised evidence is not available, preference should be given to studies where the baseline measure of the outcome variable is separate from the measurement used to select patients (Morton & Torgerson 2005).

TG 13.5 Assessment of the applicability of health outcome gains evidence

Therapeutic utlity evidence can differ from the target setting in multiple ways, and the applicability of the evidence is markedly influenced by the applicability of prior steps in the linked evidence approach.

The following applicability concerns may be considered:

- Test related applicability – are testing components in the health outcome gains evidence (proposed, comparator or standard practice) the same as the current or proposed clinical practice? (See Technical Guidance 11)
- Change in management applicability – is the interpretation of the test results, and change in management in the health outcomes evidence the same as the current / proposed clinical practice? (See Technical Guidance 12)
- Health outcome gains applicability – assess the applicability of the health outcome gains evidence in the same way as assessing applicability for a therapeutic intervention (comparison of the evidence with the Australian setting for patient characteristics (demographics, disease) and intervention characteristics (dose, duration, setting) (See Technical Guidance 13).

TG 13.6 Presentation of health outcome gains evidence

The principles for presenting health outcome gains evidence are similar to that for presenting clinical study evidence relating to a therapeutic medical service. The key results presented in this section will depend upon how the evidence for comparing the treatment outcomes across the proposed test and comparator test strategies is constructed. The assessment of health outcome gains evidence should consider the following points:

1. Explain how evidence for each of the management pathways has been constructed, and how it relates. Clearly identify where outcomes for one group are based on a relative treatment effect compared to another group, and when the evidence is derived from different sources.
2. Present the evidence for health outcome gains (for each different management strategy) in the same way as presenting evidence for a therapeutic medical service. That is, describe the

literature search (and any subsequent searches for supplementary evidence to explore uncertainties in the evidence), risk of bias, trial characteristics, present the results and meta-analyse, if appropriate.

3. When presenting the results, provide an assessment of the outcomes relating to the change in management – eg, if a test results in 20% of patients receiving Treatment A instead of Treatment B, a comparison of Treatment A vs Treatment B is appropriate.
4. Clearly describe the assumptions required to generalise evidence across groups. For example, the treatment effect is assumed to be the same for patients who test positive using both the comparator test and the proposed test.
5. Present evidence to support the generalisability of the evidence across different populations. This may include evidence to address the risk of a change in the spectrum of the disease, or a change in the prognosis associated with the biomarker. If there is a change in accuracy that may alter the spectrum of the disease, clearly discuss the implications of changing treatment on test positives (particularly the new true positives) and test negatives (particularly the new true negatives). Where appropriate, include a discussion of prognosis and over-diagnosis.
6. Discuss the applicability of the health outcome gains evidence to the target Australian setting. Include the impact of applicability issues identified at the test accuracy step or the change in management step. Explore the impact of applicability in subgroup analyses.
7. Present a summary of the quality of the body of evidence (GRADE).

If the outcomes are considered surrogates or intermediate outcomes, rather than critical outcomes of value to patients, then the link between these outcomes and patient-relevant outcomes should be assessed (See Appendix 12 on surrogate outcomes).

Preference will be given to evidence of health outcomes, where the treatments provided are available to patients in Australia. The the health benefit of a test that is only demonstrated through the subsequent use of a treatment only available in a trial setting should not be considered.

The safety of any downstream effects of the test should be discussed as part of the health outcome gains. For example, if the test is a triage test to rule out invasive testing in those who don't need it, the safety of the further testing should be discussed. If the test results in different proportions of patients receiving treatments than if they had had the comparative test strategy, then any adverse effects of treatments received should be assessed.

Technical Guidance 14 Safety of investigative technologies

An assessment of impact of health outcomes from the use of a health technology includes an assessment of relative safety versus the main comparator. The assessment of safety has three key parts for investigative technologies:

- The assessment of the direct and more immediate impacts (adverse events) of the use of the health technology (often captured to a varying degree in the included clinical studies);
- The assessment of impacts of downstream implications related to the management decisions following a test; and,
- The assessment of longer term or rarer safety events unlikely to be captured in clinical studies.

The objective of an assessment of safety is to identify the relative safety of performing the proposed test versus the main comparator, which may be an alternative test or no test. The assessment of the safety of a test (or the comparator) involves the assessment of both the direct (often immediate) harms associated with the test itself, as well as harms associated with downstream consequences of testing. It is important to present the direct harms of testing and harms of downstream management decisions separately. However, a narrative synthesis that discusses the entire safety profile may be informative, particularly in cases where there is a trade-off (e.g. the proposed test has an inferior safety profile than the comparator, however the use of the test results in improved safety outcomes for treatment).

Comparative safety data may be available from direct from test to health outcomes evidence, or from test accuracy studies. However, additional searches for test safety may be required.

TG 14.1 Test-related adverse events

The direct harm of testing is the immediate or delayed safety consequences of physically performing the test. For most tests (particularly in vitro tests), the direct safety of a test is related to the method of performing the test or retrieving samples. In many cases, the mode of testing for an investigative technology is established (eg, imaging using a CT scanner, a biopsy, or a blood test).

The direct safety of testing may not be necessary in the following circumstances:

- The proposed test uses the same modality as the comparator test it is intended to replace; and,
- The proposed test will be used in the same proportion of patients.

A discussion of why the direct safety of the proposed test versus the comparator test is considered the same (or identical) is required if no quantitative safety data are to be presented.

The direct safety of testing will be necessary in the following circumstances:

- The proposed test is used in addition to current testing;
- The proposed test will be used in a greater number of patients (due to acceptability or the need for additional samples / repeat biopsy); or,
- The proposed test is performed using a different modality to current testing. This may include a change in biopsy techniques, a change in how invasive the sample retrieval is, a change in the extent of radiation delivered for scans, or a change in patient factors due to tests being performed at different time points.

Where comparative data are available, these should be presented as dichotomous outcomes.

Report adverse event data as both the number of patients reporting an adverse event in each category and the absolute number of adverse events in each category. The absolute number of events in each category may be a more appropriate estimate for costing adverse events in an economic or financial analysis, rather than the number of patients who experience an adverse event, because the latter will not capture patients who experience two events in the same category.

For each important adverse event, present these results as for dichotomous data, and include relative risks and risk differences with their 95% CIs across the groups for each study, separately. Where appropriate, meta-analyse the results using a random effects model and provide an interpretation.

Analyse the relative adverse event rates (events per period at risk), if the average period at risk per participant varies substantially between treatment groups (eg using a straight Poisson regression or a negative binomial approach). Present the assumptions associated with statistical analyses and how they were tested.

If the evidence for comparative safety is insufficient for a test, the assessment should provide some context of the likely safety profile of the proposed test and the comparator by searching for good quality evidence of the safety of tests that use the same method. For example, studies of a CT scan of a similar region of the body in a similar aged patient may provide some indication of the safety of a proposed CT test. Another example is studies of a biopsy of the same organ in patients with similar performance status may provide some indication of the safety of a proposed somatic mutation test.

The search for supplementary evidence is only required if the identified safety evidence for the proposed or comparator test is insufficient or uncertain. The search does not need to be systematic, but should aim to identify a high quality study that is applicable to the Australian setting. If supplementary evidence is included, discuss the applicability of the evidence to the assessment of the proposed test.

TG 14.2 Downstream safety consequences

The use of the proposed test may result in patients receiving different treatments compared with the use of the comparator. This change in management will result in patients being exposed to different safety profiles associated with treatments. The assessment of safety outcomes for treatments informed by testing is similar to the assessment of efficacy outcomes for a therapeutic technology (see TG 7.1). Considerations relating to direct from test to health outcomes evidence (Technical Guidance 10) or to linked evidence of the impact of change in management (Technical Guidance 13) are relevant for the assessment of safety outcomes.

Separately summarise the safety issues for patients with a positive test result and a negative test result. If there is a reason that the type, severity or number of safety events would differ for true and false positives, or true and false negatives, explain why and describe.

The overall summary of evidence (described in Technical Guidance 16) should clearly describe the safety issues for false positives and false negatives alongside the impact of misclassification on treatment outcomes.

TG 14.3 Test safety unlikely to be captured in clinical studies

Assessment of longer term or rarer safety beyond clinical studies may be relevant if the change in management involves a therapeutic technology, particularly if the therapeutic technology is novel or being used in an new indication.

An extended assessment of the direct safety of a test may also be useful if a novel mode of testing is used, or there is uncertainty for the longer term or rare effects of testing.

Ideally, the estimate of the relative safety of a health technology is derived from high quality comparative studies. Clinical trials are often inadequate for providing data on comparative harms for a few reasons:

- Trials tend to enrol patients who are healthier, have fewer comorbidities or concomitant medications, and have more stringent monitoring than the target population.
- Trials are usually underpowered and of insufficient duration to detect important adverse events.
- Adverse events in clinical trials designed to emphasise efficacy results are often underreported (Pitrou et al. 2009)

Discuss whether the included evidence base is adequate for identifying:

- Less common adverse events or safety concerns
- Adverse events that may occur in the longer term
- Harms that may occur due to differences in the target population and the more selected population that may be enrolled in a clinical trial

If the included evidence is not sufficient to capture long term or rare adverse events, or adverse events in patients with comorbidities or receiving concomitant treatments, present additional evidence. Describe the search strategy for identifying nonrandomised studies of the proposed health technology, or registry data. If appropriate, include evidence of safety involving the proposed health technology in other indications. Where the proposed medical service is delivered in combination with an implantable device, provide an assessment of the safety of that device. Sources of safety information may include device registries, regulatory databases, complaints registries and post-market surveillance studies.

Technical Guidance 15 Special cases

TG 15.1 Screening

Screening is a form of investigative technology which may be used (in isolation or combination) to lead to early detection of a target condition which may benefit from early intervention. There are different types of screening, with different aims. Screening tests should not be confused with diagnostic tests, which are those investigative technologies (in isolation or in combination) that tend to be applied to symptomatic individuals to elucidate information that explains and/or assists in managing their current clinical presentation.

Universal or population screening

Universal screening involved the testing of all people from the population who meet certain criteria (i.e. through programs such as the Newborn bloodspot screening, BreastScreen Australia, the National Bowel Cancer Screening Program, and the National Cervical Screening Program). Delivery of preventive services (such as screening) is predominantly under the remit of the states and territories, although there are times that the Commonwealth and states/territories share responsibility. MSAC may occasionally be requested to assess universal screening tests (examples of prior assessments are for neonatal hearing screening, and digital mammography).

Deciding whether an investigative medical service should be incorporated as part of a population-based screening program is not simply a decision based on epidemiological evidence. The effectiveness of population-based screening depends on both the accuracy of the screening test and the clinical effectiveness of early detection and intervention. A good screening test must detect the target condition earlier than without screening, and with sufficient accuracy to avoid producing large numbers of false positives and false negative results. Screening and treating those who test positive should also improve the likelihood of favourable health outcomes. These potential benefits should not be measured as disease outcomes/cases diagnosed (eg 5-year survival) as these measures are skewed by lead time bias and overdiagnosis (Doust, JA, Bell & Glasziou 2020).

The prevalence of the biomarker and disease for conditions being universally screened for are low, which can mean that there is a high number of false positive screening results for every true positive screening result. The downstream effects of further investigations and intervention are therefore critical to assess, and the preferred method of assessing universal screening is through the assessment of direct from test to health outcomes evidence from screening.

Historically, conditions were only screened for if there was a method of treating the condition once diagnosed. However, there are some advocates for expanding the definition of clinical utility of screening tests to incorporate family's quality of life, which may be improved by early knowledge of their child's condition, even if no treatment is available (Burke, Laberge & Press 2010). Early detection of a genetic condition in one child may provide parents with information to inform future reproductive decision-making.

Universal screening programs are considered to have a high financial risk associated with them. MSAC therefore has a clear preference for restricting consideration of clinical utility of universal screening to health outcomes, rather than the value of the information itself.

Targeted screening

Targeted screening is testing of asymptomatic people who are at high risk of a given clinical condition/disease. Screening may be targeted so that the harms associated with the screening test (e.g. radiation exposure and over-diagnosis) are outweighed by the benefits of earlier detection of disease. The people screened may be considered high risk due to personal characteristics (age, gender,

history of known medical risk factors), family history, and/or specific exposures (e.g. workers in lead battery factories). For targeted screening, the health effect of the condition prevented may be more minor than considered for universal screening (i.e. nausea or vomiting) but the screening may be a high priority if the effect reduces the patient's ability to work. Targeted screening may be legally required (e.g. minors who work with lead or chromium) and used in follow-up to environmental health incidents.

Based on the definition established for genetic testing, the definition of high risk is a ≥10% pre-test risk of having the disease or the biomarker.

When an index patient is found to have a heritable genetic variant, family members may potentially undergo cascade testing for the familial variant, if clinical or personal utility is able to be established for the predisposition testing. More on cascade testing is shown in TG 11.6.

Targeted screening may be assessed using either the linked-evidence approach or using direct from test to health outcomes evidence.

TG 15.2 Monitoring

Some investigative technologies are intended to be used as part of a monitoring strategy. Monitoring can be summarised as consisting of five phases (Bell, KJ et al. 2014; Doust, J & Glasziou 2013):

1. Pre-treatment monitoring (surveillance): to screen individuals on the need to start treatment;
2. Initial response monitoring: to determine whether the individual's response to treatment is as expected from the mean response observed in trials;
3. On-treatment long-term monitoring: to assess whether treatment remains adequate over the long term;
4. After a significant change in the disease process or treatment has occurred; and
5. To determine if it is possible to stop treatment.

The assessment of a test used for monitoring is similar to other uses, in regards to the need to investigate the clinical utility of the test (direct from test to health outcomes evidence), how the monitoring test influences management, the impact of the change in management on patient health outcomes, and test accuracy. Increased monitoring may lead to increased anxiety, or could increase feelings of empowerment. These outcomes should be addressed either under direct from test to health outcomes evidence, or personal utility.

However, there are two additional aspects of test accuracy (in addition to diagnostic or predictive accuracy) that are relevant to address for monitoring(Bell, KJ et al. 2014):

1. Responsiveness: how much the test changes in response to an intervention/treatment change relative to background random variation (signal to noise ratio).
2. Detectability of long-term change: the size of change in the test over the long term relative to background random variation, and what frequency of monitoring therefore is logical.

Further assessment should also be performed on:

3. Practicality: the ease of use and interpretation of the test, cost and level of invasiveness.

Responsiveness describes how much the test changes in response to a therapeutic technology relative to background random variation. The responsiveness criterion is especially important for the initial response phase of monitoring soon after a new treatment has been started. Although less obvious,

this criterion is also important for both pre-treatment and long-term monitoring. For all monitoring phases, ideally the test should be responsive to treatments that alter the patient's risk of the clinical outcome. Such interventions may be lifestyle changes in the pre-treatment phase, pharmacologic treatments in the initial response phase, or measures to improve adherence in the long-term monitoring phase.

Related to the concept of responsiveness is the speed of change in response to an intervention. Preferably a test should show a rapid response to treatment. This is obviously a necessity when the change in outcome in response to the intervention is also rapid, for example, risk of hypoglycaemia for glucose-lowering medicines (monitor glucose) or bleeding risk for patients on warfarin (monitor international normalized ratio). In other situations in which the change in outcome is much slower, it is still preferred that the test response can be quickly judged whether treatment is working as expected, for example, risk of a cardiovascular event (monitor cholesterol and blood pressure). Not all responsive tests show rapid changes in response to treatment; in fact, some take months/years to change, for example, HbA1c. Because changes in the results of the tests reflect average treatment effects over a longer period of time, these services may be preferred for judging effects over the medium to long term.

If initial response monitoring is considered to be of value, the frequency of monitoring should be considered and justified.

Once a patient is stable on treatment, the frequency of monitoring can be decreased. The frequency of monitoring would depend on the within-person variability of the monitoring test, and the range of rates of long-term change (e.g. the rate of progression or regression of the disease).

The concepts of **''signal''** and **''noise''** are relevant to both response monitoring and long-term monitoring. For response monitoring, signal includes both mean change and between-person variation in response. If the between-person variation component of the signal is small, then the signal for an individual can be estimated using the population mean change without needing to monitor. If the between-person variation is not small, it is difficult to estimate signal on the basis of population mean change alone. The individual's true deviation from the mean change also needs to be estimated and this is best done where there is a favourable signal-to-noise ratio. Noise is a result of background random variation within individuals because of measurement error and biological fluctuations. The amount of noise in investigative technologies used for monitoring may not be appreciated by clinicians, and variations due to noise may be wrongly attributed to real change.

A study by Bell et al. (2008) addresses the assessment of initial response monitoring, and when it is worth monitoring initial response to treatment (Bell, KJ et al. 2008). Treatment for patients with chronic conditions is often monitored by using surrogate outcomes (such as blood pressure or cholesterol). A surrogate outcome should only be considered for monitoring if it is known to predict the treatment effect on risk of the clinical outcome. Monitoring initial response to treatment should be avoided unless it is expected that it is useful in informing clinical decision making. Monitoring is unlikely to be of value when there is no evidence of variation in the response to treatment between patients/subgroups, or when it is highly likely that therapeutic targets will be met. Variability in treatment effects between individuals can be estimated from placebo controlled randomised trials.

Detectability of long-term change describes the size of changes in the results of an investigative technology over the long term relative to background random variation. Long-term change describes the ability of the service to discern true long-term changes in the patient's condition (signal) from short-term measurement variability (noise). The signal for long-term change monitoring is the true long-term trend in level within an individual over time. This is a combination of the population mean change and the between-person variability or individual deviations from the mean change. Noise is the same as for response monitoring: short-term random variation in level within an individual. Unlike response monitoring in which the between-person variation component of the signal is often small,

rendering monitoring unnecessary, in long-term monitoring, there is usually substantial between-person variation in the long-term trends. This means that it is difficult to estimate a signal on the basis of population mean change alone and the need to also estimate the individual's true deviation from the mean change under conditions of a favourable signal-to-noise ratio. The frequency of long term monitoring should be determined based on the rate of 'drift' and the closeness to the target or threshold value (Doust, J & Glasziou 2013; Glasziou, PP et al. 2008). The closer the measurement is to the threshold, the sooner a repeat measurement is needed.

Finally, the **practicality** of the test as a monitoring tool describes its ease of use, level of invasiveness and cost. Although every assessment should consider the practicality of the proposed technology, monitoring tests are likely to be repeated, and may involve a component performed by the patient or family member than other forms of test. Ease of use and invasiveness is therefore likely to have a higher impact on compliance than for a once-off test such as used for diagnosis.

Co-dependent technologies

If an investigative medical service that is being considered by MSAC is being proposed to be used as part of a monitoring strategy that informs the use or disuse of a pharmaceutical concurrently seeking PBS listing through PBAC (and thus the pharmaceutical is co-dependent on the investigative medical service) then a paired application across both committees applies (see TG 15.4).

TG 15.3 Multifactorial algorithms

A multifactorial or multicomponent algorithm can be diagnostic, prognostic or predictive in intent. It can use static rule-based prediction models, or adaptive self-learning algorithms that create their own models. Fixed algorithms can be based on either a sequence of simple if → then statements or a sequence of more complex mathematical equations and are limited by the size of the underlying knowledge or rule-base as defined by human experts. Fixed algorithms can also be derived from adaptive self-learning algorithms that have been "locked" so that they can no longer automatically adapt or change over time.

Adaptive self-learning algorithms are derived using artificial intelligence technology and can be divided into two categories: machine-learning algorithms derived using structured data (maps input data to known outputs) in order to complete a task, and deep-learning neural network models (only input data is provided and must self-learn relationships) that are able to cope with unstructured data and unforeseen circumstances and still function (Scott et al. 2019). However, as it is not always straightforward to know whether the underlying data is structured or unstructured, there is no clear separation between machine-learning and deep-learning algorithms. Artificial intelligence is defined as any computer method that mimics human intelligence, such as pattern recognition, abstract reasoning and planning (Scott et al. 2019).

To enable assessment of the biological plausibility of the multifactorial algorithm, the relationship between basic clinical characteristics used to develop the algorithm (e.g. presence of a biomarker, age, specific imaging formations, etc) and the outcome should be examined. This should be possible even if access to the actual algorithm's decision points is restricted by the propriety of the "invention" and is not available for assessment.

Another important issue for multifactorial algorithms is the generalisability of the algorithm to the relevant Australian population.

Static or fixed algorithm

Static or fixed algorithms are developed in two steps: a training or discovery step (where the algorithm is developed) and a validation step (where the performance of algorithm is confirmed). Sampling bias

warrants additional attention when assessing the generalisability of the training cohort (Cahan et al. 2019; Park et al. 2019). There needs to be transparency about recruitment and the baseline demographic and clinical characteristics of the cohorts used in both of these steps. If certain patient subgroups are under or over represented in the input data, the algorithmic outputs may not be representative, introducing a systematic bias. Further studies using a cohort representative of the targeted Australian population may be needed to assess the performance of the algorithm in the Australian setting.

Quality appraisal of risk assessment algorithms

Fazel and Wolf (2018) developed a simple 10-point checklist covering key factors to consider when determining the quality of risk assessment algorithms. Some of these factors may relate more closely to static or fixed algorithms (such as knowledge of the included variables and the weighting of the variables). When applying the checklist to dynamic or self-learning algorithms, amend the included factors as appropriate. These key factors are:

- Study type (external validation study in new sample or a derivation and internal validation study)
- Was the study based on either a protocol or pre-specified analyses?
- Did the algorithm have a set objective and clearly defined outcomes?
- Did the study report the discrimination and calibration measures used in the algorithm?
- Were the variables included in the algorithm the same for both the validation and derivation studies?
- Was the weighting of the variables the same for both the validation and derivation studies?
- Was the study population the same as that of interest?
- Was the sample size sufficient?
- Did the study report predefined output categories or thresholds?
- Was the study published in a peer-reviewed journal? Note that this checkpoint item should not conflict with the preference for the best available evidence. In general, all evidence of the proposed health technology should be provided. The evidence would preferably include complete clinical study reports and protocols, and any additional supportive evidence that may be required.

If the tool has not been externally validated, it should not be routinely used in practice.

Self-learning algorithms

A self-learning or adaptive algorithm would extend beyond these two steps of discovery and validation. After establishing the initial performance characteristics of the algorithm using these two steps, a description is needed of how the algorithm is programed to improve performance over time. This should explain how new data would be used to improve the performance of the test result, what methods would be used to achieve this, what would be optimised, and what quality controls are applied. It should also be determined whether it is necessary to revalidate the test/intervention after implementation and how this should be done. Some argue that self-learning or adaptive algorithms should always be subject to rigorous audits after approval, as unsupervised self-learning tools will account for new variables as time goes by and therefore their predictive performance will change over time and over populations (e.g. an algorithm's systemic bias against certain population groups may only emerge when it is deployed across large populations.) If the algorithm is intended to be used across different and importantly heterogeneous target populations, it must demonstrate how its performance also improves across these populations.

It should be clear how incomplete or inaccurately entered data would be handled by the algorithm, and the robustness of the computer algorithms to all types of missing data inputs and adversaries should be demonstrated.

This should be supported by evidence of improved performance at different points of time to date and projected improvements in the future. Similar to the initial development and validation steps, there needs to be transparency about recruitment and the baseline demographic and clinical characteristics of the cohorts providing subsequent inputs to these algorithms. This should be managed to minimise any existing biases (e.g. from the use of existing historical data). For example, describe the plan to ensure that the ongoing acquisition of subsequent inputs reflects current practice and population characteristics. Similarly, the representativeness of these subsequent inputs to the targeted Australian population is an important consideration for assessing the ongoing performance of the algorithm. New methods and technology infrastructure may be needed to address potentially complex issues. As deep learning tools are able to account for new variables, their accuracy may change as time goes by (Parikh, Obermeyer & Navathe 2019). Alternatively, if action, taken as a result of the algorithm, prevents the predicted outcome, this outcome would not be observed in the post-implementation period and could bias the results in ways that are difficult to ascertain (Sendak et al. 2019).

The outcomes likely to be important for assessing the clinical utility of multifactorial algorithms include the impact of: diagnostic accuracy and therapeutic effectiveness, time savings on routine tasks, turnaround time for investigation results; costs of care reductions, and/or improvements to patient outcomes (Scott et al. 2019). Following cohorts over time (preferably prospectively rather than retrospectively) is important to support claims of prognostic or predictive performance. However, longitudinal clinical studies and cost-effectiveness analyses are likely to be sparse for these algorithms.

The prognostic or predictive capability of self-learning algorithms may raise specific ethical and legal challenges, such as privacy and discrimination (Char, Shah & Magnus 2018; Jaremko et al. 2019). An example might be when earlier and subsequently outdated prognoses about a patient's health, cognitive decline or addiction appears in their health record, may lead to discrimination, psychological harm or raise issues of equitable access. This should be discussed and considered when assessing self-learning algorithms.

Genomic predictive algorithms

An MSAC working group for the evaluation of genomic multifactorial algorithms determined that additional specific criteria need to be addressed. The process by which the algorithm/model was built and updated should be described, as remaining ignorant about the construction of machine-learning systems could lead to ethically problematic outcomes. Fixed quality control protocols should be used for the preparation of samples, and evidence of adherence to these protocols should be provided. Adherence to adequate quality control protocols should be demonstrated, applied to all aspects of the process of testing (including reagents, hardware and software). This is necessary to have confidence that the processes "locked" in place after the discovery phase are maintained throughout validation and into its regular intended use. These processes should be fixed before being evaluated by MSAC, and any static or fixed test developed and assessed should be identical to the version proposed to and approved by MSAC.

TG 15.4 Co-dependent technologies

Health technologies are co-dependent when the patient health outcomes related to the use of a therapeutic health technology (e.g. a medicine) are improved by the use of another health technology (e.g. an investigative technology). The combined use of these technologies leads to their intended clinical effect, and therefore the benefits of both technologies would be assessed together (instead of

assessing each technology in isolation). Appendix 8 provides detailed guidance on the preparation of a submission that involves co-dependent technologies (mainly a genetic test and a medicine) between MSAC and the PBAC.

Investigative technologies could be involved in co-dependent technologies for the purpose of:

- establishing a predisposition or estimating a prognosis
- identifying a patient as suitable for a therapeutic technology
- measuring an early treatment effect on a surrogate outcome as the basis for predicting more patient-relevant health outcomes
- monitoring a patient over time after an initial investigation to guide subsequent treatment decisions.

A co-dependent submission is required when the Minister for Health requires advice from two different expert advisory committees because listing of the co-dependent technologies involve two separate reimbursement schemes. For example, co-dependent technologies that require new listings or amendments to both the PBS and the MBS would need advice from both PBAC and MSAC. Co-dependent submissions can be integrated (one combined submission for the two technologies and considered jointly by both MSAC and PBAC) or streamlined (two individual submissions for each of the technologies).

Other than the combination of a genetic test to determine eligibility for a medicine, which is the majority of co-dependent technologies, there could be other co-dependent technologies; e.g. an application for an MBS service (assessed by MSAC) in combination with an implant (which would be listed on the Prostheses List and would need to go to PLAC), an imaging or blood test to determine eligibility for a therapeutic service or medicine, a medicine delivery system in combination with a medicine, or a monitoring test to determine specific therapeutic management or medicine dose.

When an application for a co-dependent technology is purely for a therapeutic purpose (e.g. the combination of an MBS service with an implant), the two interventions would most likely be assessed together, as one therapeutic intervention.

Technical Guidance 16
Interpretation of the investigative evidence

The objective of summarising the overall evidence base is to describe the results of the assessment report as they apply to the clinical claim in the specific context of the Australian setting. Follow the guidance provide in TG 8.1. In addition, the interpretation of investigative evidence requires an overview of how the evidence has been constructed, the effects on subgroups defined by testing, and an identification of areas of uncertainty, particularly when evidence has been linked. This is described below in TG 16.1.

TG 16.1 Investigative evidence interpretation

Summarising the evidence base for a test must account for the quality and strength of the evidence for each individual component, as well as an overall assessment of the effect of the proposed test on health (or other) outcomes.

Consider the following structure of the summary of the evidence base:

- A description of the actual evidence approach that was taken. Describe the direct or linked approach, and any supplementary evidence that was required.
- A summary of each evidentiary step, addressing the components listed above.
- With reference to the key uncertainties identified for each evidentiary step, provide an overall interpretation of the comparative impact of the proposed test on health outcomes (and personal utility outcomes if required).
    - This interpretation considers the implications of evidence and uncertainty at each step on the results of subsequent steps in a linked approach. For example, the impact of a test with poor specificity on treatment outcomes, or the impact of uncertainty in change in management on treatment outcomes.
    - When a linked evidence approach to evaluating a test has been used, it is critical for the assessment to link the pieces of the evidence together, to come up with a conclusion regarding the impact of the test on outcomes, compared to the comparator. One way to summarise the information would be to start with a hypothetical population who undergo testing, and follow the population through to the range of different outcomes, incorporating change in management, the proportion treated appropriately versus inappropriately (based on accuracy of test results), and the patient-relevant health outcomes. This creates a comparative effectiveness model, which could then be expanded on in the economics section, to be the basis of a cost-effectiveness model. The model, or narrative synthesis of clinical utility, should capture the trade-off inherent with test and subsequent decisions. It should also identify crucial areas of uncertainty in the existing data where more primary data collection is required (Pletcher & Pignone 2011).
- Include a summary of the health impacts on patients by test status (positive and negative) including those that are misclassified (false positive and false negative).
- Present a table of the key uncertainties for each evidentiary step, and an overall assessment of the quality and certainty of the evidence. While this table is a culmination of the summaries described above, it is reasonable to present the table in the assessment report prior to the longer summaries.

Evidence component of the assessment	Interpretation and key uncertainties
Test accuracy	
Change in management	
Health outcomes	
Safety of the test	
Safety of the treatment	
Overall assessment of the evidence	

Provide a summary of the overall evidence base (without repeating evidence from other sections). Consider:

- the level of the evidence, taking account of the directness of the comparison
- the quality of the evidence
- the clinical importance and patient relevance of the effectiveness and safety outcomes
- the statistical precision of the evidence
- the size of the effect
- the consistency of the results across the clinical studies and across subgroups
- the strength or certainty of the evidence
- the applicability of the evidence to the Australian setting
- any other uncertainties in the evidence, including missing outcomes or populations
- other relevant factors that may have an influence on decision making, particularly implementation and ethical factors.

TG 16.2 Conclusion of clinical utility

The interpretation of the clinical data presented in Section 2 is crucial in determining the success of the submission. It is important to classify the health outcomes of the proposed health technology in relation to its main comparator (ie whether it is superior, inferior or non-inferior to the comparator).

The conclusion of the clinical utility of the investigative technology should be a simple and unequivocal statement that is supported by evidence provided in the submission.

Example:

The use of [proposed health technology] results in superior/non-inferior/inferior effectiveness compared with [comparator].

The use of [proposed health technology] results in superior/non-inferior/inferior safety compared with [comparator].

Section 3 Economic evaluation

Introduction

In Section 3, an economic evaluation of substituting the proposed health technology for the main comparator in the context of the listing requested should be presented. Assessment Reports should present a full and transparent description of the economic evaluation, with sensitivity analyses to characterise the uncertainty around the results.

Separate guidance is provided whether the economic evaluation to be presented is a cost-effectiveness analysis (CEA) (including cost-utility analysis, CUA) **(Section 3A),** or for when a cost minimisation approach is to be taken **(Section 3B)**.

A cost-minimisation approach should only be used when the proposed service has been demonstrated to be no worse than its main comparator(s) in terms of both effectiveness and safety, so the difference between the service and the appropriate comparator can be reduced to a comparison of costs. However, as there may be uncertainty around such a conclusion, MSAC may subsequently request cost-consequences, CUA and/or CEA to be presented. Further, there may be circumstances where a clinical claim of non-inferiority is made in Section 2, however other supportive factors (Technical Guidance 28 and Technical Guidance 29) may be present which may justify an increase in costs to the health system. Under such circumstances, a CEA and/or CUA could be presented to support the proposed increase in costs.

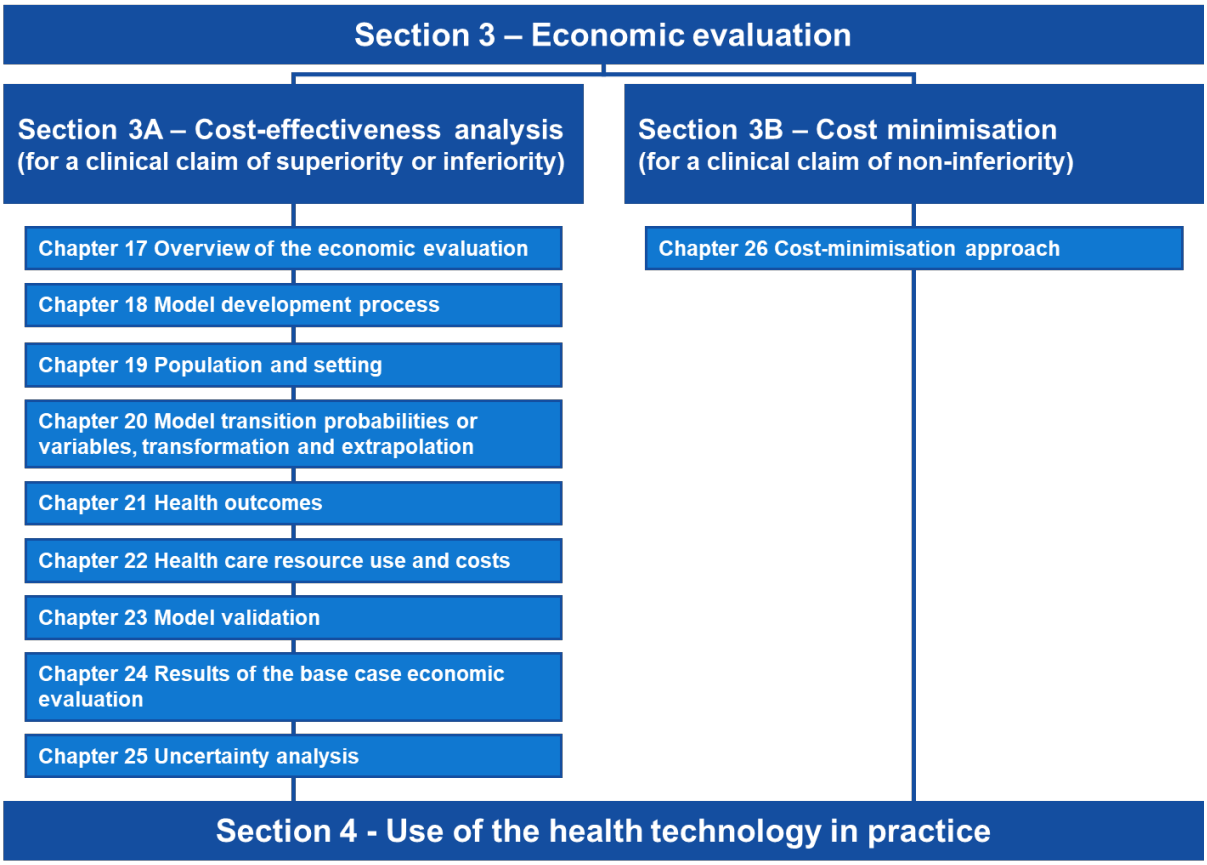
In circumstances where a high degree of uncertainty is present in the clinical assessment, such that an economic evaluation would also be associated with a high degree of uncertainty (and so may have limited usefulness to MSAC), this should be raised with the Department as soon as possible during the development of the Assessment Report. In these cases, the value of information from a model-based economic evaluation may diminish due to the quality of underlying data and greater uncertainties introduced through the process of modelling. Progression through modelling steps should continue only as long as the results generated are likely to be of value and informative to MSAC.

In the unusual circumstances where the proposed health technology is indisputably demonstrated to be therapeutically inferior, an economic evaluation may not be required as MSAC is unlikely to recommend Government subsidy of the service. However other supportive factors (see Technical Guidance 28 and Technical Guidance 29) may be present and an economic evaluation may be useful in circumstances where a therapeutically inferior service is funded at an overall lower cost to the healthcare system.

The most useful presentation of results from the economic analysis might vary with the level of evidence available. For example, in some circumstances, the evidence base might be weak (e.g. where a claim that a service is safe and 'promising' in terms of effectiveness is based on low-level evidence, such that the claim cannot yet be considered proven). In such cases, a threshold analysis that examines incremental cost-effectiveness over a range of possible benefits, might be more informative than reporting of an incremental cost-effectiveness ratio (ICER) based on a single point-estimate of incremental effectiveness.

The objective of cost-effectiveness analysis should be to provide an unbiased, plausible estimate of the incremental cost-effectiveness of the proposed health technology. Where there is considerable uncertainty around an assumption or the value of a parameter, a relatively conservative approach should be used in the base case analysis.

Flowchart 3.1 Structure of the economic evaluation guidance

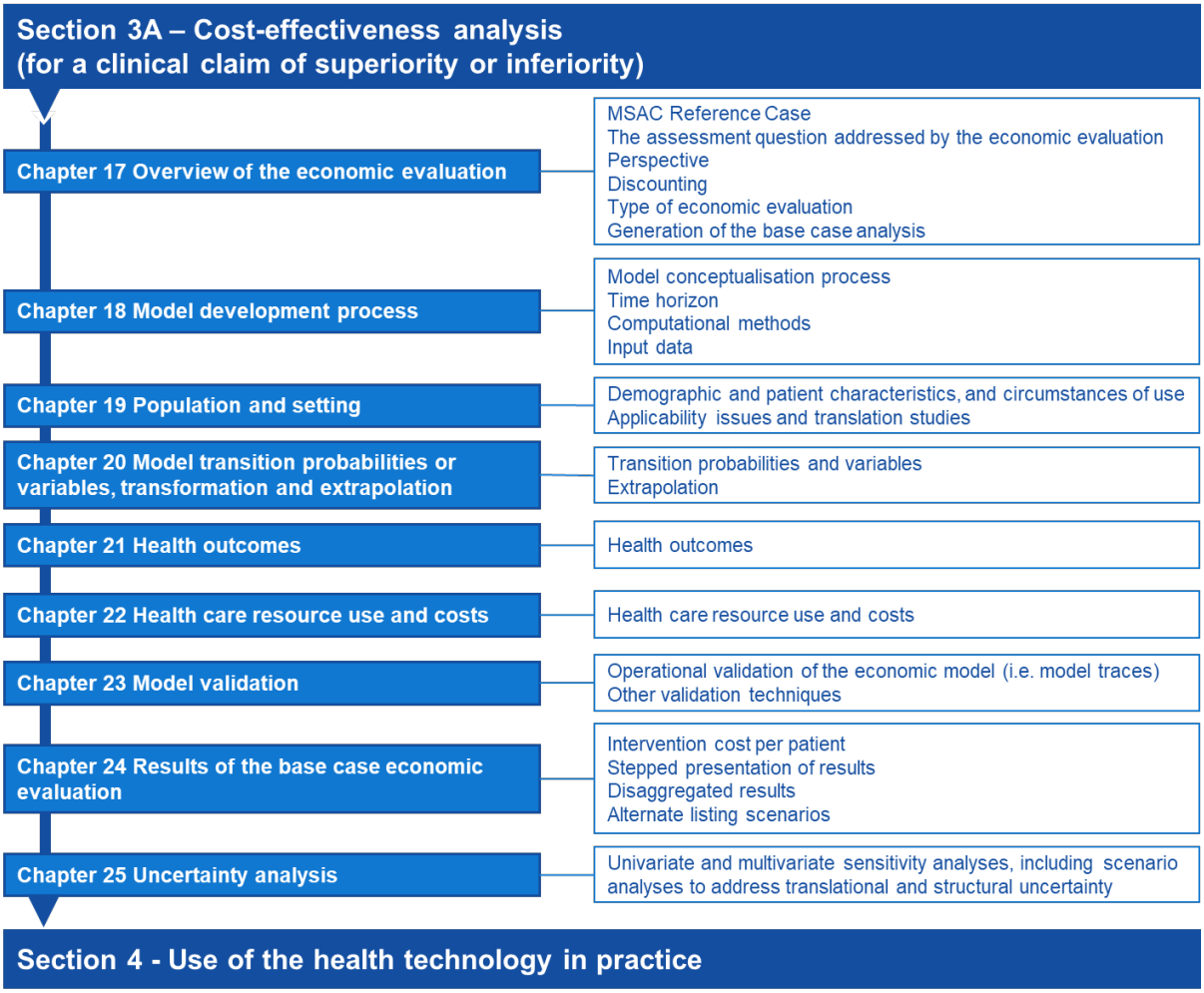


Section 3A Cost-effectiveness analysis

Section 3A provides guidance for preparing a CEA (including CUA).

MSAC prefers that the economic evaluation is based on results from direct randomised trials, with adjustments or additions to the trial data as required to account for differences in the population and setting, timeframe of analysis or outcomes of interest. Adjustments should be presented transparently in a stepped manner. For economic evaluations that rely on results from indirect comparisons of randomised trials, comparisons based on nonrandomised studies or linked analyses, an adaptation of the stepped approach is recommended.

Flowchart 3A.1 Summary of the cost-effectiveness analysis guidance



Technical Guidance 17
Overview and rationale of the economic evaluation

TG 17.1 The MSAC Reference Case

A reference case (Table 11) has been defined that specifies the preferred methods for economic evaluations to be presented to MSAC. These have been specified to promote consistency across economic evaluations of different technologies and disease areas.

Table 11 The MSAC reference case for economic evaluations

Component	Description	Relevant Guidance
The assessment question	As defined in the PICO Confirmation	Technical Guidance 17
Comparator	As defined in the PICO Confirmation (the currently available service that is most likely to be replaced by the new service)	Technical Guidance 17
Perspective on outcomes	Personal health of person receiving intervention	Technical Guidance 17
Perspective on costs	Healthcare system (health care costs incurred by the public or private (including patient) health care provider)	Technical Guidance 17
Type(s) of analysis	Cost-utility analysis, or a cost-effectiveness analysis where a cost-utility analysis is not feasible	Technical Guidance 17
Time horizon	Sufficient to capture all important differences in costs and outcomes between the intervention and the comparator	Technical Guidance 18
Source of effectiveness inputs	Derived from the systematic review conducted in Section 2, translated as necessary	Technical Guidance 19 and Technical Guidance 20
Measuring and valuing health effects	QALYs. However, where transformation to QALYs is not feasible, the outcome measure should be that which most closely and validly estimates the final health outcome from a patient perspective.	Technical Guidance 21
Evidence on resource use and costs	Where available, use the source of costs recommended in the *PBAC Manual of resource items and their associated costs.* However, for MBS-funded services, patient out-of-pocket costs, including average charges above the schedule fee, should be used where possible.	Technical Guidance 22
Discount rate	Annual rate of 5% for both costs and outcomes	Technical Guidance 17
Sensitivity analyses	Parameter uncertainty should be explored using deterministic (univariate and multivariate) analyses. Scenario analyses to address translational and structural uncertainty	Technical Guidance 25

QALY = quality-adjusted life year

Where non-reference case methods or analyses are relevant, it is preferred that these be presented as supplementary analyses. If non-reference case methods are used in the base case analysis, these should be clearly specified and justified.

TG 17.2 The assessment question addressed by the economic evaluation

Present a clear statement of the assessment question the economic evaluation aims to address, which defines the interventions being compared and the relevant patient group(s). This should be

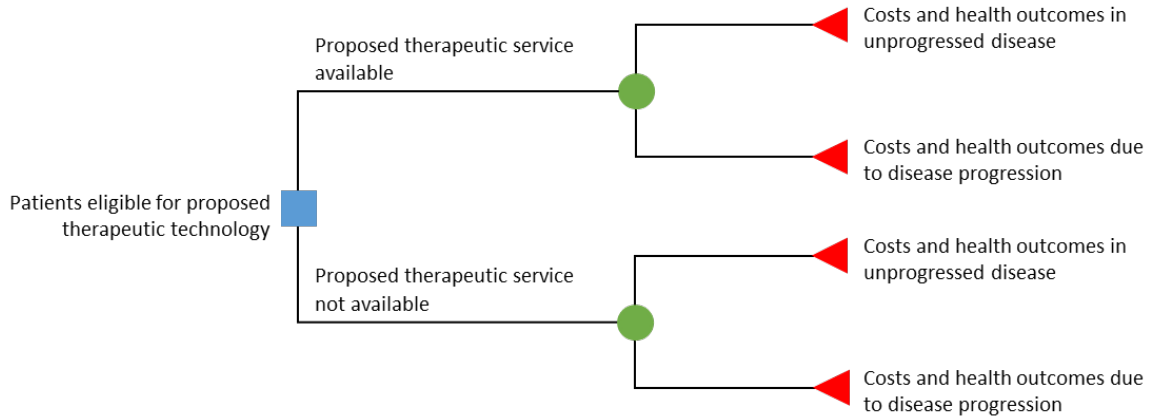
consistent with the PICO Confirmation. Any differences from the PICO Confirmation must be clearly presented and justified.

A decision-tree diagram may be presented which characterises the primary decision that the economic evaluation addresses, based on the information created in response to Technical Guidance 2 of the Guidelines. Use this diagram to provide a conceptual overview rather than the complete computational structure of the economic model. After the decision point of the tree, define alternative choices, uncertain events and outcomes. For investigative technologies, include the diagnostic decisions and outcomes, where relevant. Where the model is particularly complex, collapse and summarise branches, and clearly indicate where this has been done. Detail collapsed branches or a more suitable complete diagram of the model structure (eg a health state transition diagram) in Section 3A.1.2 of the Assessment Report.

Ensure that the pathways depicted in the decision tree are consistent with the existing and proposed clinical management algorithms presented in Section 1 of the Assessment Report. Cross-reference to these diagram(s) if they sufficiently represent the decision analytic of the economic model.

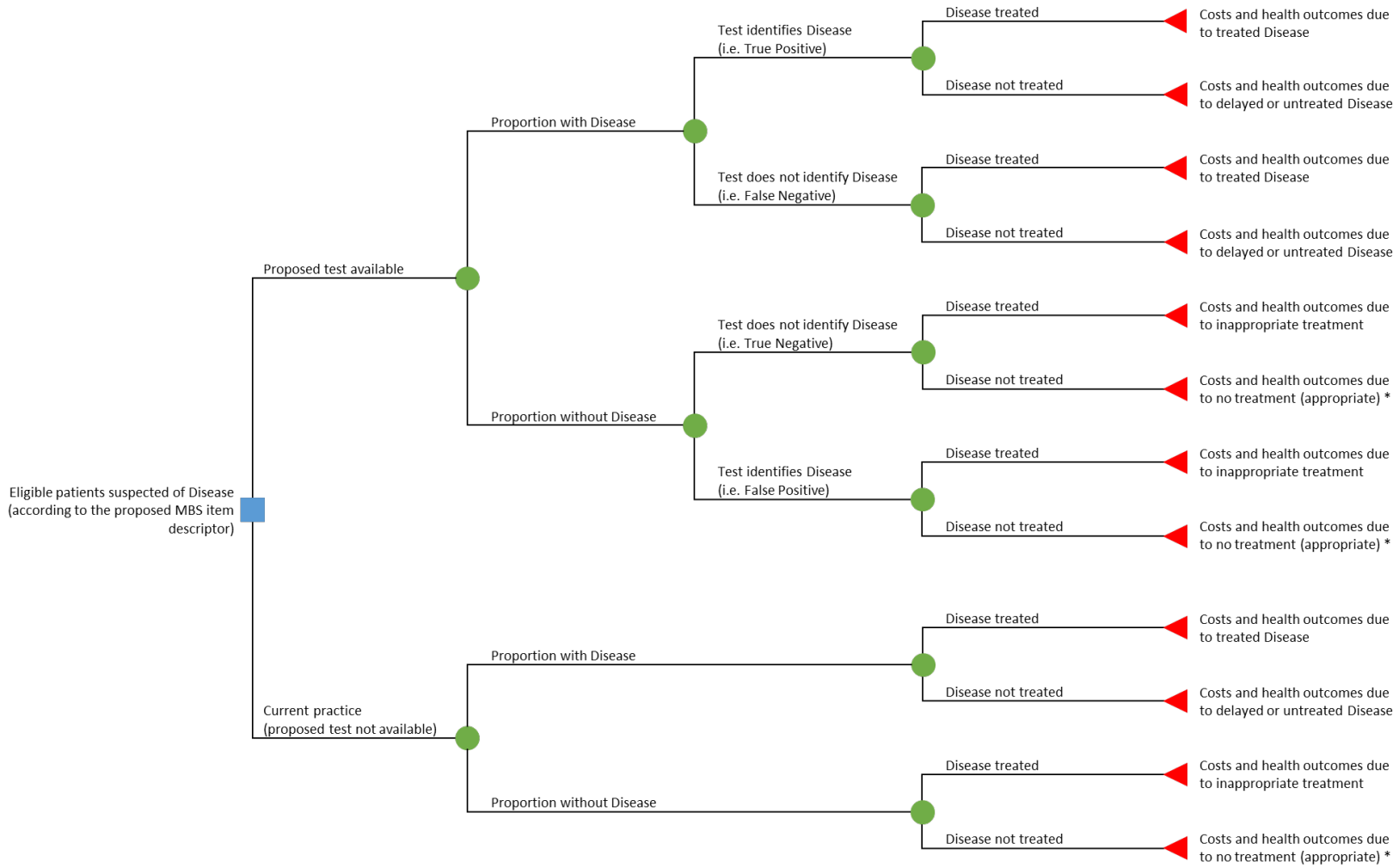
Examples of decision-tree diagrams are presented in Figure 23 and Figure 24.

Figure 23 Decision-tree diagram conceptualising the assessment question of a therapeutic technology



Note: While the conceptualised structure is the same across the arms of the economic model, with the proposed health technology, a reduction in disease progression is expected based on the results of the clinical assessment.

Figure 24 Decision-tree diagram conceptualising the assessment question of an investigative technology



* where no treatment is appropriate, this could include further investigations to diagnose underlying cause of disease, but this might not necessarily be quantified.
Note: While the conceptualised structure of the assessment question is similar across the arms of the economic model, with the proposed health technology, an increase in the proportion of appropriate treatment decisions is expected based on the results of the clinical assessment.

TG 17.3 Perspective of the economic evaluation

The perspective preferred by MSAC is a health care system perspective which includes health and health-related resource use (costs and cost offsets), and health-related outcomes. Health care costs include those incurred by the patient, and the public or private health care provider; health outcomes are those associated with the patient. For investigative technologies, this includes both the benefits and harms directly related to the service (eg an adverse event due to exposure to an imaging contrast agent) and those indirectly related to the service (such as those that arise from subsequent changes in treatment). Do not include costs and outcomes that are not specifically related to 'health and/or provision of health care' in the base case (see Technical Guidance 21 and Technical Guidance 22).

Where a broader societal perspective is relevant, quantitatively incorporate considerations beyond the patient and the health care system in a supplementary analysis. A well-justified and well-supported analysis will form a more compelling case (see Technical Guidance 21 and Technical Guidance 22 for the identification, measurement and valuation of non-health outcomes and costs, respectively).

Supplementary analyses may be appropriate where the proposed intervention has important societal implications extending beyond the health outcomes of the patient receiving the proposed health technology, and beyond the health care system. For example, costs/savings or socially relevant outcomes in domains such as education, housing or justice, or economic productivity impacts (see Appendix 10). Also, in circumstances where the beneficiaries of health or other relevant outcomes are broader than the treated patient population (eg community, carers, dependants), these generally should be included as supplementary analyses. However, where important and relevant, the omission from the base case of these costs and outcomes should be drawn to the attention of MSAC.

TG 17.4 Discounting

The values of costs and benefits incurred or received in the future are generally discounted to reflect the present value. Discount both costs and outcomes at a uniform, annual (compounding) rate of 5% per year for all costs and health outcomes that occur or extend beyond one year in the base case.

Present sensitivity analyses using fixed discount rates of 3.5%, and 0% per year (applied to both costs and outcomes). If relevant, present supplementary analyses using other discounting methodologies (eg a different uniform rate, differential rates, time-varying rates) and justify the alternative approach.

TG 17.5 Type of economic evaluation

State whether a CUA and/or CEA will be used. Identify the incremental health outcomes (quality-adjusted life years [QALYs] for a CUA or as nominated for the CEA) and incremental health costs. If no single outcome measure can be presented that appropriately captures the overall health of the patient or when the evaluation of the wider benefits of a technology is more useful, then the presentation of a cost-consequences analysis (CCA) in the base case may be reasonable.

The various types of economic evaluations are not mutually exclusive and more than one analysis can be presented to make a stronger case for cost-effectiveness (eg both a CUA and a CEA, or CUA and a CCA). (See HTA Glossary of terms[m] for definitions).

A cost-benefit analysis should not be presented in the base-case analysis.

Cost-utility analysis

A CUA presents the health outcome in terms of QALYs that represents society's preferences for the health outcome experiences relative to full health (i.e. QALY).

A CUA is preferred over a CEA, particularly where:

- there is a claim of incremental life-years gained in the economic evaluation (to assess the impact of quality adjusting that survival gain)
- there is an improvement in quality, but not quantity, of life
- relevant direct randomised trials report results using a multiattribute utility instrument (MAUI).

Where transformations or external data sources are required to estimate QALYs, present a stepped transformation from a CEA to a CUA, to transparently indicate the implications of the transformation and/or use of external data.

Other relevant factors, including prognosis, severity, age, distributional effect, context (eg emergency or prevention), and other equity and ethical issues that are ignored in measurements using a MAUI, should be considered alongside, not within, a CUA. Where this is important and relevant, the Assessment Report should draw these issues to the attention of MSAC.

Cost-effectiveness analysis

A CEA measures the incremental cost per extra unit of health outcome (expressed in natural units such as life years) achieved. Where a CEA is presented as the primary economic evaluation, justification should be provided as to why the quantified health outcomes are not translated into QALYs and presented as a CUA.

Ensure that the incremental health outcome (eg life-years, accurate diagnosis or other health events) presented in a CEA is patient-relevant. Present the outcome measure that is most closely and validly representative of the overall health of the patient, from their perspective, and in the context of the disease or condition for which they are receiving the proposed health technology. For investigative technologies where there is a personal utility with knowledge of the test result, such as those that claim to assist with reproductive planning, outcomes could include couples at-risk identified, or couples whose risk status is identified. Justify the choice of outcome and describe the extent to which the outcome captures all relevant health considerations.

Where a combination of outcomes (either intermediate or final outcomes, or both) are relevant to the patient, capture these collectively. Ideally, these would be transformed into QALYs and combined in a CUA, rather than presenting cost-effectiveness analyses for multiple outcomes. Where this is not possible, additional cost-consequences analysis may be useful.

[m] www.pbs.gov.au/info/industry/useful-resources/glossary

Cost-consequences analysis

A cost-consequences analysis compares the incremental costs of the proposed health technology with the comparator, and presents the various incremental differences in a range of relevant (disaggregated) outcomes. A cost-consequences analysis can be useful where the proposed health technology is demonstrated to have a different profile of effects that are not adequately captured by a single outcome measure, and where there might be trade-offs in effectiveness and safety between the intervention and the comparator.

Generally, a cost-consequences analysis should not be presented on its own, but it may be useful as a supplementary analysis to a CUA or a CEA. Disaggregated analyses may provide transparency in identifying changes in patterns of health care resource provision or specific health outcomes of interest that are not obvious in an aggregated evaluation.

Cost-benefit analysis

Cost-benefit analysis does not incorporate the breadth of considerations that are relevant to MSAC decision making, and there are limitations to the process of eliciting monetary valuations of health, particularly in the context of the Australian health care system where individuals do not face market prices. A cost-benefit analysis is unlikely to be helpful to the MSAC decision-making process.

TG 17.6 Generation of the base case

Within-study economic evaluation

A trial-based evaluation is sufficient to provide the base case of the economic evaluation if the evidence presented in Section 2 of the Assessment Report:

- recruited patients who are representative of those for whom listing is sought
- tested the proposed health technology in the circumstances of use expected to apply to the requested MBS listing
- directly measured and reported patient-relevant end points over an appropriate time horizon.

Modelled economic evaluation (including stepped adjustments to a trial-based evaluation)

If evidence for clinical effectiveness was synthesised across multiple sources or if the trial(s) did not provide evidence that sufficiently measures the full clinical and economic performance of the proposed health technology compared with its main comparator in the Australian setting, use modelling and/or adjustments to the trial data to generate the base-case economic evaluation.

Justify and make transparent any translations of the primary effectiveness data and additional assumptions used in the model. Construct economic models in a way that allows the results to be presented sequentially before and after key translational steps.

The stepped approach may include some or all of the following stages:

- Present the outcomes and costs as identified in the key trial(s) (see Technical Guidance 21 and Technical Guidance 22).
- Adjust treatment effects on health care resource use and health outcomes, as would be anticipated in the Australian setting and MBS population according to the proposed item descriptor (see Technical Guidance 19). This may involve one or more steps – for example:
  - re-estimate the treatment effect in the MBS population (eg use selected subgroups or weighted trial outcomes to improve applicability to the Australian demographic)

  - incorporate Australian circumstances of use or clinical practice (eg with respect to patterns of resource use)
  - incorporate other necessary and justifiable assumptions to improve the representativeness of the model (eg incorporation of resource use or outcomes associated with adverse event data, or subsequent treatment lines that are not captured in the trial data or previous translations).
- Extrapolate health care resource use and health outcomes (for the proposed MBS use) as required over the appropriate time horizon (see Technical Guidance 20).
- Transform health outcomes, if necessary, to the final outcomes used in the economic evaluation (eg using utility weights to obtain QALYs) (detailed in Technical Guidance 21).

The stages included in the stepped approach may vary depending on the nature of the available data. The base-case result is represented by the final incremental costs, outcomes and incremental cost-effectiveness ratio after the evidence from the main trial(s) has been translated.

For investigative technologies, this may include sequential incorporation of the evidence from each step of the linkages (eg assuming perfect test accuracy and change in management due to test result in the first step, then sequentially relaxing these assumptions). This enables MSAC to identify which steps of the linked evidence the cost-effectiveness of the test is most sensitive to.

A table should be presented in the Assessment Report that summarises the steps undertaken in the economic analysis.

Technical Guidance 18 Model development process

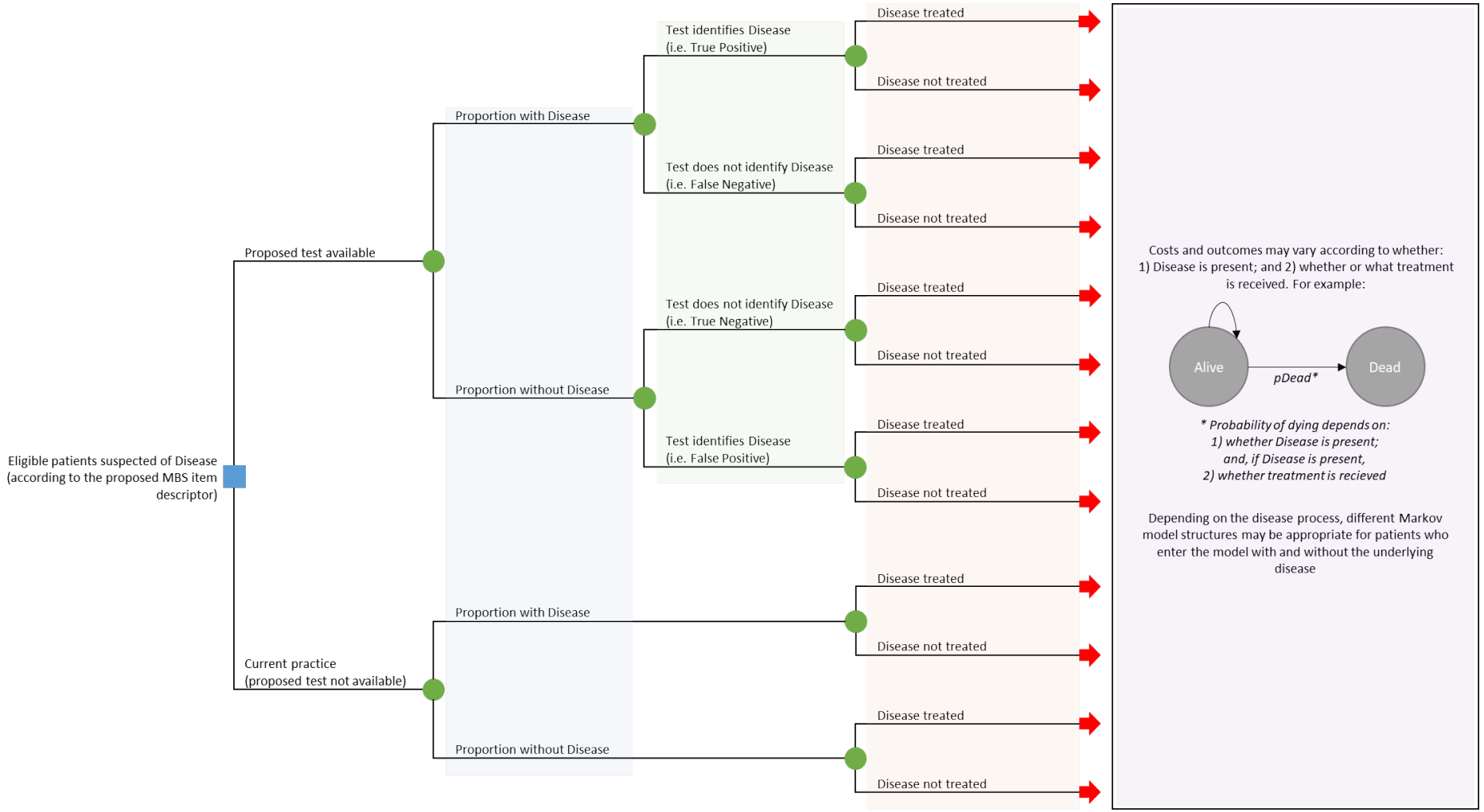
The model structure should capture all relevant and important health states or clinical events along the disease or condition pathway, and should be consistent with the treatment and disease or condition algorithms created in response to Technical Guidance 2 of the Guidelines. For investigative technologies, the model structure may need to account for prevalence of disease (or risk stratification for a prognostic technology), test accuracy (including cost and health outcome implications for patients who receive a false result or those in whom testing fails), change in management and effect of change in management, where relevant (see Figure 25 for an example of a model structure for a diagnostic test).

For evaluations in which multiple distinct populations are reasonable to model, such as where a test identifies multiple distinct diseases with differing treatments and prognoses, the model structuring process should be performed for each distinct population (with some indication as to how these are, structurally, combined). In such circumstances, the model should be structured such that the cost-effectiveness of the health technology can be determined both disaggregated and aggregated across the populations. In the case of testing for heritable diseases, the model should be structured so as to allow the cost-effectiveness to be explored under incremental expansion of the test from index cases, to index and first-degree relatives, and to index and first- and second-degree relatives, etc as considered relevant by PASC.

The model structuring process should be clearly and transparently described. This process includes: model conceptualisation, choice of computational method and consideration of other structural assumptions (Gonzalez-McQuire et al. 2019; Haji Ali Afzali, Bojke & Karnon 2018; Haji Ali Afzali et al. 2019; Kaltenthaler et al. 2011; Roberts et al. 2012; Tabberer et al. 2017; Tappenden & Chilcott 2014).

Assumptions incorporated into the model structure should be explicitly specified, with an indication as to how these have been tested in sensitivity analyses (see Technical Guidance 25).

Figure 25 Example model structure for an investigative health technology



Note: Blue shaded area denotes inputs related to prevalence. Green shaded area denotes inputs informed by the analytic validity. Orange shaded inputs denotes inputs informed by evidence of change in management. Purple shaded area denotes inputs informed by evidence for the effect of the change in management.

TG 18.1 Model conceptualisation process

The model conceptualisation process should be clearly described. The process should be driven by the assessment question, rather than by data availability. The following summarises this process to aid a transparent approach to model conceptualisation.

Literature review

Present the results of a literature search for economic evaluations of similar decision analyses (in terms of similarity to the treatment algorithm and/or the proposed and similar health technologies), focusing on the structure of the existing models. This may include Public Summary Documents or other reports of similar technologies previously considered by MSAC and models considered by other health technology assessment agencies.

Present any additional literature (eg additional clinical trials, clinical guidelines, natural history studies, burden of disease studies, surveys) that informs the model structure and that has not already been presented in Sections 1 or 2 of the Assessment Report. Provide copies of the original sources of all data not already presented in Section 2, or expert opinion used in the model, in an attachment.

Conceptual model

A figure depicting the conceptual model should be clearly presented. This should include all clinically-relevant and significant health states/events which were identified from the review of the literature. Significant health states/events are defined with respect to the strength of relationship between the condition of interest and the health state/event, as well as their potential impact on associated costs and/or economically important health outcomes such as QALYs. The health states/events should be disaggregated where there are likely to be important differences between the disaggregated states/events with respect to disease progression, associated costs, or associated health outcomes (eg QALYs). These health states/events should form the basis of the model structure used in the economic analysis presented.

The review of the literature should also identify whether there are important patient attributes that may influence the risk of experiencing subsequent events or disease progression, as this may inform the choice of computational method (see TG 18.3).

Final model structure

The conceptual model should be reviewed within the context of the available data. If adequate input data are not available to populate the model as conceptualised, alternate model structures should be identified that better conform to the available data. The face validity of these alternative model structures should be assessed (Vemer et al. 2016). The adaptations to the final model structure should be clearly justified and described, with any potential effects of these adaptations on the model outputs discussed.

If multiple plausible model structures are identified (eg alternative health states/events), these should be clearly presented and tested as part of structural sensitivity analyses. The impact of these alternative plausible structural assumptions on model predictions should be clearly discussed (see Technical Guidance 25).

Other structural choices/assumptions

Other structural assumptions used in the model should be fully documented and justified with an indication as to how these have been tested in structural sensitivity analyses. Examples other structural assumptions reported in the literature include (Haji Ali Afzali, Bojke & Karnon 2018):

- the relationship between time and transition probabilities including time dependency of probabilities (eg if an event is more likely to occur with time in a given health state)
- which model transitions the proposed health technology has an effect on (such as affecting transitions related to the initial incidence of disease or an event, but not affecting transitions subsequent to this, once the disease or event has occurred)
- the duration of treatment effects beyond the observed period in the empirical data (eg is the treatment effect assumed to continue beyond the observed period)
- the choice of statistical method for estimating health outcomes beyond the empirical data (see Technical Guidance 20).

TG 18.2 Time horizon of the evaluation

Define and justify the time horizon over which the costs and outcomes of the proposed health technology and its main comparator are estimated. Ensure that the time horizon captures all important differences in costs and outcomes between the intervention and the comparator, as a result of the choice of treatment, but does not extend unnecessarily beyond this. The same time horizon should be used for both costs and health outcomes.

Where interventions do not affect mortality and have temporary health or quality-of-life effects, a relatively short time horizon may be appropriate.

Where there is evidence that a health technology affects mortality or long-term/ongoing quality of life, then a lifetime time horizon is appropriate. Note that a lifetime time horizon relates to the life expectancy of the relevant patient population and reflects the time span required for nearly all of the model cohort to die. Consideration should be given where the patient cohort initiated in the model has a broad distribution of ages or prognoses; and the impact of this distribution on the model time horizon should be explained. The validity of the lifetime horizon is determined by the population of the model, and the inputs; it is not an independently nominated duration. Inputs that are not realistic will result in a model predicting an implausible duration of outcomes or survival and, thus, an implausible lifetime time horizon. The assessment of plausibility is also critical when considering how the model extrapolates data to reach the nominated lifetime time horizon (see Technical Guidance 20).

As a modelled time horizon extends – in absolute terms and relative to available data – it is associated with increasing inherent uncertainty. Therefore, economic claims based on models with very extended time horizons and predominantly extrapolated benefits will be less certain and are likely to be less convincing to MSAC. Technical Guidance 20 and Technical Guidance 25 address the extrapolation of costs and outcomes for an extended time horizon and associated uncertainty.

TG 18.3 Computational methods

If a trial-based economic evaluation is being undertaken using individual patient data on costs and outcomes from a clinical trial(s), describe the methods and software used to do this.

For model-based economic evaluations, identify the most appropriate modelling technique for the implementation of the final model structure(s) (Barton, Bryan & Robinson 2004). Generally, select the least complicated modelling technique for which it is feasible to implement the specified model structure, moving from decision trees to cohort-based state transition models to individual-level modelling techniques.

For some technologies (eg investigative), approaches that combine decision trees with other modelling techniques, such as cohort-based state transition models, might be appropriate.

Decision trees

Decision trees are useful for models with short time horizons. General spreadsheet software (eg Excel) or specialist software (eg TreeAge) can be used. Follow good-practice guidelines for using decision trees (Detsky et al. 1997).

Cohort-based state transition (or Markov) models

Use cohort-based state transition models to represent longer time horizons for models that can be represented using a manageable number of health states under the constraints of the Markovian (memoryless) assumption. General spreadsheet software (e.g. Excel) or specialist software (e.g. TreeAge) can be used.

Follow good-practice guidelines for using state transition models (Siebert et al. 2012). In particular, consider the following questions when implementing a cohort-based state transition model:

- Is it reasonable to assume that transition probabilities from each defined health state are independent of states that may have been experienced before entering each state? Health states that describe pathways through the model can be used to represent the effects of previous events on subsequent transition probabilities.
- Do transition probabilities vary according to how long individuals have remained in each health state? Tunnel states may be required to represent time-varying transition probabilities.
- Is the eligible population homogeneous, or is variation in patients normally distributed? This issue commonly refers to the age of the eligible population, but may include other factors. If relevant factors are not normally distributed, run separate analyses of the model and aggregate the outputs, or consider using a microsimulation model.
- What is the likely impact of alternative cycle lengths on the model outputs? Describe the factors determining the selected cycle length.

A half-cycle correction is the default approach to representing the time of transition between states, although an alternative correction factor may be proposed with justification.

Partitioned survival analysis (or area under the curve modelling)

Partitioned survival analysis models are conceptually similar to Markov models in that they are characterised by a series of health states with associated state values. However, health state membership is not estimated using transition probabilities; rather, it is derived from a set of independently modelled non-mutually exclusive survival curves (eg overall survival, progression-free survival) (Williams et al. 2017; Woods et al. 2017). Where effectiveness outcomes are reported using non-mutually exclusive survival curves, a partitioned survival approach may be used. Depending on the maturity of the data available, statistical extrapolation beyond the observed data may be required to model outcomes to the nominated time horizon (see Technical Guidance 20).

Where a partitioned survival analysis approach is used, justification for the key structural assumptions associated with this approach should be provided (ie that all endpoints, including overall survival, are modelled and extrapolated independently and that transitions between health states are not explicitly modelled) (Woods et al. 2017).

Individual-level (or microsimulation) models

Use individual-level modelling approaches only when a defined model structure cannot be feasibly implemented as a cohort-based model. Describe the characteristics of the model structure that prevents a cohort-based model being used. Potential factors include baseline heterogeneity,

continuous disease or condition markers, time-varying event rates and the influence of previous events on subsequent event rates (Karnon & Haji Ali Afzali 2014). Also describe how incorporation of these features in an individual-level model are expected to produce a more accurate representation of the disease or condition pathways, costs and patient outcomes.

The most common individual-level approaches include individual-based state transition and discrete event simulation models. Follow published guidelines on good research practices for applying these models (Karnon et al. 2012; Siebert et al. 2012). Discuss any requirements for specialist software with the Department in advance.

Other modelling techniques

If the results from simpler models are robust enough to produce plausible sensitivity and scenario analyses, it is not necessary to use more complex modelling techniques (Tsoi et al. 2015). If an alternative modelling technique is used, describe and justify how the approach leads to more accurate and valid results. For example, in the clinical area of infectious diseases, the use of dynamic transition models or agent-based models to represent herd immunity may be justified if a simple nondynamic model will not demonstrate cost-effectiveness accurately enough.

Note that more complex modelling techniques may be less transparent, and the model assumptions less certain. This might result in MSAC having less confidence in the cost-effectiveness claim. Discuss the use of complex modelling techniques (including any specialist software) with the Department in advance.

TG 18.4 Input data

Where possible, input data should be sourced from the evidence presented in Section 2 of the Assessment Report. At the 'source-of-evidence' level, identify which model inputs are derived from the clinical evidence presented in Section 2, and which were derived from alternative data sources. Where multiple sources of data were identified in Section 2 to inform a particular parameter, the justification for the input used in the base case analysis should be provided in the relevant subsection of the Technical Report, with an indication as to which alternative sources of input data are used in sensitivity analyses.

Justification to support the approach used in the economic evaluation is required if one or more studies presented in Section 2:

- Had reliability issues (due to inadequate concealment of randomisation, inadequate blinding of subjective outcomes etc);
- Reported fewer or no patient-relevant outcomes;
- Were of insufficient duration to detect the most patient-relevant outcomes; or,
- Reported outcomes that could not be translated into the economic anlaysis.

Where relevant, applicability issues with clinical data from Section 2 are identified, these are discussed and translated to the Australian population and setting, if necessary, in Section 3A.1.3 of the Assessment Report.

Describe the methods used to identify data beyond the clinical evidence identified in Section 2 to populate the model input parameters. For example, whether systematic or ad hoc reviews of the literature were undertaken, or how relevant primary data sources, including registries and observational studies, were identified. The method of identifying the data should be robust and transparent. Where multiple sources of data exist, the source of the input used in the base case should be described and justified.

Applicability concerns (and any translation) relating to additional data should be described in the relevant subsection of the Technical Report.

TG 18.5 Fully editable electronic copy of the economic evaluation

Provide access to the electronic copy of the economic evaluation. The economic evaluation should be constructed in-line with best practices (Ghabri et al. 2019). Ensure that all variables can be changed independently, including allowing the base case of the economic evaluation to be respecified and a new set of sensitivity analyses to be conducted with each respecified base case. Ensure that the economic evaluation can produce results following respecification of variables within reasonable running times. To help understand the electronic copy of the economic evaluation, apply clear and unambiguous labels to values, and cross-reference data sources.

The following software packages do not need prearrangement with the Department:

- TreeAge Pro
- Excel, including @RISK®, but not necessarily including all advanced features and plug-ins (eg Crystal Ball).

Use of other specialist software must be prearranged with the Department in advance of submission.

Technical Guidance 19 Population and setting

TG 19.1 Demographic and patient characteristics, and circumstances of use

The setting of the economic evaluation should be the Australian health care setting, with the modelled population representing the target Australian population indicated for use of the proposed health technology, and the circumstances of use consistent with the clinical management algorithm and the indication specified in the proposed item descriptor (Section 1).

Describe the demographic and clinical characteristics of the modelled population using summary statistics, including information on distributions around the central estimate (eg standard deviations, confidence intervals). Relevant patient and clinical characteristics may include age, sex, ethnicity, medical condition and severity of the medical condition, and comorbidities. Indicate which patient characteristics are incorporated explicitly and which are implicit (associated with use of other data) or not included.

For investigative technologies, the modelled population includes all patients eligible for the test – not just those that the test aims to identify. The prevalence(s) of the target (eg disease, subtypes or pathological variant(s), etc) in the tested population, used in the model should be reported and should be consistent with that identified in Section 2.

Describe and justify how heterogeneity in patient characteristics (if relevant) is represented in the cost-effectiveness analysis. Heterogeneity could include where multiple distinct populations are identified by a test (and so multiple indications are modelled), or when the test identifies a heritable disease, and so the eligible populations modelled include those suspected of being an index case, and the relatives of those in whom the heritable disease is identified.

Provide details of any additional circumstances of use relating to the proposed health technology that are relevant to the model setting or population, and detail how they are incorporated into the model. These may include:

- the position of the service in the overall algorithm for diagnosing, treating or managing the disease or condition (e.g. prevention, first-line treatment, second-line treatment);
- any limitations on the duration or frequency of delivery of the services; for example, in a 24-hour or in a 12 or 24-month period;
- any required co-delivered medical services or treatments (including any additional diagnostic tests required);
- any contra-indicated medical services or treatments;
- any unique characteristics of the referrer or provider (e.g. specific qualifications or training); and
- any specific requirements in terms of geography, facilities or location of delivery of service (e.g. limited to hospital setting or to approved laboratories; specification of any specific equipment or facilities that need to be available).

TG 19.2 Applicability issues and translation studies associated with the clinical evidence

For each difference between the clinical evidence setting(s) (including population and circumstances of use) and the Australian setting that are identified in the synthesis of the results in Section 2 of the Assessment Report as potentially important, design a translation study. These include factors relating to differences in the populations, disease or condition, circumstances or treatments as conducted in the evidence presented compared with what would be expected were the proposed health technology reimbursed according to the requested restriction and in accordance with the proposed clinical management algorithm. For investigative technologies, applicability issues arising

from a potential change in the spectrum of disease identified or the transitivity across the studies included in the linked evidence approach should be addressed.

Table 29, Appendix 6, contains a list of example factors that, when different across settings, may result in a difference in treatment effect, adverse events or patient management across those settings.

Each translation study should determine whether a quantitative adjustment to model inputs are necessary and, if so, the nature of the appropriate translation. Where there are inadequate data for a translation study, identify this as an issue that will remain a source of uncertainty in the model.

The translation study should include:

- the issue and the specific question to be addressed
- the data used and their sources (justify the choice of data where there are multiple possible sources)
- the methods of analysis, with sufficient details to enable independent verification of the analysis (common methods are described below)
- the results, which for therapeutic technologies might include an estimate of the comparative treatment effect (both relative and absolute) and the 95% confidence interval, and a description of how (or whether) the findings are applied in the model
- a description of any residual uncertainty, and sensitivity analyses that are proposed to address this uncertainty.

Take care when converting relative treatment effects or estimates of accuracy (ie use measures of sensitivity and specificity, rather than PPV and NPV) across jurisdictions with different baseline risks. Ensure that the baseline risk (ie prognostic characteristics) of patients does not differ between the trial evidence and the target population, or that patients are not expected to respond better to the proposed health technology or the main comparator in one setting than in another setting.

Common methods for translation include subgroup analyses, regression analyses, meta-regression or use of other published studies. Justify the selected approach.

Subgroup analysis

For subgroup analyses, follow the same methods outlined in Technical Guidance 6.

Regression or meta-regression

Regression analysis has an advantage compared with stratified analyses based on subgroups because it can examine more than one covariate (or difference between the clinical trial participants and the target MBS population) simultaneously. Where multiple trials are available, use a meta-regression, if appropriate. Meta-regression may be used at the study level or at the individual patient level (where the study is entered as a covariate). Only use a meta-regression at the study level if the number of trials is large (5–10 trials for each covariate examined).

Where a regression analysis is used, present and interpret the results in the main body of the Assessment Report, and provide the following additional details in an attachment:

- a clear description of the regression method, the associated assumptions, how these assumptions were tested and the results of the tests
- the statistical commands or syntax used in the analysis, with a description of the variables (including a description of the thresholds used to define categorical variables)
- the direct output from the statistical program

- the dataset used in the statistical program (or a justification, where this is not provided).

Published studies

If it is not possible to inform translation using the direct clinical evidence for the intervention, describe the reasons and seek relevant published data. Systematically identify published studies concerning the proposed health technology (or comparator) in the proposed eligible population. Present the search strategy and selection criteria in an attachment.

Report the relevant findings from the included studies. Describe the findings in relation to the proposed health technology and apply the findings to inform the translation.

Technical Guidance 20
Model transition probabilities or variables, transformation and extrapolation

TG 20.1 Transition probabilities and variables

Transition probabilities inform the movement of patients between health states in decision trees or state transition models. For investigative technologies, this also includes decision-tree parameters related to test accuracy and changes in clinical management, where relevant. In a discrete event simulation, time-to-event parameters are analogous to transition probabilities. Transition probabilities or time-to-event parameters may differ by treatment or by how long a patient has been in a particular health state (time-varying probabilities).

Transition probabilities that differ by treatment are generally estimated using the clinical evidence described in Section 2 of the Assessment Report (with applicability translation, as per Technical Guidance 19, as appropriate). Cross-reference the relevant subsections for the clinical evidence and note whether further translation studies or extrapolations are required.

Other transition probabilities may be required that describe the progression of a disease or condition following an intermediate modelled event, and for which the same transition probabilities are applied, regardless of treatment allocation. Where external sources of data (other than the clinical trials from Section 2) are used to inform transition probabilities (or other variables) in the model, assess the applicability of these sources of data with respect to the Australian setting. Note and justify whether the data are applicable, requiring translation (in which case, follow the approach detailed in Technical Guidance 19), or are a source of uncertainty within the model.

Detail where the model uses other variables instead of, or in addition to, transition probabilities, such as allocation to a medical management pathway, and justify the source of these input variables in the same manner. Do not include variables associated with the valuation of outcomes or costs; these are described in response to Technical Guidance 21 and Technical Guidance 22, respectively.

Describe and justify the methods used to identify and analyse relevant data to derive transition probabilities and variables.

For each transition probability or variable, present the point estimate and interval estimates (eg 95% confidence intervals). Follow good-practice guidelines when choosing the methods to derive interval estimates (eg using probability distributions based on agreed statistical methods for alternative types of input parameters) (Briggs et al. 2012). Ensure that values taken from all sources of evidence are appropriately adjusted to represent the transitions required by the model structure (Fleurence & Hollenbeak 2007). For example, translate reported rates or cumulative probabilities to the probabilities for timeframes associated with a model cycle, if necessary.

Occasionally, secondary outcomes and other trial-derived data (eg adverse event rates) are relevant to outcomes and/or resource use in the economic model, and point estimates are numerically different across the arms, but not statistically significantly different. This may reflect either no 'real' difference, or a difference but with insufficient power in the trial to demonstrate it statistically. Explain the approach used to inform the probability in the base-case model (eg whether it has been pooled across arms or differentiated between arms), and explain and justify with supporting evidence, if available. Examine the alternative approach in a sensitivity analysis.

Assess the potential correlation between transition probabilities and/or variables. Correlation between parameters is explored further in Technical Guidance 25 for uncertainty analysis.

TG 20.2 Extrapolation

Extrapolation may be justified when all important differences in costs and outcomes between the intervention and comparator(s) groups are not represented over the time horizon for which observed data are available. Detail any extrapolations of data that are required for the base-case economic model.

Extrapolating time-to-event data

Where extrapolation of time-to-event data is required, use observed time-to-event data in preference to modelled data up to the time point at which the observed data become unreliable as a result of small numbers of patients remaining event-free.

Describe and justify the selected time point beyond which extrapolated transition probabilities are applied. External data may be used to justify the selected time point – for example, the point at which one or more of the curves fitted to the clinical trial data deviates from a curve fitted to observational data from a similar patient cohort with a larger sample over a longer follow-up period. Test alternative truncation points in the sensitivity analysis.

Derive appropriately estimated parametric survival curves based on the observed data (using individual patient data, if available) to extrapolate transition probabilities beyond the data truncation point.

Detail each of the following:

- Whether an assumption of proportional hazards is appropriate beyond the observed data.
- Fit a range of alternative survival models to the observed data (eg exponential, Weibull, log-normal, log-logistic, gamma, Gompertz). Include more flexible extrapolation approaches with multiple points of inflexion (eg piecewise spline models) to better facilitate extrapolation based on the section of the Kaplan–Meier curve that is most representative of long-term survival (Royston & Lambert 2011).
- Assess and discuss goodness of fit using visual inspection, Akaike's information criterion and Bayesian information criterion. Justify the most appropriate model for the base case and test a number of the best-fitting models in the sensitivity analysis.
- The plausibility of the predictions in the unobserved period (eg the ongoing hazard ratio and/or treatment effect, the point of convergence and/or residual survival in each arm).

The treatment effect resulting from the independent extrapolation of the survival curves should be plotted over the time horizon of the model. If the treatment effect is maintained or increasing, and this is not clinically plausible, apply a hazard ratio such that the intervention and comparator curves converge at a plausible time point. The assessment of plausibility should be linked to the justification of the time horizon (see Technical Guidance 18).

When considering the extrapolated treatment effect, give explicit consideration to clinical decisions regarding the cessation or continuation of treatment. State and justify all assumptions in this regard, and apply them consistently when modelling respective treatment costs.

Numerous sources of advice on extrapolation techniques for economic evaluation are available in the literature (Bagust & Beale 2014; Grieve, Hawkins & Pennington 2013; Karnon & Vanni 2011; Latimer 2013; Royston 2001; Whyte, Walsh & Chilcott 2011).

Other individual patient extrapolation issues

For categorical data that describe the experience of multiple intermediate or outcome events, use a two-stage process of modelling the time to any event, combined with a multinomial logistic model

to define the probabilities of the aggregate event being each of the competing events. Include a time covariate in the multinomial logistic model to represent time-varying probabilities, if possible. The other option is to fit independent competing risks time-to-event models for each event, but this approach is likely to overestimate parameter uncertainty as a result of the assumed independence of the multiple events modelled.

For continuous variables, format the data into categories, or use a generalised estimating equation model.

Extrapolating published time-to-event data

If individual patient time-to-event data are not available, extrapolate survival probabilities from published Kaplan–Meier curves using graph digitiser software. Fit alternative constant (ie exponential), or monotonically increasing and decreasing (eg Weibull or Gompertz) hazard functions to the extracted survival data beyond the last point of inflexion to the time point at which the observed data become unreliable because of small numbers of patients remaining event-free.

Present tests of the relative and absolute goodness of fit of the alternative curves, and use the best-fitting curve in the base case. Test the alternative models in sensitivity analyses.

Use of data from other non-randomised studies to extrapolate beyond the evidence

Data from other non-randomised studies may sometimes be useful to extrapolate beyond the results of the clinical evidence presented in Section 2. This is because the included studies might have been of insufficient size or duration to capture the full impact of therapy on the outcomes of the disease, or the typical resource provision measured in an overseas trial might need adjustment to reflect patterns of resource provision in Australia. In contrast, other non-randomised studies might involve longer follow-up for an active main comparator, or the natural history of the medical condition if the main comparator is not an active intervention. Given that the data from non-randomised studies are subject to bias, assumptions based on those data made during a modelling exercise should be cautious.

When presenting data from other non-randomised studies for extrapolation purposes in a modelled economic evaluation, demonstrate that a systematic approach has been taken to search for, locate and select the non-randomised studies for presentation. The selection process should be presented and justified. Provide a report of each study in a technical document or attachment. The results of the non-randomised study might contribute to finding and justifying a variable in the economic evaluation. This variable might vary from a single point estimate to a regression formula. The results of the non-randomised study might also help identify risk factors that contribute to the expected risks of the comparator arm in a model.

When indicating which results are being extrapolated, explain how the extrapolations are achieved by the model for the streams of costs and outcomes for the proposed therapeutic health technology and the main comparator. In particular, if non-comparative data are used (e.g. from single-arm studies), it is necessary to make an assumption about how the other arm in the model would change. The usual practice, in the absence of empirical evidence to the contrary, is to assume that the comparator arm would change so that the relative risk between the two arms measured in the randomised trial(s) remains constant across the duration of therapy. Justify the use of this (or any other) assumption in the model presented in the assessment report.

Technical Guidance 21 Health outcomes

TG 21.1 Health outcomes

Nominate and justify the final health outcome that is considered to best reflect the comparative clinical performance of the interventions and will be presented as the denominator unit in the base-case ICER.

Detail the health outcome(s) (intermediate and/or final) that inform the final outcome in the economic evaluation and whether these were reported directly in the clinical evaluation (Section 2), and, if not, summarise the transformations involved to obtain the final outcome.

If available, use quality-of-life or utility data reported in Section 2 to estimate QALYs in the model, or, justify the use of alternative indirect methods to estimate QALYs when direct data are available. Present both sets of methods and results, and compare the interpretation.

Present the results of any utility study as the point estimate of the mean elicited utility weight for each health state, and include its standard deviation and 95% confidence interval, where available.

If a claim is made for a change in a non-health outcome, or the Assessment Report identifies health-related outcomes in people other than the patient receiving treatment (eg quality-of-life benefits for family, decreased carer burdens), these generally should not be included in the base-case evaluation; rather, these could be included in supplementary analyses (see Appendix 10).

Use of quality-of-life data from the clinical trials to estimate QALYs

Estimates of quality of life or utility from the evidence presented in Section 2 may inform direct estimates of QALY gains in the intervention and comparator populations or inform utility values applied to health states in a cost-effectiveness model.

If a MAUI has been used in a study included in Section 2 to estimate utility weights, state where and when the scoring algorithm was derived, and consider how applicable it is to the general Australian population. It is preferred that Australian-based preference weights are used in the scoring algorithm used to calculate utility weights.

If the initial patient-reported outcome measure is not a MAUI, provide detail of the measure and justification of its use in Section 2. Describe a validated method of mapping the results into preference weights (see below). State whether Australian-based value sets are incorporated. If there is no reliable method of transforming the patient-reported outcome data into utility weights for the model, describe why this is not possible and detail whether the patient-reported outcome data from the trial can still be used to inform or validate the economic model.

Consider the duration over which the patient-reported outcome measure informing utilities was administered compared with the duration of the condition of interest. If a generic MAUI or patient-reported outcome measure is used, consider whether it captures all important disease- or condition-specific factors that might be relevant.

Address the following questions when incorporating trial-based patient-reported outcome data into the economic model:

- Are the participants representative of the population for whom listing is requested? (Refer to Section 3A.1.3 of the Assessment Report, as needed.)

- If quality of life is not the primary outcome, is the trial adequately powered to detect a difference in the survey results? As with all secondary outcomes, assess the results with reference to the conclusion from the primary analysis of the trial.
- Is there a 'healthy cohort effect'? (ie where the sickest patients are least likely to complete patient-reported outcome data forms, and therefore the data obtained has a bias towards healthier patients). Consider the responder numbers and drop-outs. While generally associated with an overestimate of utility weights, the direction of any associated bias may depend on whether the treatment and comparator are associated with different utilities, the relative extent of the effect across different arms and health states, and the time spent in different health states. Identify any impact on the overall ICER.
- Is there potential for systematic bias where progressed health states are defined by nonsymptomatic events (ie identified by investigations that may or may not reflect clinical practice)? Provide details.
- Is it appropriate to pool patient-reported outcome data across arms of a trial? This may be appropriate where patient numbers are small and for posttreatment states, but not in other circumstances where treatment (rather than disease or condition) directly affects quality of life (eg because of serious adverse events and any associated long-term implications, or imposed limitations). Justify the approach, and, where possible, present results with and without pooling.
- Is there a risk of bias from a regression to the mean effect? (Barnett, van der Pols & Dobson 2005) This may be more likely in instances where quality of life for the control arm is drawn from a trial other than a randomised controlled trial (eg instance from a pre-intervention population).

Use of other sources of data to estimate utility weights

Where utility weights or QALY changes cannot be directly estimated from data collected in the clinical studies from Section 2, or there are significant concerns about the reliability and relevance of trial-based utility, transform the Section 2 health outcomes to estimate QALY gains (eg by applying utility weights to the time spent in different health states that represent the experience of clinical outcomes).

Additional studies (either published or commissioned for the Assessment Report) may be needed to estimate utility weights for health states in the economic model. These studies should be identified, with copies provided.

Describe the source(s) and method(s) (as described below) used to derive externally derived health state utilities and justify their inclusion in the model. Depending on the clinical context and available data, there may be more than one acceptable source of utility weights. Where this is the case, reflect the uncertainty in selecting an optimal source of weights by reporting the sensitivity of the result to switching between the various sources of weights.

Address the questions regarding quality-of-life data derived from the clinical trials (above) that are applicable to any utility estimates obtained from alternative sources and methods.

Mapping of generic and disease-specific scales

Non-preference-based patient-reported outcome measures will require a mapping algorithm to be transformed into preference-based measures to estimate utilities. Where this occurs, detail the source of the mapping algorithm. Describe the estimation sample (population demographic and clinical characteristics, sample size etc) and whether there is an external validation sample. Provide details of the source and target measures (eg index, dimensional), and the statistical model and methods used to estimate the mapping algorithm. Detail the statistical association or operations

that constitute the algorithm. Discuss methods used to measure the algorithm performance and validity. Present the resulting predicted utilities with associated uncertainty. Discuss the applicability to the data presented in the Assessment Report, particularly in relation to the sample in which the algorithm was developed.

Scenario-based methods to indirectly elicit utility weights

Scenario-based methods use vignettes to describe the symptoms of a health state to a sample population, usually a representative general population sample, from which utility weights are elicited using an accepted preference-based method. Methods to elicit preferences include the standard gamble, time trade-off and discrete choice experiments, and other stated preference methods.

If using a scenario-based utility valuation to value health outcomes beyond the time horizon of the trial, include one or more health states captured and valued within the trial in the scenario-based study to validate the commonality of the trial-based and scenario-based utility weights.

Present supporting evidence for any claim of increased sensitivity of a scenario-based approach to identify real differences in utility.

Describe all stages of a scenario-based study in detail and explain efforts to minimise potential bias. It is difficult to minimise the many sources of analyst bias that are intrinsic to the scenario-based utility approach, including the non-blinded nature of the construction and presentation of the scenarios (eg incomplete inclusion and differential focus on alternative aspects of quality of life), the design of the methods to elicit values, and the analysis and interpretation of the results.

Population matching study method to indirectly elicit utility weights

This form of utility study involves the recruitment of a separate sample of patients with characteristics similar to those enrolled in the clinical trials reported in Section 2. Matched patients complete a MAUI reflecting their current health state, which informs the estimation of utility weights for the health states in the cost-effectiveness model. See Technical Guidance 6 for further detail on MAUIs.

Potential sources of bias for such studies include the possibility of systematic differences between the clinical study participants and the matched patients, and the inability to blind the sampled patients from the objectives of the study. If there are important symptomatic toxicities, the sampled patients should possibly have been exposed to the health technology and its toxicities at the time the MAUI is completed.

Matched patients should complete other patient-reported outcome measures that were completed by the trial participants, and the results of this concurrent instrument should be used to more closely match utility study participants to the clinical study population.

Published sources of utility weights

Utility estimates may be available from the literature. The validity of the derived utility weights depends on the applied elicitation methods and the relevance of the study populations. Present details of search strategies, and inclusion and exclusion criteria used to identify relevant utility studies. Assess the validity of all identified studies, including (Brazier et al. 2019):

- how representative the health state in each identified study is of the health state in the economic evaluation (including the type and severity of symptoms, and the duration of the health state)

- how the health state was captured (eg MAUI, scenario based)
- how the preference was elicited (eg standard gamble, time trade-off)
- what sample was chosen to respond to the MAUI questionnaire or scenario (eg the general public, patients, carers, health care professionals)
- the country that the utility data were collected
- what assessment was made of the nature and direction of bias that might arise, given the sample and methods (report the variance in the utility estimates and response rates and extent of missing data or data lost to follow-up, or study type ie observational study or RCT)
- how the sensitivity analyses examined variation in the identified utility options.

The original published study for utilities should be cited and not a previous economic study that used this evidence.

Using different published studies to inform utility weights for alternative health states is discouraged because of the potential for inconsistency in the methods (eg instrument) and populations from which utilities were derived.

When estimating utilities for concurrent clinical events, multiple approaches exist, including:

- subtracting the sum of the estimated utility decrements for overlapping events from the estimated utility in the absence of an event (additive method);
- multiplying the utility in the absence of an event by the product of the ratios of the utility for individuals with the clinical events to the utility for individuals who do not experience the clinical events (multiplicative method);
- using the lowest utility for all the clinical events (minimum method).

Good-practice guidelines for using health state utilities currently recommend the multiplicative method (Brazier et al. 2019). Alternate approaches may be presented in supplementary analyses, if relevant.

Presentation of outcomes and health utility value information

If presenting a CUA, a format for summarising the minimum information on all modelled health outcomes (eg intermediate, final outcomes and events) contributing to the final health outcome in the economic evaluation, and any associated utilities or disutilities is suggested in Table 12.

Table 12 Identification of health outcomes used in the model

Health state or event	Mean utility (SD and/or 95% CI) or QALY	Nature of estimate and any translations	Source of estimate	Alternative estimates of utility value (and sources)	Average application in the model: proposed health technology	Average application in the model: comparator
[Health state 1]	[Utility estimates for health state 1]	[eg EQ5D data (Australian value set)]	[eg from Trial 001 (see Section 2)]	[eg nonpooled data from study]	[eg days/months]	[eg days/months]
[Health state 2]	[Utility estimates for health state 2]	[eg scenario-based study using standard gamble method]	[eg external publication: Smith et al 2010]	[eg external publication: Jones et al 2008]	[eg days/months]	[eg days/months]
[Event 1]	[x QALYs per event]	[eg scenario-based study using time trade-off method]	[eg commissioned study (study report provided in attachment)]	[eg external publication: Jones et al 2008]	[no. of events]	[no. of events]

CI = confidence interval; QALY = quality-adjusted life year; SD = standard deviation

Technical Guidance 22 Health care resource use and costs

TG 22.1 Health care resource use and costs

For within-trial analyses, identify the health care resource items for which there is a change in use associated with substituting the proposed health technology for the main comparator.

For model-based evaluations, estimate cost weights representing the resources used within a relevant time period (eg a model cycle for a state transition model) for every health state. Alternative health state costs may be defined for patients receiving the intervention and the comparator – for example, to account for differences in adverse event rates.

Health care resource items

Where appropriate, consider the following resource items:

- medical services (ie procedures, diagnostic and investigational services), including the proposed and comparator health technologies, if medical services
- hospital services
- medicines, including pharmaceutical benefits
- blood products
- community-based services (eg attendances by specialists, general practitioners or allied health care professionals)
- any other direct medical costs.

Consider whether there are resource differences between who can request the proposed health technology and the main comparator (eg if the proposed health technology can only be requested by a specialist, whereas the comparator can be requested by a general practitioner).

For pathology services, consider whether use of patient episode initiation, specimen referral or block retrieval services would differ substantially between the intervention and the comparator. Where relevant, include the cost of obtaining a new sample and retesting.

For each resource item, define the natural units and quantify the number of natural units provided to patients in each treatment group, or to patients remaining in a health state for a relevant time period (eg number of services provided, number of packs of medicine dispensed, number of general practitioner consultations, number of episodes of hospital admission etc).

Use of the intervention and comparator services is generally derived from the clinical studies reported in Section 2. However, in circumstances where a therapeutic health technology is provided multiple times over the treatment course and studies have incomplete follow-up, this may represent a truncated mean and require adjustment. Justify and explain any calculation of the cost per patient per year, as necessary, for therapeutic health technologies used episodically. If relevant, incorporate wastage in the model, because it is a consumption and therefore an incurred cost.

For estimates of health care resource item use, describe and justify the basis, and specify the information source. Consider the applicability of the data to the modelled setting. Measure prospectively the pattern of provision of health care resources in the course of a clinical study by:

- retrospectively reviewing relevant records or through linking data with claims data
- administering a questionnaire or survey
- using diaries.

Distinguish between data on resource use that are directly derived from the primary evidence, and extrapolations or modelling of resource use beyond that available from the primary evidence. Justify any choice to use data that are not consistent with data from the primary evidence, particularly where this has an important impact on incremental costs, as revealed in the sensitivity analyses.

If appropriate, exclude types of health care resources that would not have a material influence on the conclusion of the economic evaluation, if appropriate. This may be due to the cost being very small, or that the cost largely cancels out between the intervention and the comparator(s). If resources are excluded for this purpose, state this and justify their exclusion, and outline how the exclusion affects the incremental cost of the intervention.

Occasionally, because of the medical condition under treatment or the age of the patients, consideration of non-health care costs such as social services (home help, day care, meals on wheels, private travel to access health care, etc) or costs to other sectors might be relevant (see also Appendix 10). If incorporation of such non-health care resources is relevant for a supplementary analysis, adapt the general principles described in this TG section to generate and present these variables.

Allocation of prices (unit costs) to resources

Present all unit prices and costs in Australian dollars with a consistent year of analysis (which should be stated and be as close as possible to the submission date of the Assessment Report).

Section 3 adopts a broad perspective for the valuation of health care resources, so include all contributions to the costs of health care resources – including those paid for by patients, governments, health insurance agencies and any other part of society – in the economic evaluation. Generally, the source of costs recommended by the *PBAC Manual of resource items and their associated costs*[n] should be used. It is preferred, however, that the unit cost of MBS-funded health technologies (including the proposed service if it is to be funded through the MBS) used in the economic evaluation includes patient out-of-pocket costs (ie average charges above the schedule fee) where possible. It is recognised that there may be difficulties in obtaining or estimating these costs, and sensitivity analyses should be presented. The unit cost of blood products should be derived from the National Product Price List[o].

If there are important reasons to use different unit prices from those recommended, present these as a sensitivity analysis, justify each, and describe its source or generation. Ensure that any different unit price is consistent with the broad perspective of including all contributions to the costs of health care resources.

Detail all alternative costs, their sources and any assumptions about them. If multiple estimates are identified, justify the estimate used in the base case and present alternative plausible estimates in sensitivity analyses.

If cost conversion is required from current non-Australian prices, and is done using a prevailing exchange rate, justify the price comparability between countries.

If using historical estimates of costs, detail the information sources and the methods used to estimate them. Justify the use of the historical cost source as relevant and the best estimate available. Use the most relevant Australian price index (eg total health and health industry–specific

[n] www.pbs.gov.au/info/industry/useful-resources/manual
[o] www.blood.gov.au/national-product-list

price indexes published by the Australian Institute of Health and Welfare) to adjust for inflation and estimate current prices. If cost conversion is required from older published non-Australian prices, the historical exchange rate should be used, with the most relevant Australian price index to adjust for inflation and estimate current prices.

Value future costs at current prices (ie do not allow for future price inflation in the calculations), consistent with using a constant price year in the economic evaluation.

Presentation of resource use and cost information

A format for summarising the minimum dataset of health care resource items and their associated unit costs relevant to the economic evaluation is suggested in Table 13. These are samples for each identified category, which are consistent with the *PBAC Manual of resource items and their associated costs*,[p] but are not comprehensive of all types of health care resource items, natural units of measurement or sources of unit costs.

Present all steps taken to calculate costs in the economic evaluation in a way that allows the calculations to be independently verified.

If a complete presentation of costs is very large, present the calculations in an accompanying technical document. Cross-reference between the calculations and the main body of the Assessment Report, and include an electronic version of the detailed calculations.

[p] www.pbs.gov.au/info/industry/useful-resources/manual

Table 13 Indicative list of health care resource items, unit costs and usage included in the economic evaluation

Type of resource item	Subtype of resource item	Natural unit of measurement	Unit cost (AUD)	Source of unit cost	Usage for the proposed health technology	Usage for the comparator
Medical services	Proposed health technology	Service rendered	x	Proposed cost of the health technology	[add usage]	[add usage]
	Comparator health technology	Service rendered	x	MBS schedule fee for item code according to current MBS, if MBS-listed service	[add usage]	[add usage]
	Other medical services	Service rendered	x	MBS schedule fee for item code according to current MBS, if MBS-listed service	[add usage]	[add usage]
Medicines	Medicine	Prescription dispensed	x	PBS dispensed price for item code according to current PBS, if PBS-listed medicine	[add usage]	[add usage]
Hospital services	Hospital admission	Episode for identified AR-DRG	x	Average cost weight for DRG item code according to current AR-DRG Public Sector Estimated Cost Weights	[add usage]	[add usage]
Residential care	ACFI category	Daily	x	Daily ACFI subsidy rate plus basic daily care fee	[add usage]	[add usage]

ACFI = Aged Care Funding Instrument; AR-DRG = Australian Refined Diagnosis Related Group; AUD = Australian dollars; MBS = Medicare Benefits Schedule; PBS = Pharmaceutical Benefits Scheme

Technical Guidance 23 Model validation

Validation of an economic model to demonstrate that the generated results represent what they are intended to represent is best practice. It helps to reduce some of the uncertainty associated with modelling, and a more thoroughly validated model allows more confidence in its predictions.

TG 23.1 Operational validation of the economic model

Model traces for the proposed health technology and its comparator provide a clear depiction of the implications of the model. They can inform the face validity of the model logic, computerisation and external validity.

For models that include multiple indications or populations, model traces should be presented per indication/population. For investigative technologies, separate model traces for patients who do and do not have an appropriate change in management might also be informative (due to the dilution effect associated with investigative technologies, whereby the test may affect management in a subset of patients tested).

Use traces to track patients through the model and demonstrate that the logic of the model is correct. Present traces representing the proportions of the cohorts in each health state over time, and the cumulative sum of the undiscounted costs and outcomes (eg QALYs) over time. If applicable, state the number of events over time where patient-relevant events occur within a health state. Comment on whether each of the model traces is logical – for example, ensure that any traces of overall survival practically converge to zero at or before the time horizon of the model where lifetime models are appropriate (see Technical Guidance 18 and Technical Guidance 20).

Compare model traces with corresponding empirical data, where possible, to identify whether outcomes are consistent. Consider both data sources used in the model (dependent validation) and data sources not used in the model (independent validation). For example, compare predicted clinical events with observed data on the natural history of the medical condition. Comment on and explain any differences indicated by these comparisons.

In addition, compare modelled outcomes against outcomes from similar models identified in Section 3A.2.1 of the Assessment Report as a cross-validation tool to identify consistencies (or differences that can be explained).

TG 23.2 Other validation techniques

Present or cross-reference any other completed model validation exercises. The Assessment of the Validation Status of Health-Economic Decision Models (AdViSHE) Study Group describe a range of validation processes, and these should be considered (Vemer et al. 2016).

Technical Guidance 24
Results of the base case economic evaluation

TG 24.1 Intervention costs per patient

Present the expected costs of the proposed health technology and comparator (individually) per patient. For therapeutic health technologies, the costs per patient per course for an acute or self-limited therapy, or per patient per year for a chronic or continuing therapy, should be reported. This estimate should be consistent with estimates of per-patient use in Section 4 of the Assessment Report.

TG 24.2 Stepped presentation of results

If the model translates clinical data, present the results of the key steps involved in transforming the comparative data (from Section 2) into the modelled base-case estimate of incremental cost-effectiveness.

Begin with an analysis of costs and outcomes that are directly associated with the comparative data presented in Section 2. Where the following procedures are undertaken to estimate the base case, sequentially present re-estimated costs and outcomes (and interim results) for each step:

- transformation(s) for applicability
- extrapolation of data over longer time periods
- additional data or assumptions
- transformation of clinical outcomes to final health outcomes (QALYs).

For investigative technologies, consider aligning the initial steps sequentially incorporate evidence from each of the linkages. For example, presenting costs and outcomes associated with test use based on test analytical data, then adding change in management information, clinical outcomes and finally translated (e.g. extrapolated and/or transformed) health outcomes, with additional translations as required for applicability incorporated where relevant.

Identify the steps or assumptions of the model that have important impacts on the ICER.

Table 14 shows an example of how to present a stepped analysis incorporating evidence translations. Table 15 shows an example of steps that may be relevant where the intervention is a diagnostic and linked evidence is required to estimate identify all health outcome and resource changes.

Table 14 Presentation of the stepped derivation of the base-case economic evaluation from the clinical study data

Steps (only included if undertaken)	Proposed health technology costs	Comparator costs	Incremental costs	Proposed health technology health outcomes	Comparator health outcomes	Incremental health outcomes	Incremental cost-effectiveness ratio
Comparative study data (as presented in Section 2); Setting: (trial setting); Time horizon: (trial follow-up)	[A][a]	[B][a]	[A – B]	[C] (surrogate outcome)[b]	[D] (surrogate outcome)[b]	[C – D] (surrogate outcome)	$[A – B]/[C – D] per [surrogate outcome]
Study evidence transformed to clinical outcome and translated to the Australian population and/or Australian setting (may need multiple steps)	[modified A][d]	[modified B][d]	[modified A – modified B]	[modified E][e]	[modified F][e]	[modified E – modified F]	$[modified A – modified B]/[modified E – modified F] per [clinical outcome]
Study evidence transformed to clinical outcome, translated to the Australian population/setting, and extrapolated to the appropriate time horizon	[modified & extrapolated A] = [G]	[modified & extrapolated B] = [H]	[G – H]	[modified & extrapolated E] = [I]	[modified & extrapolated F] = [J]	[I – J]	$[G – H]/[I – J] per [clinical outcome]
Study evidence transformed to clinical outcome, translated to the Australian population/setting, extrapolated and with additional assumptions or modelled information	(G + w) = [K][f]	(H + x) = [L][f]	[K – L]	(I + y) = [M][g]	(J + z) = [N][g]	[M – N]	$[K – L]/[M – N] per [clinical outcome]
Study evidence translated to clinical outcomes, the Australian population/setting, extrapolated, with additional modelling and transformed into a relevant health outcome (eg QALYs)(M→O, N→P)	K	L	[K – L]	[O]	[P]	[O – P]	$[K – L]/[O – P] per QALY

QALY = quality-adjusted life year

a Key outcome(s) from comparative data (presented in Section 2) used to generate 'treatment effect' in the economic evaluation, without any modification.

b If resource data are not provided, estimate resource use and apply costs (Australian $) within the study period.

c Evidence to justify the transformation of the surrogate outcome to the clinical outcome and the method employed should be fully documented in Section 2.

d Include here any transformations to estimated outcomes to increase applicability to the Australian population or setting.

e Include here any modelled changes in the provision of resources that would occur in the Australian health care setting.

f Re-estimate of outcomes after including additional data or assumptions that were not captured in the key comparative clinical data (eg adverse events or second-line treatments).

g Re-estimate of costs after including additional data or assumptions that were not captured in the key comparative clinical data (eg adverse events or second-line treatments).

Table 15 Presentation of the stepped derivation of the base-case economic evaluation, investigate service example

Steps (only included if undertaken)	Proposed health technology costs	Comparator costs	Incremental costs	Proposed health technology health outcomes	Comparator health outcomes	Incremental health outcomes	ICER
Comparative diagnostic accuracy, as applied to the prevalence in the eligible Australian population [a] Time horizon: time to reach a diagnosis	Cost of the proposed test [A]	Cost of the comparator service [B]	[A – B]	TP: []%, FP: []%, TN: []%, FN: []% Total correct diagnoses: [C]%	TP: []%, FP: []%, TN: []%, FN: []% Total correct diagnoses: [D]%	[C – D] (correct diagnosis)	$[A – B]/[C – D] per [correct diagnosis]
Incorporation of repeat or confirmatory testing, which may affect final diagnostic conclusions Time horizon: time to reach a diagnosis	Include cost of additional testing, resampling, where relevant [E]	Include cost of additional testing, resampling, where relevant [F]	[E –F]	Total correct final diagnoses: [G]%	Total correct final diagnoses: [H]%	[G – H] (correct final diagnosis)	$[E – F]/[G – H] per [correct final diagnosis]
Uptake of treatment, or other change in clinical management, by final test result Time horizon: time to treatment allocation decision	Include cost of treatment [I]	Include cost of treatment [J]	[I – J]	Correct treatment allocation: [K]%	Correct treatment allocation: [L]%	[K – L]	$[I – J]/[K – L] per [correct treatment allocation]
Incorporation of effectiveness of treatment (eg survival benefit) translated to the Australian population and/or setting and extrapolated to the appropriate time horizon (may need multiple steps) Time horizon: appropriate time horizon to capture differences in costs and outcomes due to changes in treatment allocation decisions (eg lifetime)	Include costs due to time horizon extension (eg disease progression or management) [M]	Include costs due to time horizon extension (eg disease progression or management) [N]	[M − N]	Life years gained [O]	Life years gained [P]	[O − P]	$[M – N]/[O – P] per [life year gained]
Outcomes transformed into a relevant health outcome (eg QALYs)(O →Q, P→R)	M	N	[M − N]	[Q]	[R]	[Q – R]	$[M – N]/[Q – R] per QALY

ICER = incremental cost-effectiveness ratio; QALY = quality-adjusted life year

a Trial-based accuracy and prevalence estimates could be presented as a prior first step, and then translated to the proposed setting (ie most applicable estimates of accuracy and prevalence in the proposed setting)

The order of the steps for the translation of the trial-based economic evaluation may vary.

The final row of Table 14 incorporates all translation studies and additional modelling to complete the impacts of translation of the trial-based economic evaluation into a modelled economic evaluation. Ensure that this corresponds to the base-case ICER.

The stepped presentation informs the face validity of the results, and identifies assumptions and approaches to be examined in more detail in sensitivity analyses. For example, if the main impact is achieved by extrapolating the final outcome over time, then undertake comprehensive sensitivity analyses around the extrapolation methods.

Present the base-case incremental cost, incremental effectiveness and ICER (calculated as the incremental costs divided by the incremental health outcomes).

TG 24.3 Disaggregated and aggregated base-case results

If a decision-tree model is used, present a detailed disaggregation of costs incurred at each branch by resource type for the intervention and comparator groups. For state transition models, present disaggregated discounted costs by resource type for each health state for the intervention and comparator groups. In all models, report the proportions of patients predicted to experience alternative target clinical outcomes in the intervention and comparator groups.

Alternative examples of tables showing disaggregated costs are provided in Table 16 and Table 17.

Table 16 Health care resource items: disaggregated summary of cost impacts in the economic evaluation

Type of resource item	Subtype of resource item	Costs[a] for proposed health technology	Costs[a] for main comparator	Incremental cost[a]	% of total incremental cost[a]
Medical services	Type of medical service				
	Health state 1	$x1	$y1	$x1 — $y1	z1%
	[etc]	$xk	$yk	$xk – $yk	zk%
	Total	∑$x	∑$y	∑$x – ∑$y	∑z%
Medicines	PBS medicine	[as above]	[as above]	[as above]	[as above]
	Health state 1	[add]	[add]	[add]	[add]
	Health state 2	[add]	[add]	[add]	[add]
	[etc]	[add]	[add]	[add]	[add]
	Total	**[add]**	**[add]**	**[add]**	**[add]**
	Non-PBS medicine	[add]	[add]	[add]	[add]
	Health state 1	[add]	[add]	[add]	[add]
	Health state 2	[add]	[add]	[add]	[add]
	[etc]	[add]	[add]	[add]	[add]
	Total	**[add]**	**[add]**	**[add]**	**[add]**
Hospital services	Hospital admission	[add]	[add]	[add]	[add]
	Health state 1	[add]	[add]	[add]	[add]
	[etc]	[add]	[add]	[add]	[add]
	Total	**[add]**	**[add]**	**[add]**	**[add]**
Residential care	ACFI category	A$x	A$y	$x – $y	z%
	Total	**A$x**	**A$y**	**$x – $y**	**100%**

ACFI = Aged Care Funding Instrument; PBS = Pharmaceutical Benefits Scheme

a Indicate clearly whether cost values are discounted costs (use of discounted costs is appropriate).

Table 17 List of health states and disaggregated summary of cost impacts included in the economic evaluation

Health state in model	Resource use by health state (modelled)	Proposed health technology costs	Main comparator costs	Incremental cost	Total incremental cost (%)
Health state 1	Resource type 1	$x1	$y1	$x1 – $y1	z1
	Resource type 2	$x2	$y2	$x2 – $y2	z2
	[etc]	$x etc	$y etc	$x etc – $y etc	z etc
	Total for health state 1	∑$x	∑$y	∑$x – ∑$y	∑z
Health state 2	Resource type 1	$xx1	$yy1	$xx1 – $yy1	zz1
	Resource type k	$xxk	$yyk	$xxk – $yyk	zzk
	Total for health state 2	∑$xx	∑$yy	∑$xx – ∑$yy	∑zz
[etc]	[etc]	[etc]	[etc]	[etc]	[etc]
Total	**–**	**∑$x + ∑$xx etc**	**∑$y + ∑$yy etc**	**(∑$x + ∑$xx etc) – (∑$y + ∑$yy etc)**	**100**

– = not required

Similarly, an example of a table showing outcomes disaggregated by health state is given in Table 18.

Table 18 List of health states and disaggregated summary of health outcomes included in the economic evaluation

Health state in model	Outcome for proposed health technology	Outcome for main comparator	Incremental outcome	Total incremental outcome (%)
Health state 1	x1	y1	x1 –– y1	z1
Health state 2	x2	y2	x2 – y2	z2
[etc]	[x etc]	[y etc]	[x etc – y etc]	[z etc]
Total	**x**	**y**	**x – y**	**100**

Identify which health states and resources contribute to the greatest incremental differences between the proposed health technology and the comparator.

TG 24.4 Summary of base-case results

Summarise the base-case estimate of the incremental outcome(s), incremental cost and the cost-effectiveness ratio(s) obtained in the economic evaluation(s), including both CUA and CEA where relevant.

Comment on whether there is likely bias in the base case estimate of the ICER (eg an over or underestimate of costs or outcomes, that was identifiable but not quantifiable) and the likely overall direction of that potential bias.

If the ICER is based on an outcome other than life-years or QALYs gained, summarise any other health outcome effects (benefits or harms) that are associated with the intervention, but are not captured in the outcome (and may not have been able to be quantified). If additional health outcomes effects can be estimated, present a summary of relevant health outcomes in the format of a cost-consequences analysis. Compare the presented results with any previous MSAC decisions based on the same measure of outcome.

TG 24.5 Alternate listing scenarios

If there are alternate listing scenarios that may be relevant for MSAC consideration, the results for these alternate scenario analyses should be presented. For multi-indication models, this may include presenting the results of the cost-effectiveness analysis disaggregated by indication (if these can be reasonably excluded from the population eligible for the proposed health technology).

For genetic testing of heritable disease, this would include presenting alternate listing scenarios where testing is expanded incrementally across index through to first-, second- and potentially third-degree relatives. The marginal cost-effectiveness of expanding the populations eligible for the test should also be presented.

For plausible alternate listing scenarios, key sensitivity analyses (Technical Guidance 25) should be presented in Section 3A.9 of the Assessment Report.

Technical Guidance 25
Uncertainty analysis: model inputs and assumptions

TG 25.1 Identifying and defining uncertainty in the model

Present univariate deterministic sensitivity analyses for all input parameters, or natural groups of input parameters (eg cost or utility weights for all target clinical outcomes) using plausible alternatives. The following requests are based on good-practice guidelines for model parameter estimation and uncertainty analysis (Briggs et al. 2012).

Parameter uncertainty

Use commonly adopted statistical standards to represent the uncertainty around the true value of each uncertain input parameter. For example, beta distributions are a natural match for transition probabilities; log-normal for relative risks or hazard ratios; logistic distributions to calculate odds ratios; and gamma or log-normal for costs and utility parameters.

Justify using alternative distributions. Use interval estimates (eg 95% CIs) derived from fitted probability distributions to define the ranges of the parameter values tested in the deterministic sensitivity analyses.

Where there is very little information on a parameter, adopt a conservative approach by defining a broad range of possible parameter values. Never exclude parameters from uncertainty analysis on the grounds that there is insufficient information to estimate uncertainty.

Consider correlation between input parameter values. If applicable, represent the joint uncertainty around the true values of two or more input parameters in the uncertainty analyses. It is preferable to represent the joint uncertainty around transition probabilities in the intervention group and the comparator group through the application of a relative treatment effect parameter. If a relative treatment effect parameter is not applicable, individual-level data for the comparator and intervention could be bootstrapped to provide more realistic estimates of the joint uncertainty between these (Briggs et al. 2012).

The joint estimation of multiple input parameters when using regression analysis produces relevant correlation parameters. Otherwise, model calibration methods may be used to represent joint uncertainty around the true value of model input parameters.

Translational uncertainty

Where clinical data have required translation for applicability issues, transformation or extrapolation for incorporation into the model, systematically consider the assumptions incorporated into the translation and identify any uncertainty in these assumptions. Identify plausible alternatives for testing in scenario analysis.

Examples of analyses that can be used where the data or outcome translations are incorporated into base-case analysis are presented in Table 19.

Table 19 Examples of potential sources of translational uncertainty in the economic model and suggested scenario analyses

Translations incorporated into base-case analysis	Suggested uncertainty analysis
Transformation of continuous outcome data to a dichotomous outcome	Alternative thresholds
Treatment effect with adjustment for switching	Treatment effect without adjustment for switching, and/or using an alternative adjustment technique
Treatment effect based on translation (eg subgroup analysis) following applicability study	Treatment effect based on intention-to-treat population
Selected source(s) of data for treatment effect	Alternative available source(s) of data, and/or meta-analysis of data as source of treatment effect
Transformation of a surrogate to a final outcome	Range of alternative plausible values (as derived establishing STFO relationship)
Extrapolation of data beyond the trial	Alternative data truncation point(s), alternative choices of parametric model, or alternative assumptions regarding ongoing treatment effect
Pooled within-trial data to estimate utility values (or alternative approach)	Estimates based on individual arms (or the alternative approach)
Externally sourced utility values	Alternative values or sources

STFO = surrogate to final outcome

Structural uncertainty

If multiple plausible model structural choices/assumptions are identified, assess and present the potential impact of these on the model outputs. If a substantial impact is predicted, use a formal approach to characterise the structural uncertainty. Use scenario analyses to assess the impact of assumptions around the structure of the economic model, including alternate model structures identified in response to Technical Guidance 18, or alternate assumptions regarding the duration of the treatment effect or choice of parametric model used to extrapolate survival data. Report the results of each set of plausible structural assumptions. Alternatively, parameterise structural assumptions where there is sufficient clinical evidence or expert opinion to do so.

Describe and justify the inclusion and exclusion of potential scenario analyses when making alternative assumptions about model structural aspects.

Include an analysis of the impact of the time horizon.

Use other scenario analyses to assess the effects of substantial use of the proposed health technology beyond the intended population and circumstances of use defined in the requested restriction. This wider population or circumstances are expected to have demographic and patient characteristics and circumstances that differ from the target population and circumstances.

TG 25.2 Presentation of univariate sensitivity and scenario analyses

Tabulate all parameter values and assumptions included in the model, and present the results of univariate sensitivity and scenario analyses in a similar format to Table 20.

Use a tornado diagram to represent the relative effect of the uncertainty around alternative input parameters on the base-case incremental cost-effectiveness result.

Identify the input parameters and model assumptions to which the incremental cost-effectiveness results are most sensitive.

TG 25.3 Presentation of multivariate and probabilistic sensitivity analyses

Use multivariate sensitivity analyses to test the combined effects of the uncertainty around the true values of input parameters to which the base-case incremental cost-effectiveness result was shown to be sensitive in the univariate analyses. If the univariate analyses identify multiple parameters for testing in a multivariate analysis, consider incorporating changes in a stepped manner to allow MSAC to see the impact of each change on the resulting ICER.

Describe the multivariate sensitivity analyses to be undertaken, and present the results. Justify the inclusion and exclusion of parameters in these analyses.

A probabilistic sensitivity analysis (PSA) may be provided in addition to a deterministic sensitivity analysis. As translational and structural uncertainties have previously been more influential on MSAC deliberations than uncertainty regarding the precision around parameter estimates, multivariate analyses incorporating these translational and structural uncertainties should be prioritised above the conduct and presentation of PSA.

If undertaking a PSA on a cohort-based state transition model, the number of iterations (sets of randomly sampled input parameter values included in the analysis) should provide stability in the model outputs across multiple analyses using alternative random number seeds. Provide the random seed associated with the presented results to enable replication, and also ensure that the model permits alternative seeds.

If undertaking a PSA on an individual-level model (eg a discrete event simulation), the number of iterations may be selected to balance stability of model outputs and a reasonable time required to undertake a PSA (eg a few hours, rather than a few days).

Use cost-effectiveness planes and acceptability curves to present the results of a PSA, as well as the tabulated presentation of the interval estimates for the ICER or the incremental net benefits of the proposed health technology.

TG 25.4 Summary of the uncertainty analysis

Describe and justify a likely range of values within which the true estimate of the incremental cost-effectiveness of the proposed health technology is likely to lie, identifying the key sources of uncertainty. This range may be informed by a formal PSA, or by subjective interpretation of the presented deterministic sensitivity and scenario analyses.

Discuss the implications of the sensitivity and scenario analyses with respect to the certainty of the base-case ICER estimate.

Discuss the likely overall effect of deficiencies in the evidence base on the reported cost-effectiveness of the proposed health technology.

Table 20 Results of the sensitivity and scenario analyses characterising the uncertainty around the ICER

Variable or assumption	Base-case value	Plausible alternative(s) or range of values	Incremental outcomes	Incremental costs	ICER	Description of impact on ICER
Base case			[base case]	[base case]	[base case]	
Discounting rate	Outcomes and costs = 5%	Outcomes and costs = 3.5% Outcomes and costs = 0%	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Plausible range of treatment effect, if modelled as a variable (eg hazard ratio or relative risk)	[add]	[eg upper and lower 95% confidence intervals around estimate]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Altered patient characteristics, if relevant	[add]	[eg different average age, disease or condition severity]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Transition or event probabilities	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Outcome-related assumptions or variables [Recommended examples: • alternative methods or sources of utility weights]	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Cost-related assumptions or variables	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Alternative extrapolation variables or assumptions [Recommended examples: • start point • choice of parametric model • assumption regarding ongoing treatment effect]	[eg maximum follow-up]	[eg median follow-up]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Any other translation assumptions [eg use of intention-to-treat/non-adjusted data]	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Alternative assumptions regarding model structure	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Time horizon	[add]	[eg trial based; 5, 10, 20 years, as appropriate]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]
Plausible alternatives for other variables or assumptions [eg including leakage beyond the requested restriction]	[add]	[add]	[alternative estimates]	[alternative estimates]	[alternative estimates]	[describe as required]

ICER = incremental cost-effectiveness ratio

Section 3B Cost minimisation

This section provides information requests for preparing Section 3 using a cost-minimisation approach (see Section 3, Introduction).

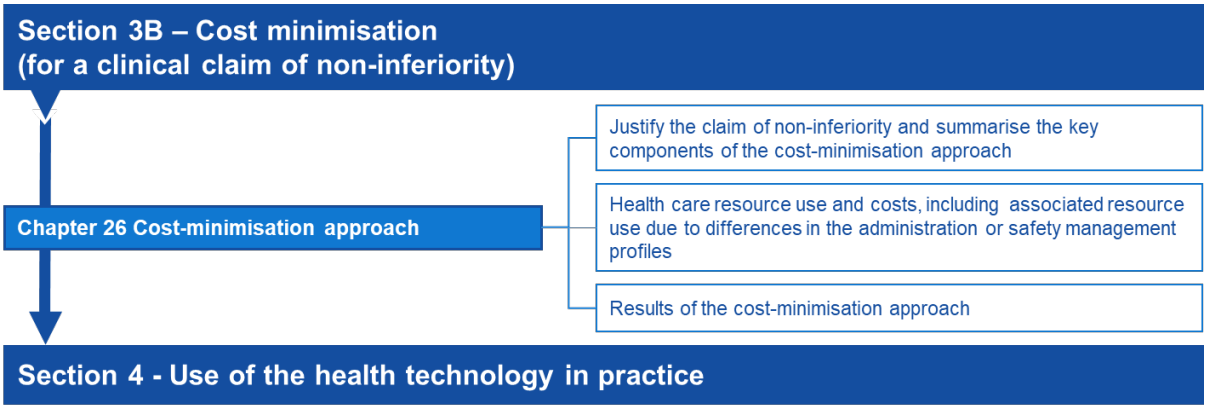
The assumption of non-inferiority, with respect to both effectiveness and safety, needs to be well justified for the cost-minimisation approach to be accepted. Irrespective of the therapeutic claim, if the adverse effect profiles of a proposed health technology and its main comparator are significantly different in nature, it is unlikely that the cost-minimisation approach will suffice. The implications of these differences, for both health outcomes (ideally, utility) and resource use, should be explored in a full economic evaluation.

The cost-minimisation approach has an abbreviated Section 3, where differences between the proposed health technology and the comparator that are likely to result in a difference in health resource use should be identified. This includes identifying differences in:

- the costs of prescribing or administering the services
- the costs of monitoring or managing associated adverse events
- anything else that may impact health resource use.

A cost analysis compares costs only and so is strictly defined as a partial rather than a full economic evaluation, because it does not quantitatively assess comparative costs in a ratio over comparative effectiveness. Although less preferred than a full economic evaluation, cost analyses have sometimes been presented and found to be acceptable if the proposed health technology is demonstrated to be no worse in terms of effectiveness but to have a superior safety profile compared with the main comparator.

Flowchart 3B.1 Summary of guidance for a cost-minimisation approach



Technical Guidance 26 Cost-minimisation approach

TG 26.1 Health care resource use and costs

Direct health technology costs

Using guidance in Technical Guidance 22, estimate the direct health technology costs per patient. For therapeutic health technologies, the costs estimated should be per patient per course for an acute or self-limited therapy, or per patient per year for a chronic or continuing therapy. Use of the intervention and comparator therapies is generally derived from the clinical studies reported in Section 2.

For diagnostic services, it would generally be sufficient to cost the health technologies in each arm to the point of diagnosis. It would be difficult to justify a cost-minimisation approach assuming final health outcomes were equivalent if the analytical test outcomes/diagnostic outcomes were not also equivalent.

Additional costs and/or cost offsets

The nature of additional costs and/or cost offsets will differ across MSAC applications. Two common areas for these are costs associated with prescribing or administration and costs of managing adverse events; however, this does not preclude other possible cost offsets. This could also include subsequent changes in resource use due to changes in management (eg further downstream testing) that result from investigative technologies, provided these do not impact downstream costs and final health outcomes. Justify any other additional costs and/or cost offsets in terms of how they are realisable and/or patient relevant, and show how they differ between the options being considered in the cost-minimisation analysis.

Comparison of prescribing and administration profiles

Identify differences in the costs of prescribing or administering the health technologies.

Listing a non-inferior health technology might have cost consequences related to its differing mode of administration. These have sometimes arisen if the proposed health technology and its main comparator are available in different forms. If this applies in an assessment report, identify the types of other resources affected, estimate the extent to which the quantity of each type of resource provided would change (in its natural units of measurement) following a listing, and multiply by the relevant unit costs. Aggregate this with the health technology cost impact to estimate the net cost impact within the cost-minimisation analysis.

See also the *PBAC Manual of resource items and their associated costs*[q] for further detail on costing administration-related resource use.

Comparison of safety management profiles

Only use the cost-minimisation approach where the proposed health technology has a safety profile that is superior (preferably) or non-inferior to the main comparator.

Identify any differences in the costs of monitoring or managing adverse events associated with the health technologies.

If the proposed health technology is demonstrated to be no worse in terms of effectiveness, but to have a superior safety profile to the main comparator, a price advantage for the proposed health

[q] www.pbs.gov.au/info/industry/useful-resources/manual

technology over its main comparator could be sought on the basis of cost offsets because of reduced costs of monitoring for, or managing of, adverse reactions. Use clinical trials and the recommendations in the Australian Instructions of Use to support a claim that monitoring costs are reduced.

Where safety profiles are similar, but the proposed health technology simply has a reduced magnitude of adverse effects (severity or incidence), present a thorough description of the quantified differences in safety, with a justified estimate of any corresponding resource-use implications.

Where the adverse effect profiles of a proposed health technology and its main comparator are different in nature, a cost-effectiveness or cost-utility analysis is likely to be preferred (Section 3A). However, a cost analysis may be acceptable to quantify a claim that the cost offsets from the reduction in health care resources required to treat the adverse events are sufficient to reduce the incremental cost to zero or a negative value.

See also the *PBAC Manual of resource items and their associated costs*[r] for further detail on resource use and costing associated with monitoring and adverse effects.

TG 26.2 Results

Results of the cost-minimisation approach

List all identified costs associated with both the proposed health technology or the comparator to estimate the net cost difference.

The economic claim should be that, at the price requested, the overall cost of therapy with the proposed health technology is the same as, or less than, the overall cost of therapy with the main comparator.

Sources of data

Provide copies of the original sources of all data (beyond those already presented in Section 2) or expert opinion used in the model in an attachment or technical document. Cross-reference data extracted from each source to the level of the page, table or figure number of the source document.

To enable independent verification of each analysis, provide an electronic copy of any computer-based calculations of the analysis.

[r] www.pbs.gov.au/info/industry/useful-resources/manual

Section 4
Use of the health technology in practice

Introduction

Section 4 presents a set of budget impact analyses, and provides the most likely extent of use and financial estimates. These analyses are relevant to MSAC, the Australian Government and, where relevant, other Committees/funding bodies that refer to MSAC. Section 4 is important for estimating the likely uptake of the proposed health technology in clinical practice and the cost impact of the service to the relevant funding program and to the Australian Government budget. Depending on the funding context, this may also be used to negotiate risk-share arrangements.

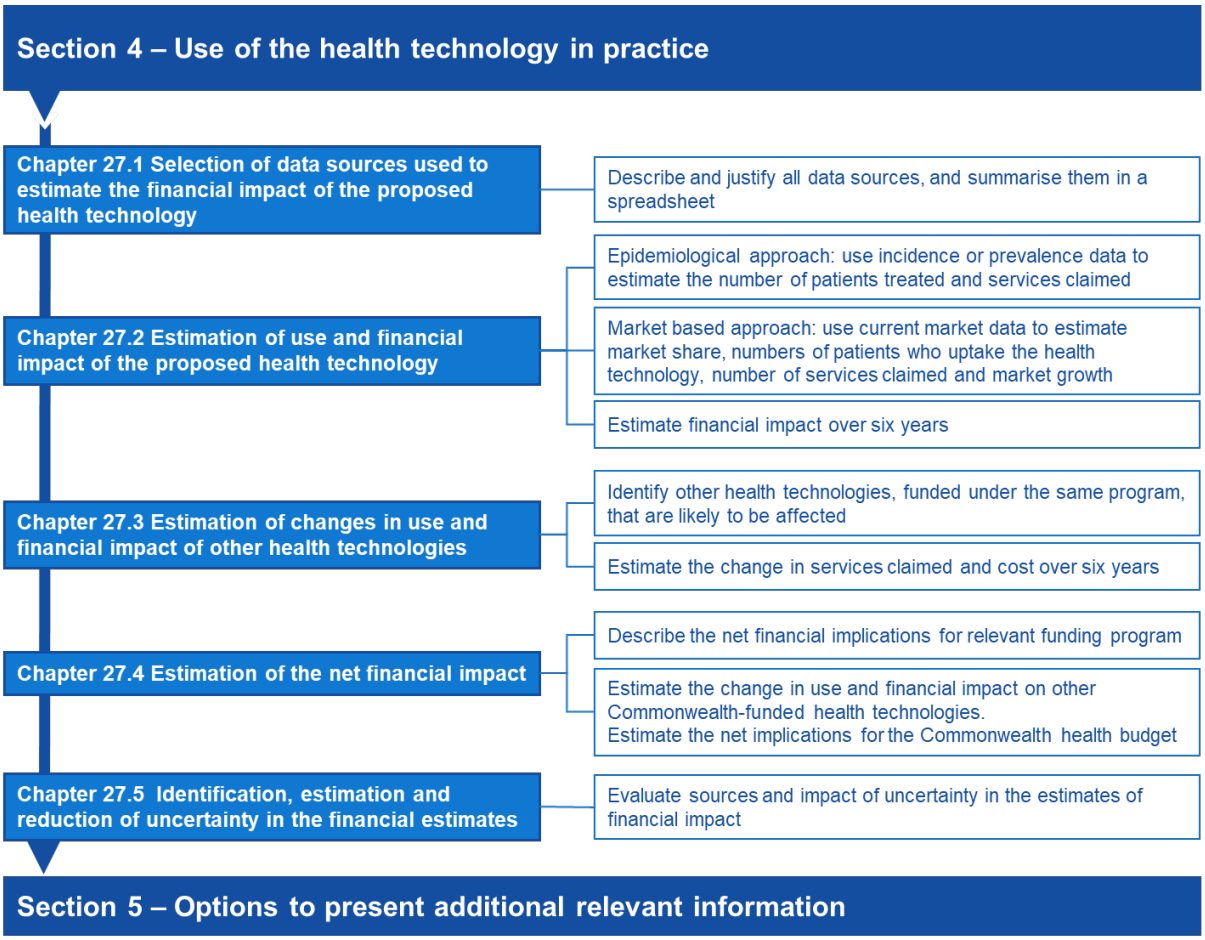
Epidemiological and market-share analyses are the two broad approaches for developing utilisation and financial estimates, although their use is not mutually exclusive. An epidemiological approach is usually preferred for generating utilisation and financial estimates if the Assessment Report indicates a superior therapeutic clinical conclusion. However, a market-share approach might be preferred if the Assessment Report indicates a non-inferior therapeutic clinical conclusion.

The approach taken should be justified in the Assessment Report. Demonstrate concordance across both approaches where data inputs from one approach (epidemiological or market share) are uncertain.

Ensure that any estimates of the extent of use of the proposed health technology (and other technologies affected by the listing of the proposed) in the Australian setting are consistent with evidence presented throughout. Ensure that uptake of the health technology, change in the use of alternate health technologies and offsets are all consistent with the clinical place (Section 1), the use of the health technology in the clinical evidence (where applicable) (Section 2) and the circumstances presented in the economic evaluation (Section 3). Any discrepancies should be explained and justified.

Provide sufficient data in Section 4 so that the steps can be interpreted. Where the calculations used to generate estimates are not transparent in the main body of the Assessment Report, present additional data.

Flowchart 4.1 Summary of the guidance for estimating the use and financial impact of the proposed health technology in practice



Specification of the relevant funding program

As it is within the remit of MSAC to consider Assessment Reports for health technologies funded through different funding programs, the relevant funding program for the proposed health technology should be identified. The utilisation and financial estimates of the proposed health technology to the relevant funding program should be presented in Section 4.2 of the Assessment Report, with any consequential utilisation and financial changes to other items funded by the same funding program presented in Section 4.3. These should be presented after the exclusion of any non-government copayments. In Section 4.4, the estimated net financial implications to the relevant funding program should be reported. Changes in the utilisation and financial estimates of other health technologies funded by Commonwealth Government health programs should be reported in Section 4.5.

Epidemiological approach

An epidemiological approach estimates the number of people with the medical condition, and then estimates the use of the proposed health technology (see TG 27.2) and consequential changes in use of other services (see TG 27.3) in the context of the patient group defined by the proposed item descriptor.

An epidemiological approach estimates the patients eligible for the proposed health technology; however, market-based data or market research may be required to establish estimates such as the rate of uptake of the health technology.

In contrast to the economic evaluation presented in Section 3 of the Assessment Report, these financial analyses exclude health outcomes, do not use discounting, and exclude any resource item or copayment from a source other than the identified budget in Section 4.4. However, in Section 4.5, financial implications to other Commonwealth budgets can be presented.

Market-share approach

The market-share approach estimates the extent of the current market represented by the proposed patient indication and, consequently, the share likely to be taken by the proposed health technology. It is likely to be the most suitable approach where the proposed health technology will completely substitute existing MBS-listed services.

In contrast with the epidemiological approach, the market-share approach allows an abbreviated presentation of information, where justified by an expectation of no market growth following listing, or provides an alternative way of generating estimates to compare with the epidemiological approach.

The key issue with estimates built on the market-share approach is whether the current market or market growth rate is expected to increase because of listing the proposed health technology on the MBS. If not, a health technology listed on a cost-minimisation basis would usually have a negligible effect on the net financial impact on the MBS, but may have financial impacts on other parts of the Australian Government health budget. If the proposed health technology is likely to increase the market size or its growth rate, it is critical to estimate the extent of this likely increase.

Fully editable electronic copy of the financial implications analysis

The analysis should be constructed in an Excel workbook to be provided with the Assessment Report to allow an independent assessment of the data. A template has been provided on the MSAC website to facilitate the presentation of these analyses. Ensure that the responses to Section 4 and the Excel workbook cross-reference the extraction of all data used to generate estimates in these analyses, from each attached data source (to the level of the page, table or figure number of each

source document). Where commissioned data have been used, include the correspondence for the data request.

Ensure that the calculations flow through the spreadsheets, so that changes to any variable flow on to the results. To help understand the spreadsheets, apply clear and unambiguous labels to spreadsheet values, and cross-reference the data source. Where relevant, complex analyses or supporting data should be presented in separate spreadsheets. Provide clear and consistent formulas in the spreadsheets, to facilitate tracing and replicating the calculation flow.

Throughout Section 4, refer to the relevant spreadsheet number/title. Describe the approach, methods, assumptions and potential biases. Where possible, add comments to the Excel workbook to describe these factors, particularly if the approach is complex. Confidence in the estimates is reduced if the interpretation of calculations in the Excel workbook cannot be reconciled with the relevant assumptions or approach.

Technical Guidance 27 Use of the health technology in practice

TG 27.1 Selection of data sources used to estimate the financial impact of the proposed health technology

Available data sources

Data sources fall under the broad headings listed in Table 21; however, there might be other suitable data sources (some examples may be listed on the PBS site: Sources of data for use in generating utilisation estimates[s]).

For the market-share approach, relevant sources of data include MBS data, including those supplied by Services Australia relating to the MBS rebates paid and patient out-of-pocket costs, or data collected through the relevant funding program.

Table 21 Categories of data sources

Data type	Examples
Disease or condition epidemiological data (provide estimates of prevalence or incidence in the population)	• Australian case or mortality registers that estimate the incidence or prevalence of a disease or condition • Large, well-designed Australian studies that estimate the incidence or prevalence of a disease or condition • Australian national health surveys that estimate the prevalence of a disease or condition • Utilisation databases, including MBS data for other services in the proposed population, or State-based utilisation data where the proposed health technology is already in use.
Market data	• Quantitative description of the existing market, including estimates of change in the size of the market over time • Estimates of relative market shares • Estimates of the impact of the requested MBS listing on current treatment paradigms, based on similar previous listings
Commissioned data	• Data requests to registries, epidemiological studies or utilisation studies • Epidemiological studies

MBS = Medicare Benefits Schedule; PBS = Pharmaceutical Benefits Scheme; RPBS = Repatriation Pharmaceutical Benefits Scheme

Different sources of data may be required. In Section 4.1 of the Assessment Report:

- describe the data and data source
- explain the purpose of the data in the analysis
- describe how the data are relevant to the present Australian setting. Where data on overseas markets are provided, clearly state that Australian data were not available and discuss the applicability of these data to the Australian setting (with particular reference to the subsidy arrangements in the overseas jurisdiction)
- where there are multiple sources of data, discuss the concordance across these sources and present sensitivity analyses for the different estimates across the sources
- for each estimate derived from source data, summarise the methods, and discuss any assumptions, limitations and biases in the approach taken.

[s] www.pbs.gov.au/info/industry/useful-resources/sources

Commissioned data

A commissioned study may be used to fill a gap in the data, and may include health technology usage surveys; data from disease or condition registries; or claims data. Clearly state the original purpose for the collection (eg the data were collected for the primary purpose of understanding treatment choices). When reporting the results of commissioned data, provide sufficient background and methodological information to adequately interpret the results.

See Appendix 9 for further guidance on presenting commission data from a survey of experts. Provide the method for identifying respondents, the reasons for collecting information, and any potential conflicts of interest of the respondents or the company undertaking the survey. Present the actual questions asked and the range of responses. Where the respondents are experts in treating specific diseases, provide an estimate of the number of patients they treat, what proportion this is of the expected numbers of patients in Australia, and the health area and setting in which the respondents practise (eg public hospital, private hospital, community, regional area, inner urban area).

When analysing administrative data and registries, provide sufficient information about the method used to sample the dataset, the proportion of the affected population included in the dataset, rules for analysis, assumptions used (particularly where elements in the dataset are used as surrogates) and statistical methods (such as censoring or use of propensity scores).

TG 27.2 Estimation of use and financial impact of the proposed health technology

Justify any estimates of the incidence, prevalence or market growth over six years. Multiple factors may influence growth, and it may not be appropriate to assume linear growth in the estimates, particularly if the proposed health technology is not the first entrant to the market for the specific indication. It is important to base projections on the number of patients, not services provided, wherever possible.

Epidemiological approach

Incidence or prevalence data

For an epidemiological approach, present the methods and assumptions for converting incidence or prevalence data to the number of patients likely to uptake the proposed health technology each year.

The choice to use incidence or prevalence data depends on several factors, including the nature of the medical condition, its treatment and the available data. In general, treatments of short duration are best suited to incidence estimates, and long-term treatments (eg for chronic diseases or conditions) may be better suited to prevalence estimates. A combination of prevalence and incidence estimates may be required (eg intermittent treatments for a chronic condition).

Consider the current prevalent patient population in addition to the incident population – for example, a cancer therapy where there are patients receiving best supportive care before the proposed health technology becomes available. Only calculating the incident population would underestimate the likely number of patients treated in the early years of listing.

Estimate of the number of patients with the medical condition

Estimate the likely number of patients in the six years following listing, using the incidence or prevalence approach, accounting for changes in disease or condition incidence or prevalence trends. If appropriate, present shorter periods (eg monthly or quarterly) in supporting spreadsheets and summarise annually for six years from listing. If using an incidence approach, also estimate the

prevalent population (from years before listing) that may add to the eligible patient pool in year 1. Justify when the addition of a prevalent population is not required.

If the medical condition has a subjective element in its diagnosis, consider the impact of misdiagnosis for the purposes of rendering patients eligible for the proposed health technology.

Estimate of the number of patients eligible for the proposed health technology

Using the annual numbers of patients with the medical condition for six years, estimate the proportions of patients who would be expected to be eligible for the proposed health technology according to the proposed eligibility criteria.

Where the proposed eligibility criteria contain subjective elements, consider whether patients might be misclassified to be eligible for the proposed health technology.

Estimate of the number of patients likely to use the proposed health technology

Using the annual numbers of eligible patients, estimate the proportions likely to take the proposed health technology in each of the six years. Ensure that the estimates reflect the rate of uptake of the proposed service and consider the impact of the use of other services/treatment options. For proposed MBS services, uptake should further be considered by the setting of use, ie private sector or public hospital sector. Analyses should account for billing of the MBS by public hospitals, where relevant.

Consider whether there are differences in out-of-pocket costs associated with the proposed health technology that may influence the rate of uptake. Justify the estimate of uptake and assess variations to this estimate in a sensitivity analysis.

Number of times the proposed health technology is delivered

The estimate of the number of services provided for each of the six years should account for, where applicable:

- the rate of uptake of the proposed health technology across the six years from listing (described previously)
- the number and frequency of use of proposed health technology per patient

Present each of the steps for estimating the units dispensed separately.

Market-share approach

Describe the market

To generate estimates of expected utilisation and costs, ensure that the market-share approach relies on health technology utilisation data or studies for currently available services that are likely to be substituted. This is the basis for predicting whether the market will change because of listing the proposed health technology.

Number of services provided by currently listed items

Estimate the units dispensed in the most recent 12 months of the relevant market.

Where possible, present the services provided **and** the number of patients this represents according to the evidence provided in Section 2. This will be particularly important where a market-share approach is being compared or used in conjunction with an epidemiological approach. It may also be required where the Assessment Report is providing information on services that increase or decrease in usage, because this is often calculated from patient-level data rather than units

dispensed. However, if the number of services per patient per course of treatment is uncertain, do not back-calculate to patients, as it can introduce significant errors into the patient numbers.

Estimate the rate of growth in this market over six years following listing. Base this on historical trends in the market or other influences, but ensure that it is unrelated to the listing of the proposed health technology. Justify the estimate of market growth in the absence of the listing of the proposed service.

Where more than one service within the funding program is likely to be substituted, present the market share and rate of growth for each item, if required. Disaggregating the estimated growth according to each service is important if they are likely to have different rates of growth, are likely to be substituted differentially by the proposed health technology or have a different cost.

Estimate of the market share

Estimate the rate of substitution in the market by the proposed health technology for each year over six years. Provide evidence, such as market uptake rates from other markets and the applicability of these markets to the Australian setting, to justify the estimate of market share. Clearly communicate and justify the likely extent of market uptake following listing of the proposed service.

Present a table in the Assessment Report for overall estimates, if appropriate. Also present a table in the Excel workbook, stratified by individual health technologies, and clearly show the steps for aggregating the data. Ensure that the proportions of each health technology likely to be substituted by the proposed service are clear on the spreadsheet.

Estimate of growth in the market after listing

Estimate the units dispensed for the proposed health technology for each year that is above the growth projected in the market, using historical data. Report both the expected increase in patient numbers, and expected number of services for the proposed health technology.

Justify when no additional growth in the market is predicted. When the proposed service may be used in clinical practice to treat people who are intolerant to an existing listed service, or following failure with that service, it is likely that entry of the proposed service into the market will increase the overall number of people treated.

Provide references to data of similar circumstances in similar markets, and discuss risks associated with market growth, to increase the certainty of the financial implications of listing the proposed service.

Financial impact over six years

Present the total estimated financial impact for listing the proposed health technology, with appropriate patient co-payments subtracted. For proposed MBS items, the MBS rebate paid depends on a number of factors:

- whether the service is provided as part of an episode of hospital or hospital-substitute treatment;
- whether patients are bulk-billed; and
- for high cost outpatient (non-admitted patient) services, the patient co-payment is capped at the Greatest Permissible Gap amount (and so the proportion covered by the MBS increases with service cost). Additional rebates under the Extended Medicare Safety Net may also apply for outpatient (non-admitted patient) services, if a patient is eligible.

Ensure that these are considered, where important, in the estimated financial impact of listing the proposed health technology. If the proportion of services that attract the different levels of MBS benefit (i.e. 75% or 85%) are known, or if there is adequate justification to support that only one rebate would apply, these proportions can be used. However, if the proportions are unknown, then a pragmatic approach assuming an 80% level of MBS benefit may be used.

TG 27.3 Estimation of changes in use and financial impact of other health technologies

Identify health technologies likely to be affected

If using a market-share approach, services funded under the same program that are likely to be substituted will have been identified in Section 4.2 of the Assessment Report. However, identifying other affected services within the program that will increase or decrease in usage may still be relevant.

Health technologies funded within the same program likely to be affected by the listing of the proposed health technology include:

- health technologies substituted by the proposed health technology;
- other health technologies with decreased usage; and/or
- other health technologies with increased usage.

List all health technologies that fall into each of these three categories. Include those identified as comparators and as other relevant therapies in Section 1 of the Assessment Report. Where the proposed health technology is replacing a technology funded through a different program, or where patients are receiving best supportive care in the absence of the proposed technology, there will be no substituted technologies.

Health technologies funded within the same program with expected increased or decreased usage after the listing of the proposed health technology include those that are:

- co-administered with substituted therapies or with the proposed service;
- used to treat adverse reactions to substituted therapies or the proposed service; or,
- used to treat the clinical end points that might be increased or reduced after the proposed health technology.

The impact of adverse reactions might have less weight if the evidence shows that they are of insufficient clinical importance to require management, or if they are similar for the proposed health technology and its comparator. Note if there is insufficient information available from trial results or extended assessment of comparative harms to include the impact of adverse reactions on expenditure.

Change in other health technologies funded within the same program provided over six years

Discuss the extent of change for each health technology within the same funding program that will be substituted, and for those that are expected to increase or decrease in usage after listing of the proposed health technology. Present and justify the change in the number of services provided for each of these over six years. Reference how the estimates were generated and the data on which the estimates are based.

Justify any inconsistencies between Sections 3 and 4 in terms of the identified health technologies or the estimated extent of change of usage over the six years following listing of the proposed health technology.

Financial impact over six years

Based on estimated utilisation changes, estimate the financial impact in each year over six years for each health technology funded within the same program that is substituted, decreased (ie cost offsets) or increased (ie on costs). Refer to TG 27.2 for the suggested approach.

TG 27.4 Estimation of the net financial impact

Net financial implications for the relevant funding program

The net financial implications for the relevant funding program over six years should be presented in Section 4.4, accounting for the estimated cost of the proposed health technology (estimated in response to TG 27.2), the increased usage of other health technologies and cost offsets for substituted health technologies with a likely reduction in usage (estimated in response to TG 27.3).

Net financial implications for the Commonwealth health budget

Change in use and financial impact on other Commonwealth health budgets

Use the approach in TG 27.3 to identify health services funded through other Commonwealth Government health budgets that are likely to be affected by the listing of the proposed health technology.

Based on estimated utilisation changes, estimate the financial impact in each year over six years for each affected health service and per program (eg if multiple PBS items are expected to be affected by the listing of the proposed health technology, estimate the financial impact for the change in each item, and then overall to the PBS). Refer to TG 27.2 for the suggested approach to estimate the financial impact. Present costs with, where relevant, the appropriate patient copayment subtracted.

Net implications for the Commonwealth health budget

Present the net financial implications for the health budget over six years, incorporating the changes in use and financial implications on the Commonwealth Government budget estimated in Section 4.5 of the Assessment Report, to the budget impact estimated in Section 4.4.

TG 27.5 Identification, estimation and reduction of uncertainty in the financial estimates

Sources of uncertainty

Uncertainty arises when estimating utilisation and financial implications because of the potential for usage that differs from expectations, and usage that extends beyond the restriction.

Address these sources of uncertainty and clearly differentiate the two. Where there is substantial uncertainty in the utilisation and financial estimates, particularly when this uncertainty is a result of usage beyond the restriction (‘leakage’), minimise the impact of the uncertainty by proposing a risk-sharing arrangement.

Factors affecting uncertainty

The following subsections list some factors to consider when assessing uncertainties in predicted utilisation patterns and financial implications resulting from listing of a proposed health technology as requested. The lists are not exhaustive. Factors may arise from epidemiological data, expert opinion and assumptions used in generating the quantified predictions. Present any of these factors to increase understanding of the uncertainties present in utilisation estimates. It might not be necessary to address any or all of these factors, because the uncertainties might be very small or of little importance to the overall cost to the MBS, so consider how relevant each of the factors might be.

Factors that could affect the extent of usage within the requested restriction

Consideration of the following factors might provide relevant information on uncertainties within the requested restriction. Some factors might not be relevant in all Assessment Reports or might have a negligible impact on the overall estimates:

- Promotion might result in greater identification of the proposed health technology, resulting in more health care practitioners considering patients for treatment.
- Indirect media exposure to consumers might result in some consumers being more aware of, and seek to use, the proposed health technology. These patients might not be identified if a treated prevalence approach has been used.
- Outcomes of related research might have an impact on uptake of the proposed health technology. This could be positive or negative, and could emerge at the time the Assessment Report is lodged or be expected to occur within five years of listing.
- More health care practitioners and patients might seek treatment if the proposed health technology treats a medical condition for which the alternatives are considered to be substantially inferior to the proposed health technology (e.g. in terms of effectiveness, tolerability, or patient acceptability and convenience).
- Limited access to designated types of health care practitioners or to designated diagnostic procedures in a requested restriction might limit uptake and utilisation.
- The duration of treatment might be longer than expected, compared to the time frame of the randomised trials, particularly when trials are truncated.
- Utilisation might be greater than expected, particularly in the case of medical conditions with episodic manifestations.
- There might be a likelihood of usage increasing over time.

Factors that could affect the likelihood of usage beyond the requested restriction

Some of the factors listed above might also affect the likelihood of usage beyond the requested restriction. More detailed guidance is given in Section 1 about ways of designing a restriction to minimise usage beyond its intention, however, the following factors might be considered:

- The requested restriction is for a subset of the types of patients who are eligible according to the TGA-approved indication(s).
- The requested restriction is for a subset of the types of patients who were eligible for the randomised trial(s) published for the proposed health technology, or there are randomised trials demonstrating evidence in other medical conditions.
- The requested restriction is for a subset of the types of patients who have been subsidised by the applicant before lodgement of the Assessment Report (e.g. on compassionate grounds or as part of clinical studies).
- The requested restriction is for a subset of the types of patients for whom the applicant plans to promote use of the proposed health technology before or after the listing for MBS funding is implemented.
- The requested restriction is for a subset of the types of patients who have the underlying medical condition, in this case identify whether:
  - there are any likely difficulties for health care practitioners in determining eligibility for the proposed health technology (e.g. a difficult differential diagnosis, ambiguity in the wording of the restriction, or poor precision or accuracy in a diagnostic test) that might result in misclassifications of eligible patients from the population with the underlying condition; and /or
  - patient advocacy groups are likely to have an influence on determination of eligibility by health care practitioners.

Impact of uncertainty

Address the following factors in any uncertainty consideration:

- The direction of impact on the estimate (underestimate or overestimate);
- The impact on the magnitude of the estimate (small or large); and,
- The likelihood that another estimate should replace the base-case estimate (probable or improbable).

Although quantitative estimates of uncertainty are preferred, provide approximate assessments, if required. Note where the effects of some uncertainties are difficult to quantify. As a general principle, the more sensitive the overall financial implications are to a particular source of uncertainty, the more important it is to minimise that uncertainty.

Reducing the uncertainty

Uncertainty can be reduced by using data from multiple sources, if available, which is sometimes referred to as 'triangulation' (the use of multiple sources of data or multiple approaches to determine the consistency or otherwise of the conclusions from those sources or approaches). Where estimates derived from different sources are concordant, there might be more confidence, and less uncertainty, in the resulting estimates. Where estimates are discordant, the disparity between the estimates might contribute to the estimate of uncertainty. A similar approach can be taken when more than one methodological approach has been applied (eg estimates based on a market-share base as well as an epidemiological base; or treated prevalence, where the prevalence of patients treated for a disease or condition, determined from an epidemiological database, is used as a surrogate for the true prevalence).

Summary of calculations

Summarise the results of any calculations (eg sensitivity or scenario analyses), to quantitatively examine the impact of uncertainty.

Section 5
Options to present additional relevant information

MSAC considers factors beyond the clinical, economic and financial implications of the proposed health technology. The purpose of the 'Other relevant considerations' guidance is to discuss concepts which may affect implementation of the proposed health technology or influence the decision-making, but has not been captured through the evaluation of the comparative safety, effectiveness and cost-effectiveness of the technology.

There are two types of additional information that may be relevant to MSAC decision making.

Other or personal utility

For the purposes of these guidelines, other or personal utility is well-being or benefit derived by a subject (or a subject's family or carers) from knowing the results of a test. The completion of this section is only required if the proposed test is more costly than the comparator, and the additional cost is not adequately justified by an impact on health. Examples of where the other or personal utility of a test may be necessary are:

- A test that can detect a disease for which there is no available treatment (the test results in an increase in cost but does not result in health gains).
- A test that can provide a prognosis although there is no clinical management that would alter the prognosis.

Benefits or impacts beyond individuals, family members or carers are not considered other or personal utility.

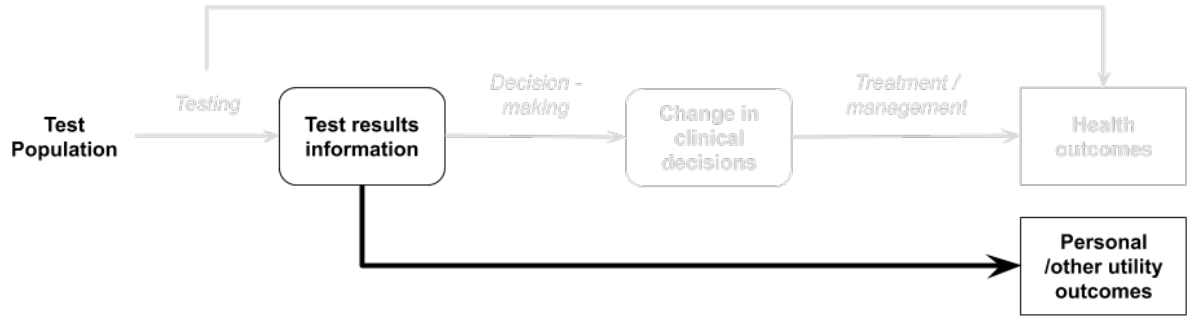
Other relevant considerations

These may include ethical principles such as equity, rule of rescue, and other factors (organisational issues, social issues, legal issues and environmental issues).

While other relevant considerations may include benefits that fall within the other or personal utility category, it also includes impacts that are broader than the individual, family members or carers.

There may be additional impacts of the proposed medical service beyond the health care system. Some non-health care system costs and outcomes may be quantified and included in supplementary analyses in the economic analysis (see Appendix 6). However, some impacts may be less readily quantified (such as impacts on educational attainment).

Technical Guidance 28 Other utility



TG 28.1 Introduction

In the absence of being able to establish the clinical utility of a test, MSAC may consider other utilities, such as personal utility, familial utility or carer utility.

For the purposes of these guidelines, personal utility is defined as encompassing any consequence for the well-being of a patient which does not arise from changes in health outcomes attributed to subsequent changes in the provision of healthcare resources. These outcomes may or may not be able to be demonstrated in quantitative data. If the health technology provides benefits for people other than the patient receiving the health technology (e.g. their family members or carers) then the utility to these people may also be considered.

When Wilson and Jungner first described their criteria for screening tests in 1968, it was deemed essential that there should be an effective treatment available in order for screening to be worthwhile (Becker et al. 2011). However, since this time, there has been an understanding that information derived from tests may be used in a range of ways, such as reproductive planning or changing behaviours relating to sectors outside of health.

The majority of tests provide some value to patients and/or family members through providing a greater degree of certainty regarding a diagnosis, risk level, prognosis etc. In most assessments, the qualitative benefit of the information itself need not be demonstrated, as the clinical claim rests on how the information is used by a clinician to alter the clinical management of the patient. However, there are occasions where information may not lead to any change in clinical management, or health outcomes. Value may still be derived from the test results, through the ability to avoid a lengthy diagnostic odyssey, plan for end of life, allow social support from others with a similar diagnosis, or financial support due to a diagnosis allowing access to disability schemes etc. Tests may therefore have value to patients, families and caregivers, from the “value of knowing” and may provide peace of mind, or reassurance. In cases where the claim is that the test provides personal utility, but no change in clinical utility (health outcomes), the personal utility must be demonstrated so MSAC can appropriately consider whether there is evidence to support the claim.

For example, genetic testing to identify X-linked retinitisu pigmentosa may be important to identify whether a child is likely to become blind or not. Although no treatment is currently available to prevent or treat the vision loss, the prognostic information provided by a genetic test identifying the risk of this condition, may be used for social reasons, to allow education and career planning (Burke, Laberge & Press 2010).

Knowledge may also have a negative direct impact on patients and their family members. Psychological distress is common, and in the absence of an appropriate treatment or preventive measure for the identified condition, could mean that a test is more harmful than beneficial. Pre- and post-testing information on the level of distress for those with positive results, negative results, and

ambiguous results may be important to consider. For example, research has demonstrated that testing for Huntington disease and hereditary cancers causes more stress if there is an ambiguous result, than if there is a positive result (Botkin et al. 2010; Korngiebel et al. 2019). There may also be guilt for passing on heritable diseases, or survivor guilt, for unaffected siblings (Botkin et al. 2010). The assessment must therefore not presume that additional information is always beneficial, and provide evidence of the impact. If this cannot be provided for the target condition and intervention, consider incorporating evidence from other populations, and discuss the applicability.

A prognostic test may be used to identify a subtype of cancer, but not alter the treatment used by the patients. However, knowing of the prognosis of the patient is of value for the patient and family in planning their lives (such as accommodation decisions, employment decisions, and end of life planning).

TG 28.2 How to assess other utility evidence

If a claim is made that the key benefit of a test is for the personal utility of the test, the evidence supporting this claim needs to be provided. Quantitative evidence which allows MSAC to consider the proportion of patients who experience this benefit (and the magnitude of such benefit), should be provided where available. It is acknowledged that this form of evidence may not always be possible to generate or identify.

If a claim is made that information gained from the proposed test is of value for the information itself, then a review of qualitative research should be undertaken and discussed. If no quantitative or qualitative evidence is identified to support these statements, then stakeholder input (submitted during the consultation phase) may be of use.

If the benefits of the test are seen in sectors outside of health, they would be unable to be incorporated into a cost-utility analysis, but could potentially be incorporated into a cost-consequences analysis. If the evidence is qualitative in nature, without being able to quantify the proportion of patients/families etc who experience the outcome, the personal utility could still be outlined for MSAC to consider. Various strategies are available for synthesising qualitative evidence (Dixon-Woods et al. 2005).

Where a test may have ethically debatable implications (such as termination of an affected foetus), MSAC have considered that the value of the test is to provide the prospective parents with information, rather than the avoidance of the birth of a child with a disorder. If there is any evidence that parents feel the test result is valuable to them, then this should be presented, although the actual data on the number of prospective parents who choose one option over another is not as relevant.

Benefits and harms to the individual patient (personal utility) should be presented separately from benefits and harms to other people.

For advice regarding the cost-consequences approach to claims of other utility, see Technical Guidance 17.

Technical Guidance 29 Other relevant considerations

TG 29.1 Introduction

Additional information relevant to decision making may be captured in Section 5. Evidence presented in this section should be clearly presented and reasoned. Where possible, evidence should be generated using high-quality methods or sourced systematically. Inadequately supported claims, or the presentation of evidence prone to bias as a result of the methods of generation or collection, will be difficult to interpret.

TG 29.2 Ethical analysis

"Implementing new technologies in health care can have morally relevant consequences. Technologies carry with them values that can challenge the current mores and attitudes of society. Every HTA requires many value-based decisions to be made during the assessment process" (Saarni et al. 2011).

Ethical issues could also be thought of as respects in which the technology may be more or less valuable in ways not captured by standard measures of safety, effectiveness and cost-effectiveness. They may also include sensitivities that stakeholders should keep in mind when funding, producing, delivering or using the health technology. These sensitivities or complexities may mean that attached to funding are guidelines for practice that address the sensitivities.

According to the EUnetHTA core model[t], ethical issues that may be relevant in the assessment of the health technology include the following.

- **The balance of benefits and harms**. Consider whether there may be hidden or unintended consequences or implications for patients or other groups that are not captured in assessments of safety and effectiveness. Consider whether there may have been any ethical obstacles impeding those assessments.
- **Autonomy**. Consider whether the value of the health technology is augmented by its impact on the autonomy of patients or other groups (i.e. the right and capacity of people to direct their own lives). Is the health technology of particular value because it helps to restore or promote the autonomy of patients who are particularly vulnerable or who may have a reduced capacity for exercising autonomy? Conversely, could the health technology result in a reduction in autonomy and thereby be of lesser value? Are there additional interventions that may be required to ensure the target population can provide valid (i.e. informed and voluntary) consent to receive or refuse the health technology?
- **Respect**. Could the health technology have implications for matters of human dignity, stigma, privacy, or moral, religious or cultural conviction or tradition? Could widespread use of the health technology change our conception of certain persons? (Hofmann 2005)
- **Equity**. Could the implementation of the health technology have impacts on equitable access to care across the target population? Would government subsidy of the medical intervention affect the distribution of health care resources in problematic ways?
- **The HTA and its implications**. Are there ethical issues or implications relating to the choice of endpoints, populations or comparators in the assessment? Are there ethical problems relating to the assumptions in the economic evaluation? In particular, are there important respects in which the health technology may be of greater or lesser value that the economic evaluation has not captured, as per Table 22?

[t] https://eunethta.eu/hta-core-model/

Table 22 Questions to prompt consideration of whether there is likely to be greater or lesser value from the proposed technology than captured in Section 2 and 3 of the assessment (Norheim et al. 2014).

Disease and intervention	
Severity	Have you considered whether the intervention has special value because of the severity of the health condition (present and future health gap) that the intervention targets?
Realization of potential	Have you considered whether the intervention has more value than the effect size alone suggests on the grounds that it does the best possible for a patient group for whom restoration to full health is not possible?
Past health loss	Have you considered whether the intervention has special value because it targets a group that has suffered significant past health loss (e.g. chronic disability)?
Social groups	
Socioeconomic status	Have you considered whether the intervention has special value because it can reduce disparities in health associated with unfair inequalities in wealth, income or level of education?
Area of living	Have you considered whether the intervention has special value because it can reduce disparities in health associated with area of living (e.g. rural and remote areas)?
Gender	Have you considered whether the intervention will reduce disparities in health associated with gender?
Race, ethnicity, religion and sexual orientation	Have you considered whether the intervention may disproportionally affect groups characterised by race, ethnicity, religion, and sexual orientation?
Protection against the financial and social effects of ill health	
Economic productivity	Have you considered whether the intervention has special value because it enhances welfare to the individual and society by protecting the target population's productivity?
Care for others	Have you considered whether the intervention has special value because it enhances welfare by protecting the target population's ability to take care of others?
Catastrophic health expenditures	Have you considered whether the intervention has special value because it reduces catastrophic health expenditures for the target population?

Report any ethical issues that were identified by the applicant or in the PICO Confirmation, and raise any further ethical issues identified during the assessment.

Provide a description of each of the issues. For each issue, state whether the issue is unique to the proposed medical service, or whether it would arise with other health technologies that are already available. An issue may be unique even if it is associated with another health technology if the proposed technology is intended for a different population (for example, in terms of disease, gender, age, life-expectancy, stage of disease, quality of life, co-morbidities).

For issues deemed to be unique, briefly summarise any recent literature describing an ethical analysis that has been performed in a health care system / population that is applicable to Australia.

A formal ethical analysis is not likely to be required for the Assessment Report. Where ethical concerns are considered to be substantial, these should be identified for the consideration of the committee and raised as major concerns.

TG 29.3 Organisational aspects

In the domain of Organisational Aspects (according to EUnetHTA), ways in which different kinds of resources need to be organised when implementing a technology, and the consequences that flow from it in the organisation in the health care system, are considered. These resources may include human skills, attitudes, material artefacts, money, work culture, etc. Examples of organisational issues that arise are work processes and patient/participant flow, quality and sustainability assurance, communication and co-operation, centralisation, acceptance of a technology, and managerial structure.

Until recently, in many countries, organisational aspects have not been a visible part of the HTAs (the focus has been on clinical aspects). However, there is a growing focus that organisational issues may be of crucial importance (e.g. in the introduction of digital mammography to replace film mammography, the benefits of digital storage versus storage of film, and the fact that producers and developers of film were unlikely to be available for much longer were key considerations for policy makers). Organisational aspects in HTA may identify challenges and barriers in implementing health technologies, which could influence decision making by policy makers. The assessment of organisational issues faces a challenge in the aspect that the findings are more context-dependent and less transferable than other aspects of an HTA, due to the complexity of different health care systems and processes. The objectives and criteria in an organisational analysis are less pre-determined than for example an analysis on clinical effectiveness.

EUnetHTA have suggested that the organisational domain should include five topics, each containing two to six issues (15 in total). These topics generally represent the most important organisational issues, however their relevance depends on the specific intervention and needs.

The different topics discussed in the organisational domain in the EUnetHTA core model are:

1. **Health delivery process**. How does the technology affect the current work processes? What kind of patient/participant flow is associated with the new technology? What kind of involvement must be mobilised for patients, caregivers and others? What kind of process ensures proper education and training of staff? What kinds of co-operation and communication of activities must be mobilised? And in what way is the quality assurance and monitoring system of the new technology organised?
2. **Structure of the health care system.** How do de-centralisation or centralisation requirements influence the implementation of the technology? What are the processes ensuring access to the new technology for patients?
3. **Process-related costs.** What are the costs of processes relating to acquisition and setting up the new technology? How does the technology modify the need for other technologies and use of resources? What is the likely budget impact of implementing the technology?
4. **Management.** What management problems and opportunities are attached to the technology? Who decides which people are eligible to receive the test/treatment and on what basis?
5. **Culture.** How is the technology accepted? How are other interest groups taken into consideration during the implementation of this technology?

Each type of health technology being assessed comes with its own specific organisational challenges. The implementation of a new diagnostic test could significantly change the number of patients diagnosed with a certain condition. This could substantially influence the number of patients who need to be treated, which poses challenges/changes to the health care system. If the intervention is a screening test, it also has various specific implications depending on the objective (e.g. is the focus on a new screening test, on the population eligible for screening, or on changing test delivery?).

It may not be feasible identifying and addressing all organisational issues by doing a systematic literature search or doing research in the form of surveys or interviews (although this would be ideal). However, potential organisational issues that are identified during the various steps in the assessment process (e.g. in the application phase, during the development of the PICO, during the literature search, during the assessment of safety and effectiveness) should be discussed. Qualitative research identified during the literature searches may assist in understanding how patients perceive health, how they make decisions regarding health service usage, and in understanding the culture of communities in relation to implementing changes and overcoming barriers. Guidelines could also be a source of information regarding identifying possible implementation and organisational issues.

TG 29.4 Patient and social aspects

EUnetHTA describes the patients and social aspects domain in their core model (EUnetHTA core model 3.0). In the model, patient aspects relate to issues relevant to patients / individuals and caregivers, whereas social aspects are related to social groups (i.e. specific groupings of patients or individuals that may be of specific interest in an HTA), e.g. elderly, ethnic minorities, immigrants, people living in rural areas, people with disabilities, etc. Individuals who receive the intervention and their caregivers can provide unique perspectives on the experiences, attitudes, expectations and values regarding the intervention and regarding health, illness, service delivery and treatments. This can inform HTA.

A common component of HTAs internationally is the incorporation of primary qualitative research into the HTA process, which does not fit within the current process of MSAC HTAs. If assessment groups or applicants wish to incorporate qualitative evidence into “other relevant considerations” they are free to do so.

The parallel process of consulting with the public (patients, carers, citizens etc) is outside the scope of the Technical Guidelines. However, public consultation processes used during PICO Confirmation development seek to gather the perspective of affected parties (patients, caregivers, healthcare workers). Their input should be included in the assessment report, as well as any other information gathered from the literature or through consultation with those with “lived experience”.

The patient perspective can contribute to an understanding of the value (positive or negative) of the proposed health technology. This is discussed in more detail under ‘personal utility’ (Technical Guidance 28).

If resources allow, a summary of qualitative and quantitative evidence on the perspective of the patient and other stakeholders would be beneficial. This could be evidence from patients, individuals, caregivers and social groups about the burden of living with the condition being studied, experiences of current practice or current health technologies, or experiences with and expectations of the health technology being studied. This evidence could be found by searching for published (mostly qualitative) systematic reviews or studies. When searching for this kind of evidence, it’s useful to include psychological and/or sociological databases in the search strategy (e.g. Applied Social Sciences Index and Abstracts, PsycInfo, ISI Web of Science).

TG 29.5 Legal aspects

The objective of the legal aspects domain is to assist HTA assessors in detecting rules and regulations which should be taken into consideration when evaluation the implications and consequences of implementing a health technology. As technologies rapidly change, policy and decision-makers are required to know the legal implications of implementing or not implementing a technology. Some of the legal aspects outlined by the EUnetHTA core model may be especially important to assess for digital technologies (i.e. ensuring patient data are appropriately secure).

Some elements relating to legal issues are also relevant in the ethical domain. The different topics that may be addressed when discussing legal aspects of the health technology are:

1. **Autonomy of the patient.** Who is allowed to give informed consent for the technology for minors and incompetent persons? What legal requirements are in place for providing appropriate information to the user? How should this be addressed when implementing the health technology?
2. **Privacy of the patient.** Does the use of the health technology produce additional information that is not directly related to the patient's care and may violate their right to privacy? What do the laws say regarding informing relatives about the results? What are the laws regarding the security of patient data and how should this be addressed when implementing the health technology?
3. **Equality in health care.** What do laws require regarding processes or resources which would facilitate equal access to the health technology? What are the consequences of rules and regulations around equal access to the technology?
4. **Ethical aspects.** Does the implementation of the technology affect the realisation of basic human rights? And can the implementation of the health technology give rise to ethical challenges that have not yet been considered in existing laws and regulations?
5. **Authorisation and safety.** What rules and laws are present around safety of the technology and how should this be addressed when implementing the health technology?
6. **Ownership and liability.** What should be known and reported about intellectual property rights and potential licencing fees, and about the regulations regarding the manufacturers guarantee? Who would be responsible if the health technology fails or provides false results? What would the medicolegal consequences be for 'overrelying' on test results?
7. **Regulation of the market.** What are the laws surrounding price control mechanisms of the technology? Are there laws or regulation regarding acquisition and use of the technology? Are there legal restrictions for marketing the health technology to users? What should be known around legal issues in cases of new technologies where current legislation is not directly applicable? Are there concerns about conflicts of interest regarding the preparation of binding rules and their implementation?

Consider which of the topics concerning legal issues stated above are relevant for the proposed health intervention (if any) and these should be discussed.

Information on legal aspects of health technologies may be found in journals (e.g. Health Economics, Policy and Law, Medical Law International, Medical Law review, Medicine and Law), or websites such as the Federal Register of Legislation (https://www.legislation.gov.au/Home).

TG 29.6 Environmental aspects

If there are particular concerns regarding the environmental impact of the proposed health technology, or a key benefit in the way the proposed technology reduces the environmental impact of the comparator, these should be outlined (e.g. if there is a reduction in the amount of radioactive waste generated, or a reduction of emissions related to transportation or in the manufacturing process).

TG 29.7 Basis for any claim for the 'rule of rescue'

The four factors described below apply in exceptional circumstances and are particularly influential in favour of listing. When all four factors apply concurrently, this is called the 'rule of rescue':

- No alternative exists in Australia to treat patients with the specific circumstances of the medical condition meeting the criteria of the restriction. This means that there are no nonpharmacological or pharmacological interventions for these patients.

- The medical condition defined by the requested restriction is severe, progressive and expected to lead to premature death. The more severe the condition, or the younger the age at which a person with the condition might die, or the closer a person with the condition is to death, the more influential the rule of rescue might be in MSAC's consideration.
- The medical condition defined by the requested restriction applies to only a very small number of patients. Again, the fewer the patients, the more influential the rule of rescue might be in MSAC's consideration. However, MSAC is also mindful that the MBS is a community-based scheme and cannot cater for individual circumstances.
- The proposed technology provides a worthwhile clinical improvement sufficient to qualify as a rescue from the medical condition. The greater the rescue, the more influential the rule of rescue might be in MSAC's consideration.

As with other relevant factors, the rule of rescue supplements, rather than substitutes for, the evidence-based consideration of comparative cost-effectiveness. A decision on whether the rule of rescue is relevant is only necessary if MSAC would be inclined to reject a submission because of its consideration of comparative cost-effectiveness (and any other relevant factors). In such a circumstance, if MSAC concludes the rule of rescue is relevant, it would then consider whether this is sufficiently influential in favour of a recommendation to list, that MSAC would then reverse a decision not to recommend listing if the rule of rescue were not relevant.

This guidance on the rule of rescue is deliberately kept narrow. Although there are relevant arguments for broadening the guidance, MSAC is concerned that doing so would reduce the relative influence of the rule of rescue if it is applied to a broader set of eligible submissions. In other words, the greater the proportion of submissions that the rule of rescue is applied to, the smaller its average impact in favour of listing across the identified submissions.

One issue that has arisen concerning the rule of rescue is that a second health technology to treat the medical condition (that is considered to meet requirements of the rule) would not be suitable for this consideration. This is because, by definition, the second technology does not meet the essential first factor (i.e. that there is currently no alternative intervention). This causes difficulty if listing of the second technology is sought on a cost-minimisation basis.

References

Altman, DG 1991, 'Practical statistics for medical research Chapman and Hall', *London and New York*.

Altman, DG, McShane, LM, Sauerbrei, W & Taube, SE 2012, 'Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): explanation and elaboration', *PLoS Med*, vol. 9, no. 5, p. e1001216.

Bagust, A & Beale, S 2014, 'Survival analysis and extrapolation modeling of time-to-event clinical trial data for economic evaluation: an alternative approach', *Med Decis Making*, vol. 34, no. 3, Apr, pp. 343-351.

Balshem, H, Helfand, M, Schunemann, HJ, Oxman, AD, Kunz, R, Brozek, J, Vist, GE, Falck-Ytter, Y, Meerpohl, J, Norris, S & Guyatt, GH 2011, 'GRADE guidelines: 3. Rating the quality of evidence', *J Clin Epidemiol*, vol. 64, no. 4, Apr, pp. 401-406.

Barnett, AG, van der Pols, JC & Dobson, AJ 2005, 'Regression to the mean: what it is and how to deal with it', *Int J Epidemiol*, vol. 34, no. 1, Feb, pp. 215-220.

Barton, P, Bryan, S & Robinson, S 2004, 'Modelling in the economic evaluation of health care: selecting the appropriate approach', *J Health Serv Res Policy*, vol. 9, no. 2, Apr, pp. 110-118.

Becker, F, Van El, CG, Ibarreta, D, Zika, E, Hogarth, S, Borry, P, Cambon-Thomsen, A, Cassiman, JJ, Evers-Kiebooms, G, Hodgson, S, Janssens, ACJW, Kaariainen, H, Krawczak, M, Kristoffersson, U, Lubinski, J, Patch, C, Penchaszadeh, VB, Read, A, Rogowski, W, Sequeiros, J, Tranebjaerg, L, Van Langen, IM, Wallace, H, Zimmern, R, Schmidtke, J & Cornel, MC 2011, 'Genetic testing and common disorders in a public health framework: How to assess relevance and possibilities', *European Journal of Human Genetics*, vol. 19, no. SUPPL. 1, pp. S6-S44.

Begg, CB & Mazumdar, M 1994, 'Operating characteristics of a rank correlation test for publication bias', *Biometrics*, pp. 1088-1101.

Bell, KJ, Glasziou, PP, Hayen, A & Irwig, L 2014, 'Criteria for monitoring tests were described: validity, responsiveness, detectability of long-term change, and practicality', *J Clin Epidemiol*, vol. 67, no. 2, Feb, pp. 152-159.

Bell, KJ, Irwig, L, Craig, JC & Macaskill, P 2008, 'Use of randomised trials to decide when to monitor response to new treatment', *BMJ*, vol. 336, no. 7640, Feb 16, pp. 361-365.

Bell, KJL, Doust, J, Glasziou, P, Cullen, L, Harris, IA, Smith, L, Buchbinder, R & Barratt, A 2019, 'Recognizing the Potential for Overdiagnosis: Are High-Sensitivity Cardiac Troponin Assays an Example?', *Ann Intern Med*, vol. 170, no. 4, Feb 19, pp. 259-261.

Bonetti, M & Gelber, RD 2000, 'A graphical method to assess treatment-covariate interactions using the Cox model on subsets of the data', *Stat Med*, vol. 19, no. 19, Oct 15, pp. 2595-2609.

Booth, A, Lewin, S, Glenton, C, Munthe-Kaas, H, Toews, I, Noyes, J, Rashidian, A, Berg, RC, Nyakang'o, B, Meerpohl, JJ & Team, GR-CC 2018, 'Applying GRADE-CERQual to qualitative evidence synthesis findings-paper 7: understanding the potential impacts of dissemination bias', *Implement Sci*, vol. 13, no. Suppl 1, Jan 25, p. 12.

Bossuyt, P & Leeflang, M 2008, 'Chapter 6: developing criteria for including studies', *Cochrane handbook for systematic reviews of diagnostic test accuracy version 0.4 [updated September 2008]. The Cochrane Collaboration*.

Botkin, JR, Teutsch, SM, Kaye, CI, Hayes, M, Haddow, JE, Bradley, LA, Szegda, K & Dotson, WD 2010, 'Outcomes of interest in evidence-based evaluations of genetic tests', *Genetics in Medicine*, vol. 12, no. 4, pp. 228-235.

Brazier, J, Ara, R, Azzabi, I, Busschbach, J, Chevrou-Severac, H, Crawford, B, Cruz, L, Karnon, J, Lloyd, A, Paisley, S & Pickard, AS 2019, 'Identification, Review, and Use of Health State Utilities in Cost-Effectiveness Models: An ISPOR Good Practices for Outcomes Research Task Force Report', *Value Health*, vol. 22, no. 3, Mar, pp. 267-275.

Briggs, AH, Weinstein, MC, Fenwick, EA, Karnon, J, Sculpher, MJ, Paltiel, AD & Force, I-SMGRPT 2012, 'Model parameter estimation and uncertainty analysis: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force Working Group-6', *Med Decis Making*, vol. 32, no. 5, Sep-Oct, pp. 722-732.

Burke, W, Laberge, AM & Press, N 2010, 'Debating clinical utility', *Public Health Genomics*, vol. 13, no. 4, pp. 215-223.

Cahan, EM, Hern, ez-Boussard, T, Thadaney-Israni, S & Rubin, DL 2019, 'Putting the data before the algorithm in big data addressing personalized healthcare', *NPJ Digit Med*, vol. 2, 2019, p. 78.

Carter, SM, Degeling, C, Doust, J & Barratt, A 2016, 'A definition and ethical evaluation of overdiagnosis', *Journal of Medical Ethics*, vol. 42, no. 11, pp. 705-714.

Ciani, O, Buyse, M, Drummond, M, Rasi, G, Saad, ED & Taylor, RS 2017, 'Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward', *Value Health*, vol. 20, no. 3, Mar, pp. 487-495.

Cleophas, TJ & Zwinderman, AH 2009, 'Meta-analyses of diagnostic studies', *Clin Chem Lab Med*, vol. 47, no. 11, pp. 1351-1354.

Cochran, WG 1954, 'The combination of estimates from different experiments', *Biometrics*, vol. 10, no. 1, pp. 101-129.

Croft, P, Altman, DG, Deeks, JJ, Dunn, KM, Hay, AD, Hemingway, H, LeResche, L, Peat, G, Perel, P, Petersen, SE, Riley, RD, Roberts, I, Sharpe, M, Stevens, RJ, Van Der Windt, DA, Von Korff, M & Timmis, A 2015, 'The science of clinical practice: Disease diagnosis or patient prognosis? Evidence about "what is likely to happen" should shape clinical practice', *BMC Medicine*, vol. 13, no. 1.

De Groot, S, Rijnsburger, AJ, Versteegh, MM, Heymans, JM, Kleijnen, S, Redekop, WK & Verstijnen, IM 2015, 'Which factors may determine the necessary and feasible type of effectiveness evidence? A mixed methods approach to develop an instrument to help coverage decision-makers', *BMJ Open*, vol. 5, no. 7.

Deeks, JJ, Macaskill, P & Irwig, L 2005, 'The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed', *Journal of clinical epidemiology*, vol. 58, no. 9, pp. 882-893.

Detsky, AS, Naglie, G, Krahn, MD, Redelmeier, DA & Naimark, D 1997, 'Primer on medical decision analysis: Part 2--Building a tree', *Med Decis Making*, vol. 17, no. 2, Apr-Jun, pp. 126-135.

Dixon-Woods, M, Agarwal, S, Jones, D, Young, B & Sutton, A 2005, 'Synthesising qualitative and quantitative evidence: A review of possible methods', *Journal of Health Services Research and Policy*, vol. 10, no. 1, pp. 45-53.

Doust, J 2010, 'Qualification versus validation of biomarkers', *Scand J Clin Lab Invest Suppl*, vol. 242, pp. 40-43.

Doust, J & Glasziou, P 2013, 'Monitoring in clinical biochemistry', *Clinical Biochemist Reviews*, vol. 34, no. 2, pp. 85-92.

Doust, JA, Bell, KJ & Glasziou, PP 2020, 'Potential Consequences of Changing Disease Classifications', *JAMA*, vol. 323, no. 10, pp. 921-922.

Egger, M, Smith, GD, Schneider, M & Minder, C 1997, 'Bias in meta-analysis detected by a simple, graphical test', *Bmj*, vol. 315, no. 7109, pp. 629-634.

Fazel, S & Wolf, A 2018, 'Selecting a risk assessment tool to use in practice:a 10-point guide', *Evid Based Ment Health*, vol. 21, no. 2, May, pp. 41-43.

Fleurence, RL & Hollenbeak, CS 2007, 'Rates and probabilities in economic modelling', *PharmacoEconomics*, vol. 25, no. 1, pp. 3-6.

Fuchs, S, Olberg, B, Panteli, D, Perleth, M & Busse, R 2017, 'HTA of medical devices: Challenges and ideas for the future from a European perspective', *Health Policy*, vol. 121, no. 3, pp. 215-229.

Ghabri, S, Stevenson, M, Moller, J & Caro, JJ 2019, 'Trusting the Results of Model-Based Economic Analyses: Is there a Pragmatic Validation Solution?', *PharmacoEconomics*, vol. 37, no. 1, Jan, pp. 1-6.

Glas, AS, Lijmer, JG, Prins, MH, Bonsel, GJ & Bossuyt, PM 2003, 'The diagnostic odds ratio: a single indicator of test performance', *J Clin Epidemiol*, vol. 56, no. 11, Nov, pp. 1129-1135.

Glasziou, P, Chalmers, I, Rawlins, M & McCulloch, P 2007, 'When are randomised trials unnecessary? Picking signal from noise', *BMJ*, vol. 334, no. 7589, pp. 349-351.

Glasziou, P, Irwig, L & Deeks, JJ 2008, 'When should a new test become the current reference standard?', *Annals of Internal Medicine*, vol. 149, no. 11, pp. 816-821.

Glasziou, PP, Irwig, L, Heritier, S, Simes, RJ, Tonkin, A & Investigators, LS 2008, 'Monitoring cholesterol levels: measurement error or true change?', *Ann Intern Med*, vol. 148, no. 9, May 6, pp. 656-661.

Gonzalez-McQuire, S, Dimopoulos, MA, Weisel, K, Bouwmeester, W, Hajek, R, Campioni, M, Bennison, C, Xu, W, Pantiri, K, Hensen, M, Terpos, E & Knop, S 2019, 'Development of an Initial Conceptual Model of Multiple Myeloma to Support Clinical and Health Economics Decision Making', *MDM Policy Pract*, vol. 4, no. 1, Jan-Jun, p. 2381468318814253.

Grieve, R, Hawkins, N & Pennington, M 2013, 'Extrapolation of survival data in cost-effectiveness analyses: improving the current state of play', *Med Decis Making*, vol. 33, no. 6, Aug, pp. 740-742.

Guyatt, GH, Oxman, AD, Kunz, R, Brozek, J, Alonso-Coello, P, Rind, D, Devereaux, PJ, Montori, VM, Freyschuss, B, Vist, G, Jaeschke, R, Williams, JW, Jr., Murad, MH, Sinclair, D, Falck-Ytter, Y, Meerpohl, J, Whittington, C, Thorlund, K, Andrews, J & Schunemann, HJ 2011, 'GRADE guidelines: 6. Rating the quality of evidence--imprecision', *J Clin Epidemiol*, vol. 64, no. 12, Dec, pp. 1283-1293.

Guyatt, GH, Oxman, AD, Kunz, R, Woodcock, J, Brozek, J, Helfand, M, Alonso-Coello, P, Falck-Ytter, Y, Jaeschke, R, Vist, G, Akl, EA, Post, PN, Norris, S, Meerpohl, J, Shukla, VK, Nasser, M, Schunemann, HJ & Group, GW 2011, 'GRADE guidelines: 8. Rating the quality of evidence--indirectness', *J Clin Epidemiol*, vol. 64, no. 12, Dec, pp. 1303-1310.

Guyatt, GH, Oxman, AD, Kunz, R, Woodcock, J, Brozek, J, Helfand, M, Alonso-Coello, P, Glasziou, P, Jaeschke, R, Akl, EA, Norris, S, Vist, G, Dahm, P, Shukla, VK, Higgins, J, Falck-Ytter, Y, Schunemann, HJ & Group, GW 2011, 'GRADE guidelines: 7. Rating the quality of evidence--inconsistency', *J Clin Epidemiol*, vol. 64, no. 12, Dec, pp. 1294-1302.

Guyatt, GH, Oxman, AD, Montori, V, Vist, G, Kunz, R, Brozek, J, Alonso-Coello, P, Djulbegovic, B, Atkins, D, Falck-Ytter, Y, Williams, JW, Jr., Meerpohl, J, Norris, SL, Akl, EA & Schunemann, HJ 2011, 'GRADE guidelines: 5. Rating the quality of evidence--publication bias', *J Clin Epidemiol*, vol. 64, no. 12, Dec, pp. 1277-1282.

Guyatt, GH, Oxman, AD, Sultan, S, Glasziou, P, Akl, EA, Alonso-Coello, P, Atkins, D, Kunz, R, Brozek, J, Montori, V, Jaeschke, R, Rind, D, Dahm, P, Meerpohl, J, Vist, G, Berliner, E, Norris, S, Falck-Ytter, Y, Murad, MH, Schunemann, HJ & Group, GW 2011, 'GRADE guidelines: 9. Rating up the quality of evidence', *J Clin Epidemiol*, vol. 64, no. 12, Dec, pp. 1311-1316.

Guyatt, GH, Oxman, AD, Vist, G, Kunz, R, Brozek, J, Alonso-Coello, P, Montori, V, Akl, EA, Djulbegovic, B, Falck-Ytter, Y, Norris, SL, Williams, JW, Jr., Atkins, D, Meerpohl, J & Schunemann, HJ 2011, 'GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias)', *J Clin Epidemiol*, vol. 64, no. 4, Apr, pp. 407-415.

Haddow, JE & Palomaki, GE 2004, 'ACCE: a model process for evaluating data on emerging genetic tests', *Human genome epidemiology: A scientific foundation for using genetic information to improve health and prevent disease*, pp. 217-233.

Haji Ali Afzali, H, Bojke, L & Karnon, J 2018, 'Model Structuring for Economic Evaluations of New Health Technologies', *PharmacoEconomics*, vol. 36, no. 11, Nov, pp. 1309-1319.

Haji Ali Afzali, H, Karnon, J, Theou, O, Beilby, J, Cesari, M & Visvanathan, R 2019, 'Structuring a conceptual model for cost-effectiveness analysis of frailty interventions', *PloS one*, vol. 14, no. 9, p. e0222049.

Hajjaj, FM, Salek, MS, Basra, MK & Finlay, AY 2010, 'Non-clinical influences on clinical decision-making: a major challenge to evidence-based practice', *Journal of the Royal Society of Medicine*, vol. 103, no. 5, pp. 178-187.

Hayden, JA, van der Windt, DA, Cartwright, JL, Cote, P & Bombardier, C 2013, 'Assessing bias in studies of prognostic factors', *Ann Intern Med*, vol. 158, no. 4, Feb 19, pp. 280-286.

Higgins, JP, Thompson, SG, Deeks, JJ & Altman, DG 2003, 'Measuring inconsistency in meta-analyses', *Bmj*, vol. 327, no. 7414, pp. 557-560.

Hinde, S & Spackman, E 2015, 'Bidirectional Citation Searching to Completion: An Exploration of Literature Searching Methods', *PharmacoEconomics*, vol. 33, no. 1, January 01, pp. 5-11.

Hofmann, B 2005, 'Toward a procedure for integrating moral issues in health technology assessment', *Int J Technol Assess Health Care*, vol. 21, no. 3, Summer, pp. 312-318.

Hoyer, A & Kuss, O 2015, 'Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas', *Stat Med*, vol. 34, no. 11, May 20, pp. 1912-1924.

Hoyer, A & Kuss, O 2018, 'Meta-analysis for the comparison of two diagnostic tests to a common gold standard: A generalized linear mixed model approach', *Stat Methods Med Res*, vol. 27, no. 5, May, pp. 1410-1421.

Janes, H, Pepe, MS, Bossuyt, PM & Barlow, WE 2011, 'Measuring the performance of markers for guiding treatment decisions', *Ann Intern Med*, vol. 154, no. 4, Feb 15, pp. 253-259.

Jonas, DE, Wilt, TJ, Taylor, BC, Wilkins, TM & Matchar, DB 2012, 'Chapter 11: challenges in and principles for conducting systematic reviews of genetic tests used as predictive indicators', *J Gen Intern Med*, vol. 27 Suppl 1, Jun, pp. S83-93.

Justice, AC, Covinsky, KE & Berlin, JA 1999, 'Assessing the generalizability of prognostic information', *Annals of Internal Medicine*, vol. 130, no. 6, pp. 515-524.

Kaltenthaler, E, Tappenden, P, Paisley, S & Squires, H 2011, *NICE DSU Technical Support Document 13: Identifying and reviewing evidence to inform the conceptualisation and population of cost-effectiveness models*, <http://nicedsu.org.uk/wp-content/uploads/2016/03/TSD-13-model-parameters.pdf>.

Karnon, J & Haji Ali Afzali, H 2014, 'When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES', *PharmacoEconomics*, vol. 32, no. 6, Jun, pp. 547-558.

Karnon, J, Stahl, J, Brennan, A, Caro, JJ, Mar, J & Moller, J 2012, 'Modeling using discrete event simulation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-4', *Med Decis Making*, vol. 32, no. 5, Sep-Oct, pp. 701-711.

Karnon, J & Vanni, T 2011, 'Calibrating models in economic evaluation', *PharmacoEconomics*, vol. 29, no. 1, pp. 51-62.

Korngiebel, DM, Zech, JM, Chappelle, A, Burke, W, Carline, JD, Gallagher, TH & Fullerton, SM 2019, 'Practice Implications of Expanded Genetic Testing in Oncology', *Cancer Invest*, vol. 37, no. 1, pp. 39-45.

Latimer, NR 2013, 'Survival analysis for economic evaluations alongside clinical trials--extrapolation with patient-level data: inconsistencies, limitations, and a practical guide', *Med Decis Making*, vol. 33, no. 6, Aug, pp. 743-754.

Latimer, NR, Abrams, KR, Lambert, PC, Crowther, MJ, Wailoo, AJ, Morden, JP, Akehurst, RL & Campbell, MJ 2014, 'Adjusting survival time estimates to account for treatment switching in

randomized controlled trials--an economic evaluation context: methods, limitations, and recommendations', *Med Decis Making*, vol. 34, no. 3, Apr, pp. 387-402.

Lee, J, Kim, KW, Choi, SH, Huh, J & Park, SH 2015, 'Systematic Review and Meta-Analysis of Studies Evaluating Diagnostic Test Accuracy: A Practical Review for Clinical Researchers-Part II. Statistical Methods of Meta-Analysis', *Korean J Radiol*, vol. 16, no. 6, Nov-Dec, pp. 1188-1196.

Lewin, S, Bohren, M, Rashidian, A, Munthe-Kaas, H, Glenton, C, Colvin, CJ, Garside, R, Noyes, J, Booth, A & Tunçalp, Ö 2018, 'Applying GRADE-CERQual to qualitative evidence synthesis findings—paper 2: how to make an overall CERQual assessment of confidence and create a Summary of Qualitative Findings table', *Implementation Science*, vol. 13, no. 1, p. 10.

Lewin, S, Glenton, C, Munthe-Kaas, H, Carlsen, B, Colvin, CJ, Gülmezoglu, M, Noyes, J, Booth, A, Garside, R & Rashidian, A 2015, 'Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual)', *PLoS Medicine*, vol. 12, no. 10.

Liberati, A, Altman, DG, Tetzlaff, J, Mulrow, C, Gotzsche, PC, Ioannidis, JP, Clarke, M, Devereaux, PJ, Kleijnen, J & Moher, D 2009, 'The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration', *J Clin Epidemiol*, vol. 62, no. 10, Oct, pp. e1-34.

Macaskill, P, Gatsonis, C, Deeks, J, Harbord, R & Takwoingi, Y 2010, 'Chapter 10: Analysing and Presenting Results', in J Deeks, P Bossuyt & C Gatsonis (eds), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0*, The Cochrane Collaboration.

Mathes, T & Pieper, D 2019, 'An algorithm for the classification of study designs to assess diagnostic, prognostic and predictive test accuracy in systematic reviews', *Systematic reviews*, vol. 8, no. 1, p. 226.

Merlin, T, Lehman, S, Hiller, JE & Ryan, P 2013, 'The "linked evidence approach" to assess medical tests: a critical analysis', *Int J Technol Assess Health Care*, vol. 29, no. 3, Jul, pp. 343-350.

Merlin, T, Weston, A & Tooher, R 2009, 'Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'', *BMC Med Res Methodol*, vol. 9, p. 34.

Millard, LA, Flach, PA & Higgins, JP 2016, 'Machine learning to assist risk-of-bias assessments in systematic reviews', *Int J Epidemiol*, vol. 45, no. 1, Feb, pp. 266-277.

Moher, D, Liberati, A, Tetzlaff, J, Altman, DG & Group, P 2009, 'Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement', *J Clin Epidemiol*, vol. 62, no. 10, Oct, pp. 1006-1012.

Moons, KGM 2010, 'Criteria for Scientific Evaluation of Novel Markers: A Perspective', *Clinical Chemistry*, vol. 56, no. 4, pp. 537-541.

Morel, T & Cano, SJ 2017, 'Measuring what matters to rare disease patients - reflections on the work by the IRDiRC taskforce on patient-centered outcome measures', *Orphanet J Rare Dis*, vol. 12, no. 1, Nov 2, p. 171.

Morton, V & Torgerson, DJ 2005, 'Regression to the mean: Treatment effect without the intervention', *Journal of Evaluation in Clinical Practice*, vol. 11, no. 1, pp. 59-65.

Moshi, MR, Tooher, R & Merlin, T In submission, 'A health technology assessement evaluative module for evaluating mobile medical applications'.

Moynihan, R, Barratt, AL, Buchbinder, R, Carter, SM, Dakin, T, Donovan, J, Elshaug, AG, Glasziou, PP, Maher, CG, McCaffery, KJ & Scott, IA 2018, 'Australia is responding to the complex challenge of overdiagnosis', *Med J Aust*, vol. 209, no. 8, Oct 15, pp. 332-334.

NHMRC 2009, *NHMRC additional levels of evidence and grades for recommendations for developers of guidelines.*, National Health and Medical Research Council, Canberra, ACT, viewed 1/10/10 2010, <http://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/evidence_statement_form.pdf>.

Norheim, OF, Baltussen, R, Johri, M, Chisholm, D, Nord, E, Brock, D, Carlsson, P, Cookson, R, Daniels, N, Danis, M, Fleurbaey, M, Johansson, KA, Kapiriri, L, Littlejohns, P, Mbeeli, T, Rao, KD, Edejer, TT & Wikler, D 2014, 'Guidance on priority setting in health care (GPS-Health): the inclusion of equity criteria not captured by cost-effectiveness analysis', *Cost Eff Resour Alloc*, vol. 12, p. 18.

Nyaga, VN, Arbyn, M & Aerts, M 2014, 'Metaprop: a Stata command to perform meta-analysis of binomial data', *Arch Public Health*, vol. 72, no. 1, p. 39.

Parikh, RB, Obermeyer, Z & Navathe, AS 2019, 'Regulation of predictive analytics in medicine', *Science*, vol. 363, no. 6429, Feb 22, pp. 810-812.

Park, S, Kim, Y, Lee, J, Yoo, S & Kim, C 2019, 'Ethical challenges regarding artificial intelligence in medicine from the perspective of scientific editing and peer review', *Sci Ed*, vol. 6, no. 2, pp. 91-98.

Pepe, MS, Janes, H, Longton, G, Leisenring, W & Newcomb, P 2004, 'Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker', *American Journal of Epidemiology*, vol. 159, no. 9, pp. 882-890.

Pitrou, I, Boutron, I, Ahmad, N & Ravaud, P 2009, 'Reporting of Safety Results in Published Reports of Randomized Controlled Trials', *Archives of Internal Medicine*, vol. 169, no. 19, pp. 1756-1761.

Pletcher, MJ & Pignone, M 2011, 'Evaluating the clinical utility of a biomarker: A review of methods for estimating health impact', *Circulation*, vol. 123, no. 10, pp. 1116-1124.

Reitsma, JB, Glas, AS, Rutjes, AW, Scholten, RJ, Bossuyt, PM & Zwinderman, AH 2005, 'Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews', *J Clin Epidemiol*, vol. 58, no. 10, Oct, pp. 982-990.

Riley, RD, Sauerbrei, W & Altman, DG 2009, 'Prognostic markers in cancer: The evolution of evidence from single studies to meta-analysis, and beyond', *British Journal of Cancer*, vol. 100, no. 8, pp. 1219-1229.

Roberts, M, Russell, LB, Paltiel, AD, Chambers, M, McEwan, P, Krahn, M & Force, I-SMGRPT 2012, 'Conceptualizing a model: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--2', *Value Health*, vol. 15, no. 6, Sep-Oct, pp. 804-811.

Rotily, M & Roze, S 2013, 'What is the impact of disease prevalence upon health technology assessment?', *Best Practice & Research Clinical Gastroenterology*, vol. 27, no. 6, pp. 853-865.

Royston, P 2001, 'Flexible parametric alternatives to the Cox model, and more', *Stata Journal*, vol. 1, no. 1, pp. 1-28.

Royston, P & Lambert, PC 2011, 'Flexible parametric survival analysis using Stata: beyond the Cox model'.

Saarni, SI, Braunack-Mayer, A, Hofmann, B & van der Wilt, GJ 2011, 'Different methods for ethical analysis in health technology assessment: an empirical study', *Int J Technol Assess Health Care*, vol. 27, no. 4, Oct, pp. 305-312.

Scott, IA, Cook, D, Coiera, EW & Richards, B 2019, 'Machine learning in clinical practice: prospects and pitfalls', *Med J Aust*, vol. 211, no. 5, Sep, pp. 203-205 e201.

Segal, JB 2012, 'Chapter 3: choosing the important outcomes for a systematic review of a medical test', *J Gen Intern Med*, vol. 27 Suppl 1, Jun, pp. S20-27.

Sendak, M, Gao, M, Nichols, M, Lin, A & Balu, S 2019, 'Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities', *EGEMS (Wash DC)*, vol. 7, no. 1, Jan 24, p. 1.

Shea, BJ, Reeves, BC, Wells, G, Thuku, M, Hamel, C, Moran, J, Moher, D, Tugwell, P, Welch, V, Kristjansson, E & Henry, DA 2017, 'AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both', *BMJ*, vol. 358, Sep 21, p. j4008.

Siebert, U, Alagoz, O, Bayoumi, AM, Jahn, B, Owens, DK, Cohen, DJ & Kuntz, KM 2012, 'State-transition modeling: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-3', *Med Decis Making*, vol. 32, no. 5, Sep-Oct, pp. 690-700.

Singh, S, Chang, SM, Matchar, DB & Bass, EB 2012, 'Chapter 7: grading a body of evidence on diagnostic tests', *J Gen Intern Med*, vol. 27 Suppl 1, Jun, pp. S47-55.

Siontis, KC, Siontis, GC, Contopoulos-Ioannidis, DG & Ioannidis, JP 2014, 'Diagnostic tests often fail to lead to changes in patient outcomes', *Journal of Clinical Epidemiology*, vol. 67, no. 6, pp. 612-621.

Staub, LP, Lord, SJ, Simes, RJ, Dyer, S, Houssami, N, Chen, RY & Irwig, L 2012, 'Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation', *BMC Med Res Methodol*, vol. 12, Feb 14, p. 12.

Sterne, JA, Hernan, MA, Reeves, BC, Savovic, J, Berkman, ND, Viswanathan, M, Henry, D, Altman, DG, Ansari, MT, Boutron, I, Carpenter, JR, Chan, AW, Churchill, R, Deeks, JJ, Hrobjartsson, A, Kirkham, J, Juni, P, Loke, YK, Pigott, TD, Ramsay, CR, Regidor, D, Rothstein, HR, Sandhu, L, Santaguida, PL, Schunemann, HJ, Shea, B, Shrier, I, Tugwell, P, Turner, L, Valentine, JC, Waddington, H, Waters, E, Wells, GA, Whiting, PF & Higgins, JP 2016, 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions', *BMJ*, vol. 355, Oct 12, p. i4919.

Steyerberg, EW, Moons, KG, van der Windt, DA, Hayden, JA, Perel, P, Schroter, S, Riley, RD, Hemingway, H, Altman, DG & Group, P 2013, 'Prognosis Research Strategy (PROGRESS) 3: prognostic model research', *PLoS Med*, vol. 10, no. 2, p. e1001381.

Swets, JA 1988, 'Measuring the accuracy of diagnostic systems', *Science*, vol. 240, no. 4857, Jun 3, pp. 1285-1293.

Tabberer, M, Gonzalez-McQuire, S, Muellerova, H, Briggs, AH, Rutten-van Molken, M, Chambers, M & Lomas, DA 2017, 'Development of a Conceptual Model of Disease Progression for Use in Economic Modeling of Chronic Obstructive Pulmonary Disease', *Med Decis Making*, vol. 37, no. 4, May, pp. 440-452.

Tacconelli, E 2010, 'Systematic reviews: CRD's guidance for undertaking reviews in health care', *The Lancet Infectious Diseases*, vol. 10, no. 4, p. 226.

Takwoingi, Y, Guo, B, Riley, RD & Deeks, JJ 2017, 'Performance of methods for meta-analysis of diagnostic test accuracy with few studies or sparse data', *Stat Methods Med Res*, vol. 26, no. 4, Aug, pp. 1896-1911.

Tappenden, P & Chilcott, JB 2014, 'Avoiding and identifying errors and other threats to the credibility of health economic models', *PharmacoEconomics*, vol. 32, no. 10, Oct, pp. 967-979.

Trikalinos, TA & Balion, CM 2012, 'Chapter 9: options for summarizing medical test performance in the absence of a "gold standard"', *J Gen Intern Med*, vol. 27 Suppl 1, Jun, pp. S67-75.

Trikalinos, TA, Balion, CM, Coleman, CI, Griffith, L, Santaguida, PL, Vandermeer, B & Fu, R 2012, 'Chapter 8: meta-analysis of test performance when there is a "gold standard"', *J Gen Intern Med*, vol. 27 Suppl 1, Jun, pp. S56-66.

Tsoi, B, O'Reilly, D, Jegathisawaran, J, Tarride, JE, Blackhouse, G & Goeree, R 2015, 'Systematic narrative review of decision frameworks to select the appropriate modelling approaches for health economic evaluations', *BMC Res Notes*, vol. 8, p. 244.

van Enst, WA, Ochodo, E, Scholten, RJPM, Hooft, L & Leeflang, MM 2014, 'Investigation of publication bias in meta-analyses of diagnostic test accuracy: a meta-epidemiological study', *BMC Medical Research Methodology*, vol. 14, no. 1, 2014/05/23, p. 70.

Vemer, P, Corro Ramos, I, van Voorn, GA, Al, MJ & Feenstra, TL 2016, 'AdViSHE: A Validation-Assessment Tool of Health-Economic Models for Decision Makers and Model Users', *PharmacoEconomics*, vol. 34, no. 4, Apr, pp. 349-361.

Whiting, PF, Rutjes, AW, Westwood, ME, Mallett, S, Deeks, JJ, Reitsma, JB, Leeflang, MM, Sterne, JA & Bossuyt, PM 2011, 'QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies', *Ann Intern Med*, vol. 155, no. 8, Oct 18, pp. 529-536.

Whyte, S, Walsh, C & Chilcott, J 2011, 'Bayesian calibration of a natural history model with application to a population model for colorectal cancer', *Medical Decision Making*, vol. 31, no. 4, pp. 625-641.

Williams, C, Lewsey, JD, Mackay, DF & Briggs, AH 2017, 'Estimation of Survival Probabilities for Use in Cost-effectiveness Analyses: A Comparison of a Multi-state Modeling Survival Analysis Approach with Partitioned Survival and Markov Decision-Analytic Modeling', *Med Decis Making*, vol. 37, no. 4, May, pp. 427-439.

Woods, B, Sideris, E, Palmer, S, Latimer, N & Soares, M 2017, *NICE DSU Technical Support Document 19. Partitioned Survival Analysis for Decision Modelling in Health Care: A Critical Review*, <http://nicedsu.org.uk/technical-support-documents/partitioned-survival-analysis-tsd/>.

Wurcel, V, Cicchetti, A, Garrison, L, Kip, MMA, Koffijberg, H, Kolbe, A, Leeflang, MMG, Merlin, T, Mestre-Ferrandiz, J, Oortwijn, W, Oosterwijk, C, Tunis, S & Zamora, B 2019, 'The Value of Diagnostic Information in Personalised Healthcare: A Comprehensive Concept to Facilitate Bringing This Technology into Healthcare Systems', *Public Health Genomics*, vol. 22, no. 1-2, pp. 8-15.

Yuste, R, Goering, S, Arcas, BAY, Bi, G, Carmena, JM, Carter, A, Fins, JJ, Friesen, P, Gallant, J, Huggins, JE, Illes, J, Kellmeyer, P, Klein, E, Marblestone, A, Mitchell, C, Parens, E, Pham, M, Rubel, A, Sadato,

N, Sullivan, LS, Teicher, M, Wasserman, D, Wexler, A, Whittaker, M & Wolpaw, J 2017, 'Four ethical priorities for neurotechnologies and AI', *Nature*, vol. 551, no. 7679, Nov 8, pp. 159-163.

Appendix 1 Assessment frameworks

Assessment framework for non-inferiority based on change in management

If the proposed test reports on a different parameter, an assessment of comparative test accuracy is not possible. However, if the clinical interpretation of the results is the same, then concordance on the clinical interpretation (or categorisation) is required. This would be evident if the same management decisions were made for the same test subjects regardless of the test used. Evidence that there would be no changes in management as compared with the comparator test may permit non-inferior health outcomes to be inferred (assuming there are no differences in test safety).

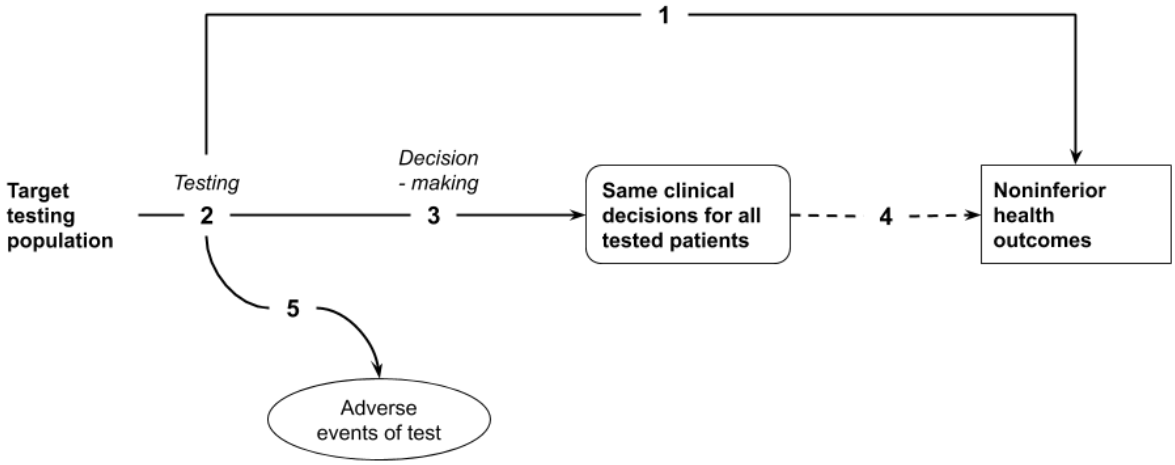


Figure 26 Assessment framework that has been truncated at decision-making with the inference that concordant decision making will result in the same health outcomes.

Assessment questions for a claim of non-inferiority for a test based on concordance of decision making (Figure 26)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of [the proposed test] in the [target testing population] result in the same or better [key health outcomes - e.g. survival, quality of life] compared with [the main comparator]? (If adequate direct from test to health outcomes evidence is available, go to Assessment question 5).

LINKED EVIDENCE

2. *Not able to compare test accuracy as tests report on different parameter or biomarker.*
3. Does the use of [the proposed test] in the [target testing population] result in the same clinical decisions compared with [the main comparator]?
    a. Is [the proposed test strategy] concordant with [the main comparator test strategy] in the categorisation of the test results? (If different parameters are measured by the proposed test and the main comparator, they may still be categorised similarly – ie, into low, moderate or high risk, for example). Is the categorisation of test results a validated tool for decision making? Is there evidence that clinicians will behave the same way regardless of the test used to inform the categorisation?

b. Is [the proposed test strategy] concordant with [the main comparator test strategy] for clinical decision making. How would decision-making non-concordance impact on patient health outcomes?

4. *Inference that concordant decision making with existing comparator test strategy will result in non-inferior health outcomes.*
5. What are the harms of [the proposed test] and of [the main comparator]?

Assessment frameworks for triage testing

If the proposed test is a triage test, such that it reduces the number of subjects that will require a more definitive test, then the final categorisation of patients or the final decision making following the proposed test stragety is of interest. This means that those ruled out from having the condition by the triage test will need to be followed to see whether they ultimately have the condition and/or present for the definitive test at a later date.

Where the final classification of patients following triage testing plus definitive testing versus definitive testing alone is identical (or concordant), the framework may be truncated at the final classification (and a claim of non-inferiority is appropriate). If the final classification of patients is not concordant, or there are differences in the timing of when the final classification is made, then the framework will need to include the impacts of subsequent steps in a linked evidence approach to establish the impact of the triage test on health outcomes.

A key consideration with triage testing might be that a greater proportion of the test population will adhere to the triage test, particularly if the triage test is less invasive than the definitive test (i.e. a blood test versus a biopsy). Uptake or adherence or compliance is unlikely to be informed by a direct trial of the proposed testing strategy vs the comparator test (as adherence within a trial setting is often artificially high). A second consideration will relate to the pathway for patients with the condition (true positives) who are determined to be negative by the triage test. Typically, there is a delay in the eventual diagnosis, and the impact of this delay due to the reduced accuracy of the triage test should be explored.

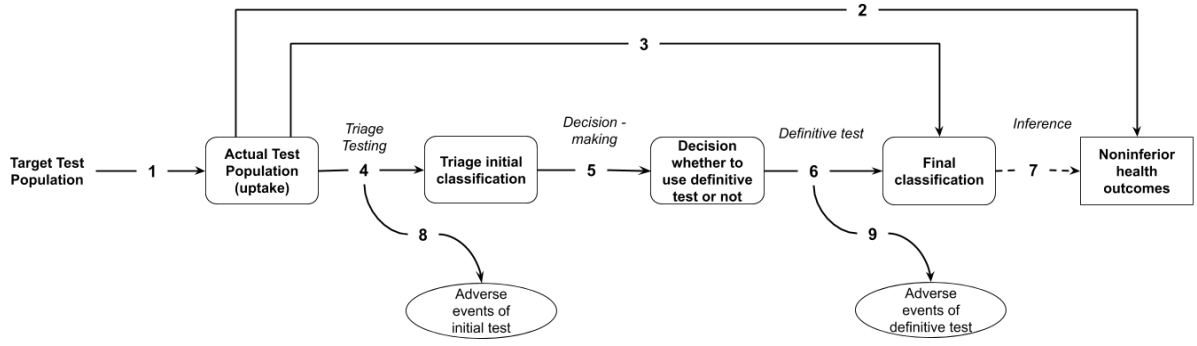


Figure 27 Assessment framework that has been truncated at the final classification of test results (following a triage and definitive test) with the inference that the final classification will result in the same health outcomes.

Assessment questions for a claim of non-inferiority for the use of a triage test compared with the main comparator (definitive test) (Figure 27)

1. Does the use of the proposed triage test change the uptake rate for testing compared with the current testing regimen?

a. If there are differences in the populations who receive the triage test compared with the comparator, are these differences likely to be associated with the result of the tests (ie, are high risk subjects more likely to receive the test, or low risk subjects more likely to be non-compliant).
b. What is the impact of differences in uptake rates on the final test results?

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

2. Does the use of [the proposed triage test] in the [target / uptake testing population] result in the same or better [key health outcomes – e.g. survival, quality of life] compared with [the main comparator / definitive test]? (If adequate direct from test to health outcomes evidence is available, go to Assessment question 8).

LINKED EVIDENCE

3. Does the use of [the proposed triage test] in the [testing population] result in the same categorisation (positive / negative, presence / absence, high risk / low risk) or the same clinical decisions compared with [the main comparator]?
4. What is the test accuracy of [the proposed triage test] compared with [a reference standard / the comparator definitive test]? What is the nature of the incorrect classifications ie ratio of false positives to false negatives from using the triage test? What are the clinical consequences of the false negative triage result? *In an asymptomatic population, when the triage test is a screening test, the consequences of false positive testing is also important.*
5. Does information from [the proposed triage test] result in a change in investigative thinking and change in the individuals who are referred for the definitive test?
6. *[If the definitive test is not established practice or is not the reference standard]* What is the test accuracy of [the definitive test] in terms of sensitivity and specificity?
7. *Inference that the same final classification of patients (all patients classified using the definitive test / comparator are classified similarly if the triage test were introduced) will result in the same health outcomes.*
8. What are the harms of [the proposed triage test]?
9. What are the harms of [the definitive test / comparator]?

Assessment framework for a more definitive test

If the proposed test is replacing more than one test (ie, the proposed test is more definitive or able to test multiple parameters concurrently), then the added value to decision-making relates to either the final categorisation of patients (described above for triage testing), or the decision making following the proposed test compared with the decision making following the full test strategy it is intended to replace.

Assessment framework for monitoring

An investigative technology intended to be used for monitoring is assessed in a similar way to diagnostic tests. However, a key difference is that monitoring tests are commonly repeated at intervals to detect a condition that would affect clinical decision making.

The assessment framework remains similar to that for other types of investigative technologies; however, the assessment questions are expanded to incorporate the characteristics of a monitoring technology.

If monitoring is followed by a confirmatory test, the first change in management would be to undertake this test, and a subsequent step would include management decisions for treatment. Both steps in the change in management are necessary for the evaluation of a monitoring test.

A key uncertainty regarding monitoring involves whether the monitoring test results in a change in management compared with current clinical practice. For this reason, more emphasis should be placed on robust evidence to support a change in management compared with current practice. As the change in management may have occurred following an earlier detection of a condition than would be detected using standard clinical practice, there is a risk of a change in the spectrum of the disease being detected, and the applicability of treatment evidence that has been derived from standard clinical practice is a concern. As with the assessment of all tests, comparative direct from test to health outcomes evidence is preferred.

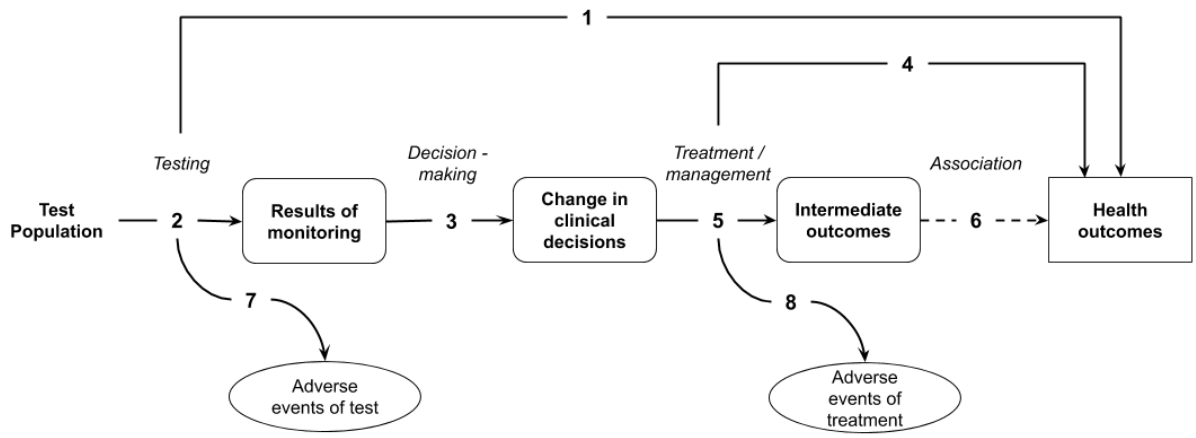


Figure 28 Assessment framework adapted for a monitoring test

Assessment questions for a claim of superiority relating to the use of a monitoring test (Figure 28)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of the test strategy in place of the current test strategy (comparator) result in the claimed superior health outcomes?
    a. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or clinical utility standard?

LINKED EVIDENCE

2. How does the information from the proposed test differ from that of the comparator? What is the concordance of the findings from the proposed test relative to the comparator? What is the accuracy of the proposed test (against a relevant reference standard) compared with the comparator?
    a. If there is a change in the timing at which information becomes available, the results of monitoring will contain a time component. Is there evidence that the proposed test will result in a change in the timing of the detection of a condition compared with the comparator?
    b. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or

clinical utility standard? Concordance of both the test results and the testing protocol (periodicity of the testing).

3. Does the availability of new information from the proposed test result in a change in management of the patient (compared to the information gained from the comparator)?
    a. The change in information provided by the proposed test may represent different test results and/or different timing of test results.
4. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention, or differences in the timing of treatment / intervention), result in the claimed health outcomes?
5. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention), result in the claimed surrogate outcomes?
    a. Has the treatment/management been provided to a population with the same spectrum of disease that the proposed test identifies? Is it biologically plausible that the treatment/management will be as effective in the population with this spectrum of disease? *For monitoring tests, there is some concern when the proposed test detects patients earlier than the comparator as the treatment effect evidence may be based on population that was identified at a later time point.*
6. Is the observed change in surrogate outcomes associated with a concomitant change in the claimed health outcomes?
7. What are the adverse events associated with the proposed test strategy and the comparative test strategy?
    a. Include downstream adverse events associated with any changes to subsequent testing (such as confirmatory testing).
8. What are the adverse events associated with the treatments / interventions that lead from the management decisions informed by the test and by the comparator?

Assessment framework for multifactorial algorithms, black-box and self-learning algorithms

Algorithms are a broad category of investigative technologies that commonly include risk scores, nomograms, prognostic scores, and more recently, self-learning software that may process genomic data, physiological data or imaging data to provide a diagnosis or an estimate of risk of a condition. The key characteristic of an algorithm is that the method of categorisation of patients (how the algorithm weights measured parameters to provide an estimate) is not easily, or cannot be understood. For this reason, in some circumstances, there is no obvious reference standard against which the algorithm can be compared. For all types of algorithms, and particularly those for which the final step of the linked evidence approach (treatment) has uncertain applicability, direct from test to health outcomes evidence is preferred.

The approach to the assessment of an algorithm varies depending on:

- The clinical claim and the purpose of the test – if the test is prognostic or predictive, it will require longitudinal data, whereas a diagnostic test *may* require cross-sectional data.
- The presence of a reference standard – if the algorithm is being used to detect something that can be verified clinically (such as the presence of a tumour on imaging), then the test accuracy can be established, and the applicability of downstream changes may be assessed.

In the absence of a relevant reference standard, the accuracy of the test cannot be determined, and direct from test to health outcomes evidence will be required.

- The applicability of the training and validation dataset to the Australian population. Algorithm results by subgroups of interest are required to establish whether there is a risk of the algorithm failing in different populations.
- The applicability of the final step (intervention or treatment) in a linked evidence approach to the population identified by the algorithm (eg, change in spectrum of disease).

A subsequent step to the assessment of self-learning dynamic algorithms will relate to the safe guards that are in place to ensure that the algorithm remains applicable as it continues to evolve once it is available in clinical practice.

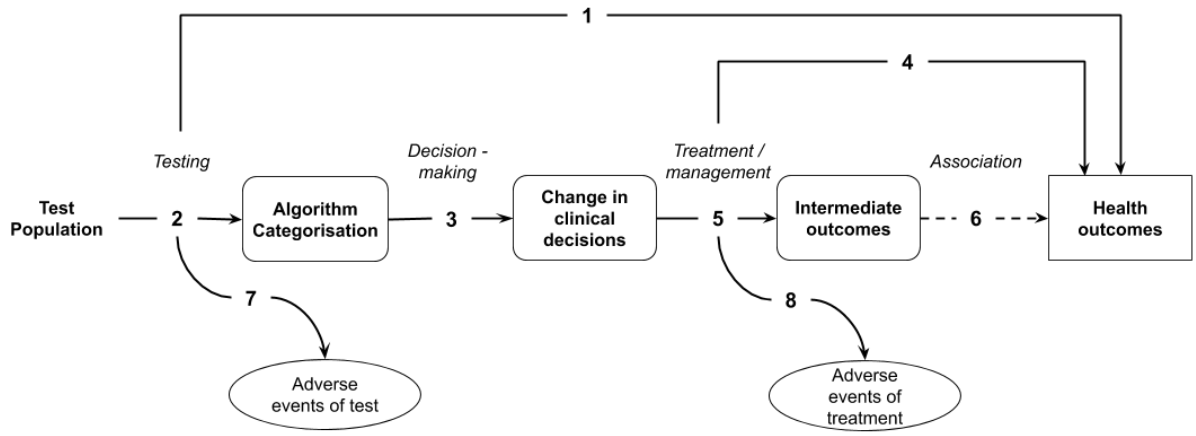


Figure 29 Assessment framework adapted for a multifactorial algorithm

Assessment questions for a claim of superiority relating to the use of a multifactorial algorithm (Figure 29)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of the test strategy in place of the current test strategy (comparator) result in the claimed superior health outcomes?
    a. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or clinical utility standard? Concordance is measured on the categorisation made by the algorithm. Where an appropriate reference standard is unavailable, a very high concordance would be required to determine that additional tests should be eligible for the same funding arrangements. In the absence of very high concordance or robust concordance data, alternative tests cannot leverage the direct from test to health outcomes evidence of the clinical utility standard.

LINKED EVIDENCE

2. How does the information from the proposed test differ from that of the comparator? That is, how do patient classifications differ using the algorithm versus standard practice?
    a. If there are multiple tests in clinical practice likely to be able to utilise the same funding arrangements, are these tests concordant with the proposed test and/or clinical utility standard?

    b. Is the data that was used to construct and validate the algorithm applicable to the target setting? Are there any populations missing from the training or validation datasets? Are there any key differences in the test accuracy across population subgroups?
    c. Is there a risk that the classification of patients will change over time (is the algorithm dynamic)? What safeguards are in place to ensure that changes to the algorithm are appropriate, or represent an improvement in accuracy?
3. Does the availability of new information from the proposed test result in a change in management of the patient (compared to the information gained from the comparator / standard practice)?
4. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention), result in the claimed health outcomes?
    a. Has the treatment/management been provided to a population with the same spectrum of disease that the proposed test identifies?
5. Do the differences in the management derived from the proposed test, relative to the comparator (eg differences in treatment / intervention), result in the claimed surrogate outcomes?
6. Is the observed change in surrogate outcomes associated with a concomitant change in the claimed health outcomes?
7. What are the adverse events associated with the proposed test strategy and the comparative test strategy?
8. What are the adverse events associated with the treatments / interventions that lead from the management decisions informed by the test and by the comparator?

Assessment framework for universal screening tests

Due to the low prevalence of conditions that are tested for in population or universal screening, there is a high risk that the harms of the tests and the harms associated with false positives may outweigh the value of earlier detection. A further complication relates to the detection of conditions prior to clinical suspicion, based on symptoms or high risk parameters. Earlier detection of the disease may have little influence on treatment outcomes, or may result in earlier treatments without any evidence that earlier intervention is more effective.

Consequently, universal or asymptomatic screening tests require direct from test to health outcomes evidence of the utility of the screening test. This may include direct from test to health outcomes evidence, or direct from test to an intermediate outcome evidence that can be robustly translated to a health outcome.

Further considerations for screening tests are presented in TG 15.1.

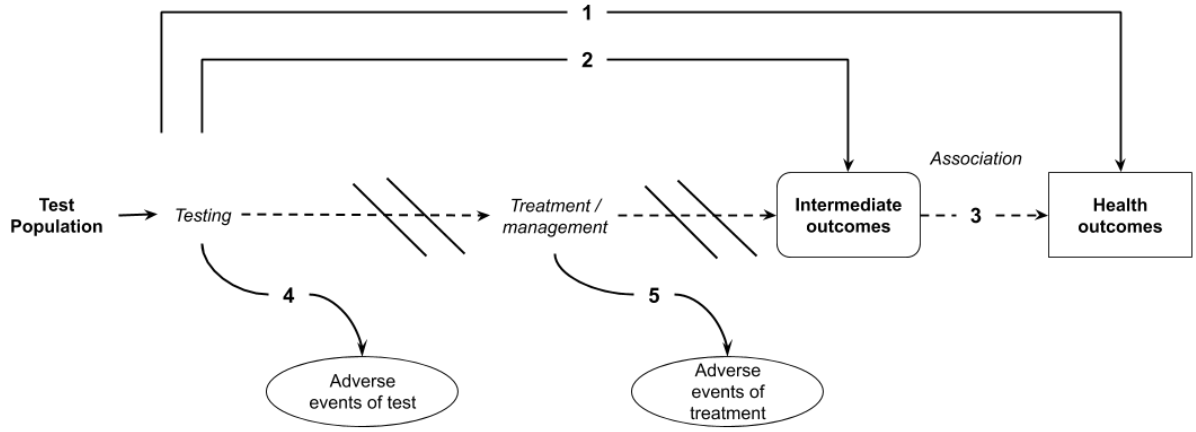


Figure 30 Assessment framework adapted for a population or universal screening test

Assessment questions for a claim of superiority relating to the use of a population or universal screening test (Figure 30)

DIRECT FROM TEST TO HEALTH OUTCOMES EVIDENCE

1. Does the use of the test strategy in place of the current test strategy (comparator) result in the claimed superior health outcomes?
2. Does the use of the test strategy in place of the current test strategy (comparator) result in a change in intermediate or surrogate outcomes?
3. Is there evidence to support the validity of the translation of the intermediate outcomes to health outcomes for the populations identified by the proposed test?
4. What are the adverse events associated with the proposed test strategy and the comparative test strategy?
5. What are the adverse events associated with the treatments / interventions that lead from the management decisions informed by the test and by the comparator?

Appendix 2 Literature search methods

The primary objective of Section 2 is to provide the "best evidence" to answer the assessment questions presented in Section 1. The purpose of this appendix is to detail the search methods for ensuring all relevant studies have been included in the clinical evaluation, as is appropriate for a full HTA.

Abbreviated search methods may be appropriate for a facilitated listing or streamlined approach (see TG 5.2).

Search terms for therapeutic technologies

In most cases, the process of identifying the "best evidence" is to identify all randomised trials that compare the proposed therapeutic technology with the main comparator(s). However, there are situations where no RCTs will be available and other 'lower level' study designs are acceptable[u]. For instance, if the intervention has already been used for a number of years, if it is unfeasible to perform randomised studies (e.g. if equipoise is lacking), if patients will not participate in a randomised study (given the already available data), if there is no alternative treatment, or if it is a rare condition, only lower level evidence may be available (De Groot et al. 2015). If no direct randomised comparisons are located, indirect comparisons of randomised trials and/or nonrandomised studies will be required.

If no comparative evidence is identified, then non-comparative literature should be assessed for both the intervention and the comparator to allow MSAC to make conclusions regarding the comparative effectiveness and safety of the technologies.

Search strategy

Develop a search strategy to address the assessment questions presented in Section 1.

An appropriate search strategy should have the following characteristics:

- Involve a search for studies involving either the proposed therapeutic technology or the main comparator(s). This approach would permit the identification of trials required to perform an indirect comparison. If comparative evidence is identified (a direct RCT), it may not be necessary to perform a search for the main comparator.
- Is not restricted by study type.
- When a device is involved in the proposed therapeutic technology, the search is not restricted by a particular manufacturer's device. In some circumstances, where the MBS item is intended to be restricted to a specific device, restricting the search to this device may be appropriate.

Search filters and additional search terms should be used with caution. Additional search filters or terms may include:

- Randomised studies or systematic reviews
    - It is generally not appropriate to limit searches to include only randomised trials or systematic reviews.

[u] For information on Levels of Evidence, see the NHMRC website.

    - If, during the PICO confirmation or during scoping searches, high quality randomised studies are identified that adequately address the assessment questions, a filter that excludes nonrandomised studies from the search may be appropriate (e.g. Cochrane Highly Sensitive Search Strategies[v]).
    - Limiting to only randomised studies or systematic reviews should only be necessary if the number of references retrieved is otherwise unmanageable.
    - A justification for restricting the search strategy to randomised studies or systematic reviews would include, 1) the number of citations retrieved in the absence of the filter, 2) careful consideration of the applicability of the high level evidence to the Australian setting (see Supplementary Evidence below).
- Population
    - The study includes participants with characteristics that overlap with those of the target population. In general, this approach is only relevant if the proposed medical service and/or the main comparator(s) are used across multiple populations / indications that are not relevant to the assessment. Care should be taken when excluding studies in different populations or indications, particularly if adverse events may be generalisable to the proposed use of the health technology.
- Date range
    - The search period may be limited to the earliest use of the proposed health technology and the main comparator(s). If comparative evidence is identified, the search period may only need to extend to the earliest use of the proposed medical service.
    - If the technology or comparator have changed over time, consider whether limiting the search period to recent literature is justifiable (such as the last 10 years).
    - If a relevant and high quality systematic review is identified, a search period designed to identify new information may be appropriate.
- Language
    - Articles published in English or with reliable translations
- Publication type
    - Conference abstracts would only be accepted as evidence under exceptional circumstances

For most assessments, a broad search strategy is appropriate. In circumstances where a large number of studies are identified which address the critical outcomes, studies of lower quality may be excluded from the search results. Search strategies that are limited by study type are generally not appropriate to remove lower quality studies.

If a focused search strategy is used, explain why. Justify why the included literature is adequate to address the effectiveness and safety of the proposed medical service and that important studies have not been missed. Higher level evidence (randomised controlled trials or systematic reviews) may not report long term safety or all relevant patient outcomes, or may not be applicable to the Australian clinical setting. Therefore, describe any gaps in the evidence, or uncertainties, associated with applying a focused search. Describe methods used to supplement the high level evidence, if required.

Supplementary evidence

Although randomised trials may provide the most robust estimates of comparative effectiveness and safety, they may not, by themselves, provide the "best evidence" or complete evidence. Well conducted nonrandomised studies or indirect comparisons may be informative and/or constitute the "best evidence" for addressing the assessment question(s).

[v] http://handbook.cochrane.org/chapter_6/6_4_11_1_the_cochrane_highly_sensitive_search_strategies_for.htm

Consider including supplementary evidence if the use of the included highest level of evidence has the following concerns:

- Inadequate applicability to the Australian setting
- Omission of important population groups
- Differences in the circumstances of use of the proposed medical service or comparator(s)
- Omission of important patient relevant outcomes
- Does not report long term effects or safety
- Does not assess user-proficiency

Explain the decision to include supplementary evidence beyond the highest level of evidence.

Present all the relevant search strategies in a technical appendix to the Assessment Report.

Presentation of the search strategy

The clear presentation of the search terms improves the transparency of the approach. Tabulating the search terms can assist with presentation. Present a table of the search terms for each bibliographic database or data source, and for each search (if more than one search is performed). The presentation of the search terms must explain how each of the terms interacts with other terms (i.e. Boolean operators).

Table 23 Search terms for the literature review

Category	Description	Search terms
Study design (if justified)	[insert description of category]	[e.g. Cochrane Highly Sensitive Search Strategies for identifying randomised trials in MEDLINE, or MeSH and text word terms for nonrandomised study designs]
Population	[insert description of category]	[include MeSH terms, text words and synonyms for the target population/disease/condition]
Intervention	[insert description of category]	[include known proprietary and nonproprietary names, MeSH terms]
Comparator	[insert description of category]	[include known proprietary and nonproprietary names, MeSH terms]

MeSH = medical subject headings

Search terms for investigative technologies

The "best evidence" for assessing a test would include studies that randomise participants to receive the proposed test or the test comparator and report on final health outcomes. However, these direct from test to health outcomes evidence studies are uncommon, and additional searches are likely to be required to complete a linked evidence approach.

Search strategy

Develop a search strategy to address the assessment questions presented in Section 1.

PICO assessment questions, based on the assessment framework, will usually include questions relating to direct evidence of the test impact on health outcomes, as well as linked steps, including test accuracy, test concordance, change in management, and the impact of change in management on health outcomes.

A broad search strategy that includes the terms to identify the index test will identify studies (if available) that report on direct from test to health outcomes evidence compared to the comparator,

and, if taking a linked evidence approach, the test related evidence up to change in management. A separate search will be required to identify the impact of change in management (treatment, interventions) on health outcomes (see treatment related search below). If direct from test to health outcomes evidence is available for the intervention, but does not provide information on the incremental clinical utility (i.e. direct evidence of health outcomes after the comparator test strategy), then a separate search could be performed for this information, in order to perform an indirect comparison.

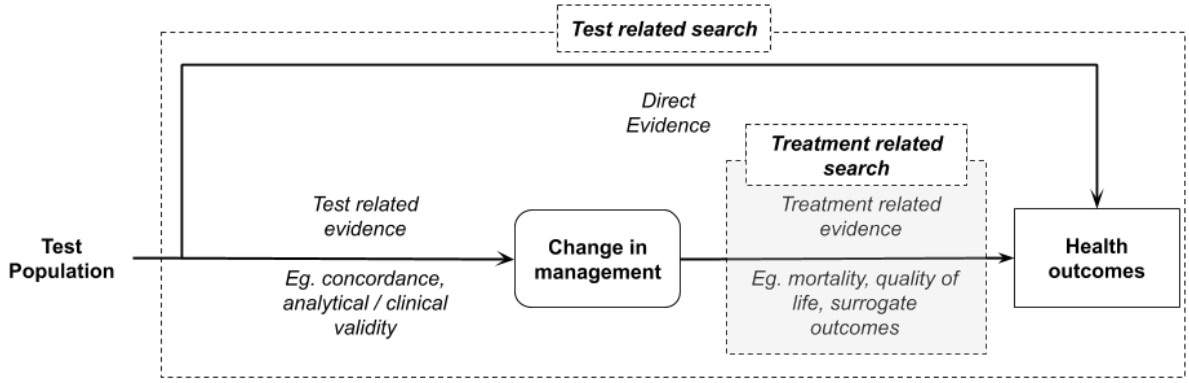


Figure 31 Topics covered by search involving the terms to identify the index test

Test related search (health impact of test, impact test has on management of patient, and accuracy of test)

An appropriate search strategy would usually have the following characteristics:

- Include terms to identify the proposed test
- Include terms to identify the comparator (if studies directly comparing the proposed test and comparator are not available)
- Include terms to identify the target condition to be detected
- Is not restricted by study type.

It is preferable to start with a broad search strategy and narrow the number of relevant includes during the screening phase of the assessment.

Search filters and additional search terms should be used with caution. In circumstances where the number of results retrieved are large and unmanageable in the timeframes available, there are several options for using search filters:

- Randomised studies or systematic reviews
    - It is generally not appropriate to limit searches to include only randomised trials or systematic reviews. These study types are unlikely to provide information required to undertake a linked approach.
- Date range
    - The search period may be limited to the earliest use of the proposed test and the main comparator(s). If comparative evidence is identified, the search period may only need to extend to the earliest use of the proposed technology.
    - If the proposed technology or comparator have changed over time, consider whether limiting the search period to recent literature is justifiable (such as the last 10 years).
    - If a relevant and high quality systematic review is identified, a search period designed to identify only new information may be appropriate.
- Language

    - Articles published in English or with reliable translations.
- Publication type
    - Conference abstracts would only be accepted as evidence under exceptional circumstances.

It is generally not appropriate to limit search strategies by study type. If the number of relevant studies identified is large, it is preferable to exclude studies of lower quality at the screening phase of the literature review.

The following guidance is relevant to searches designed for identifying test related articles:

- Use a wide range of text words for each of the concepts (including synonyms, related terms and variant spellings). Filters for specific terms should be avoided.
- Use truncations and wildcards to capture variations in terms.
- Customise search strategies for each database (either manually or using a tool such as Polyglot Search Syntax Translator[w])
- Do not rely on controlled vocabulary (subject headings) alone, and do not limit searches by filters for test performance (sensitivity, specificity, concordance etc.) as they do not capture change in management studies or direct from test to health outcomes evidence.
- Explode terms when the option is available.
- Use preliminary searches to identify a range of search terms.

All search strategies used should be saved and reported separately for each database searched, including which filters were used (if any). The date that the search was conducted should also be reported, and how many records were retrieved for each database searched.

Search for the impact of a change in management

Approaches that truncate the assessment framework (e.g. claims of non-inferiority that can be established by comparing test characteristics), do not need to provide evidence of treatment effectiveness following the test.

If a full linked evidence approach is required, evidence of the impact of a change in management on health outcomes is unlikely to be identified with a search that applies the proposed test as a search term[x]. A separate search for the impact of the change in management is therefore required. The type of searches will need to be influenced by the change in management identified in the PICO Confirmation clinical management algorithm or identified in the assessment of change in management. Different sets of patients are likely to vary in the changes in management resulting from the test. A separate search may be required for each change in management that would occur following the receipt of the test results.

[w] http://sr-accelerator.com/#/polyglot

[x] As by its definition, if this were available, it would be considered direct from test to health outcomes evidence of clinical utility, hence not requiring the linked approach.

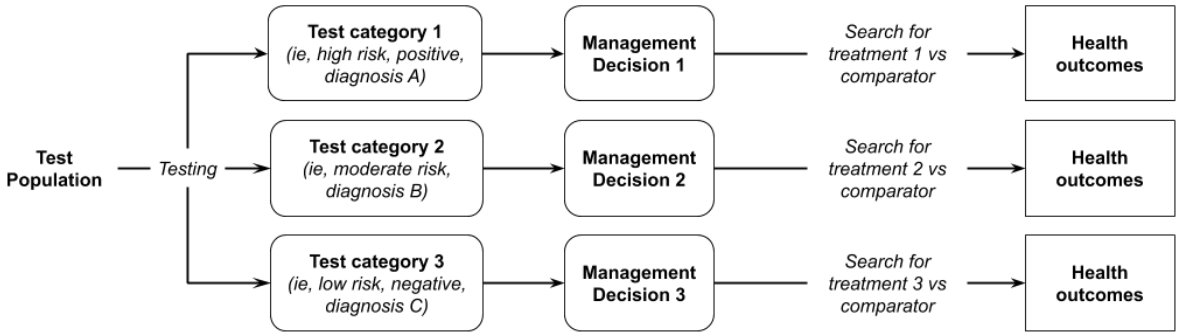


Figure 32 Demonstration of the requirement for multiple searches for the impact of the change in management

Some examples of changes in management and the types of searches required to assess the harms/benefits of these are provided below:

- If the proposed test provides an earlier diagnosis than the comparative test strategy, then the benefits/harms of early versus late treatment would be appropriate. Search terms should include terms relevant for the target population and the treatment, as well as terms that relate to the timing of treatment, e.g. early, late, misclassified, delayed etc.
- If the proposed test reclassifies the stage of disease, influencing the type of interventions chosen, then the effectiveness of different interventions for that stage of disease of disease could be assessed (ideally comparing the treatment that the patient would have received in the absence of the proposed test classification). Search terms should include the target population (possibly with additional terms for stages of disease), and the treatment and/or comparator terms. If scoping searches find high level evidence is available, then searches could be limited to systematic reviews or RCTs.
- If the proposed test results in a patient receiving a different diagnosis than they would have received otherwise, then the effectiveness of treatment for that disease should be compared against the treatment the patient would have received in the absence of that diagnosis (if possible). Limit searches to the highest level of evidence identified in scoping searches.
- If the proposed test results in the avoidance of invasive testing for some patients, then search terms related to the harms of that subsequent test in a relevant population would be required.

In general, it is important that judgement is used to determine and document relevant search strategies used to assess this step. Although best practice would be to perform a systematic review to address the impact of the change in management, judgement may be used to determine whether existing systematic reviews are sufficient, or whether a rapid review of high level evidence (e.g. using one database, with study design filters) may be appropriate. Preference should be given to higher quality and more recent evidence.

Search for personal utility of testing

The methods for assessing the personal utility of testing are still in development. Studies which directly discuss the personal utility of the proposed test in the correct population may possibly be identified through the test related search, as described above. However, it is suggested that bibliographic databases with a psychological focus should be searched to supplement the medical bibliographic databases if the clinical claim relies on personal utility.

In the absence of directly relevant literature, strategies could include:

- Broadening the type of evidence to include qualitative studies, opinion pieces, editorials, consumer input, etc.
- Generalising from broader or other populations/interventions which are likely to have similar consequences.

Presentation of the search strategy

The clear presentation of the search terms improves the transparency of the approach. Tabulating the search terms can assist with presentation. Present a table of the search terms for each bibliographic database or data source, and for each search (if more than one search is performed). The presentation of the search terms must explain how each of the terms interacts with other terms (i.e. Boolean operators).

Table 24 Search terms for the literature review

Category	Description	Search terms
Study design (if justified)	[insert description of category]	[e.g. Cochrane Highly Sensitive Search Strategies for identifying randomised trials in MEDLINE, or MeSH and text word terms for nonrandomised study designs]
Population	[insert description of category]	[include MeSH terms, text words and synonyms for the target population/disease/condition]
Intervention	[insert description of category]	[include known proprietary and nonproprietary names, MeSH terms]
Comparator (if required)	[insert description of category]	[include known proprietary and nonproprietary names, MeSH terms]

MeSH = medical subject headings

Sources of evidence

Search the following sources:

- the published literature in bibliographic databases (at least MEDLINE, EMBASE and Cochrane library)
- registers of randomised trials
- HTA agency websites or the HTA database
- if an applicant developed assessment, any unpublished studies on file
- reference lists of all relevant articles that are obtained.

The selection of data sources to search will be guided by the review topic. Include additional databases that may be relevant (e.g. PsycInfo for mental health literature).

In addition to bibliographic databases, trial registers and internal study reports from manufacturers / sponsors are an important source of identifying studies. Manually searching the reference lists of included studies (also called pearling or backward citation searching) may also identify relevant studies. Furthermore, search for studies that have since cited an included study to identify potentially relevant studies (forward citation searching) (Hinde & Spackman 2015) by looking up the included study and searching the "Cited by" (Google Scholar) or "Times Cited" (Web of Science) list of the study.

As a minimum, search the following sources:

- the published literature
- registers of randomised trials
- if an applicant developed the assessment, any unpublished studies on file
- reference lists of all relevant articles that are obtained (backward citation searching).

The methodological standards for the conduct of new Cochrane Intervention Reviews are an appropriate source of guidance for performing a high-quality systematic literature search.[64]

Table 25 Record of search strategies

Source	Date searched	Date span of search
MEDLINE (via PubMed)	[insert date]	[insert dates]
EMBASE (e.g. Embase.com)	[insert date]	[insert dates]
Cochrane Library[a]	[insert date]	[insert dates]
ClinicalTrials.gov	[insert date]	[insert dates]
International Clinical Trials Registry Platform[b]	[insert date]	[insert dates]
Australian Clinical Trials Registry	[insert date]	[insert dates]
Prospective Register of Systematic Reviews (PROSPERO)	[insert date]	[insert dates]
Internal registries	[insert date]	[insert dates]
Other (state other sources[c])	[insert date]	[insert dates]

a Includes the Cochrane Database of Systematic Reviews, the Cochrane Central Register of Controlled Trials and the Health Technology Assessment database

b International Clinical Trials Registry Platform[y]

c Report on the details of supplementary searches, including manual checking of the references in retrieved papers, searches of the TGA dossier and searches of grey literature.

Study exclusion

Following the literature search, exclude studies that:

A Describe an incorrect intervention (such as when the medical service is used beyond the use described in the requested MBS item descriptor)

B Do not include the target population (not enough patients are enrolled who would be eligible for the proposed medical service according to the requested MBS item descriptor)

C Do not make comparisons with the relevant comparator(s). This step is not relevant if comparative evidence is not identified and noncomparative studies are required.

D Do not report a relevant outcome

After studies have been excluded on the basis of the exclusion criteria (such as the PICO criteria mentioned above, incorrect study type, language, or publication year), consider excluding studies on the basis of quality. This approach may be difficult to justify if the number of relevant includes is not large. Studies excluded on the basis of quality (or other reasons, such as being unable to extract data) should be presented separately from those that are excluded on the basis of not meeting the PICO criteria.

For large reference lists that include a variety of study designs and qualities, identify the studies that represent the highest quality evidence and determine whether they are adequate to answer the

[y] www.who.int/ictrp/en/

assessment question. Where the higher quality studies are inadequate, include lower quality studies to supplement the evidence base. Quality may relate to study design, conduct or size.

Describe and justify the stepwise approach to study exclusion that is taken. Studies that are otherwise eligible for inclusion, but are excluded due to study design, conduct or size, that contradict the results of the included RCTs should be identified and discussed. If randomised studies are included, consider providing results from large comparative observational studies as supplementary evidence.

Published systematic reviews and meta-analyses

When assessing a therapeutic intervention, it is preferable to extract individual studies from published meta-analyses and compare each study against the study selection criteria. Exclude any studies that do not meet the criteria. Discuss the decision to include the treatment effect from a published systematic review.

When assessing a test and searching for the possible (health) impact of change in management, systematic reviews and meta-analyses are often included. During this step of the linked evidence approach, a rapid review is often performed and preference is given to higher quality and more recent evidence. Therefore, in this situation, it may be acceptable to extract the results of the systematic review without extracting data from the individual studies.

PRISMA Flowchart

For an investigative technology, if a linked evidence approach is taken, consider whether a single or multiple PRISMA flowchart(s) are necessary. Typically, at least two separate searches will be required for a linked evidence approach (one to capture test related articles and one to capture the impact of change in management), and it may be more appropriate to present separate PRISMA flowcharts for each search. However, it may be appropriate to use a single PRISMA flowchart for presenting all test related includes. If this is done, present in the flowchart the number of studies included for each assessment question (such as diagnostic accuracy, predictive accuracy, concordance, safety, change in management etc).

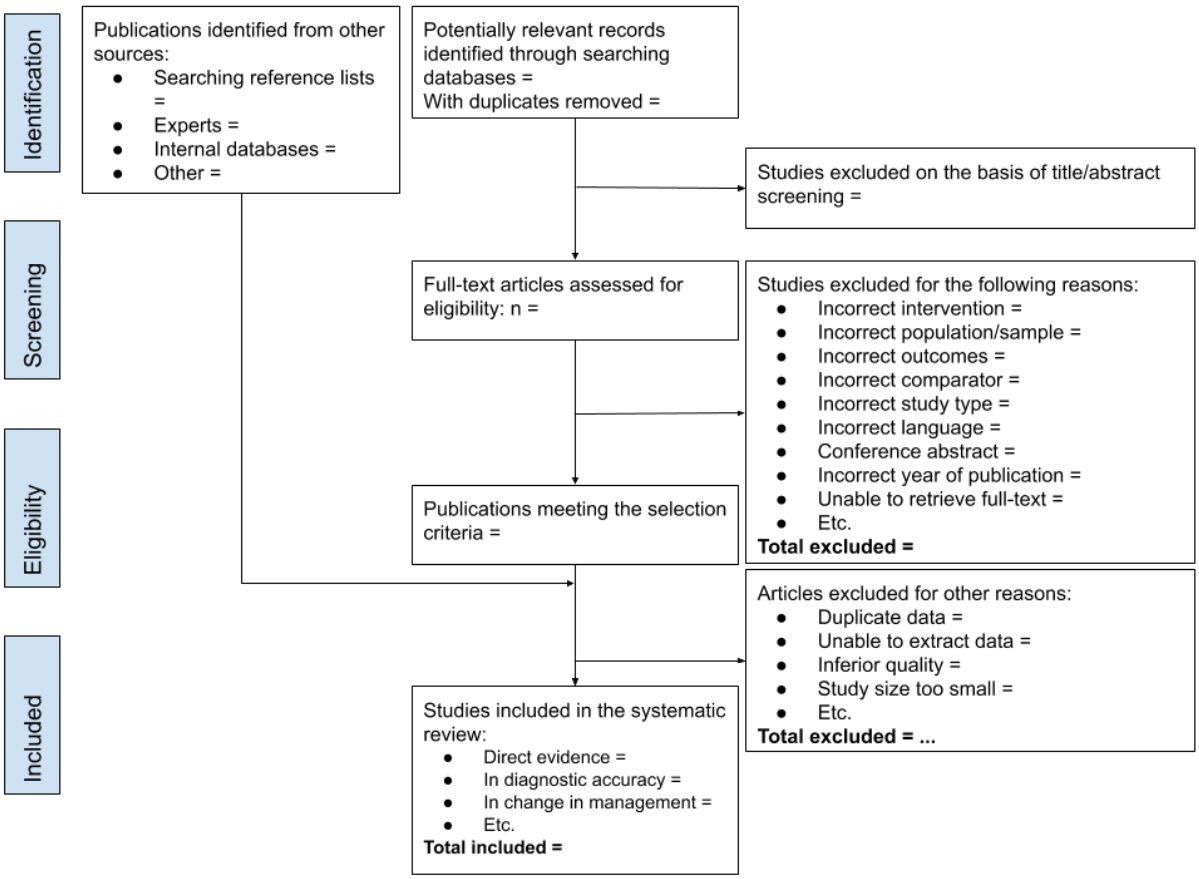


Figure 33 Adapted PRISMA flowchart for presenting screening of studies for MSAC assessment reports, (Liberati et al. 2009; Moher et al. 2009)

Clearly depict the reasons for study exclusion in the PRISMA flowchart.

The adapted PRISMA flowchart has a three-step process for study selection, where studies are excluded:

1. Based on title and abstract, or when the article cannot be retrieved
2. After retrieval of full-text articles
3. Based on clearly specified reasons other than the exclusion criteria described in the "Study Exclusion" section above. Provide justification for each exclusion at this point.

Copies of included studies

For assessment reports that will undergo a commentary route (i.e. be reviewed by an independent review group) prior to consideration by MSAC, to facilitate the critique, provide full text copies of all the included studies. For assessments reports contracted by the Department of Health, provide full text articles from three key studies.

If internal reports (commonly manufacturer led studies) have been included, the full study report is required.

Provide reputable translations of trial reports that are not published in English.

Appendix 3 Risk of bias

The objective of Appendix 3 is to describe appropriate approaches for considering the risk of bias in the studies identified in the assessment report.

Bias is a deviation from the true underlying effect of an intervention as a consequence of issues in study design, study conduct, data collection, data analysis, interpretation of the results, reporting and publication (Tacconelli 2010).

The key purpose of assessing the risk of bias of the included studies is to:

- Provide MSAC with a clear idea of which studies are of greater scientific rigour.
- Assist in the discussion and interpretation of the results.

All studies that are included in the assessment report for the purpose of answering the PICO assessment questions should be assessed for risk of bias.

While the assessment of bias may culminate in a summary statement for each study (i.e. low or high risk of bias), this is not the sole or most important output from the assessment of risk of bias. The assessment of the risk of bias assists in the identification of key issues that may have affected the treatment effect observed in the studies. These issues are then raised during the interpretation of the synthesis of the evidence.

The choice of risk of bias tool should be appropriate for the study design, should be published, structured and (ideally) validated. A list of risk of bias tools developed for different study designs is included in Table 26. These are intended as examples of tools that are commonly used, and not to state what should be used in the assessment. Many risk of bias tools do not differentiate between the possible impact of bias on different outcomes. For example, subjective outcomes may be more susceptible to unmasking or open label designs than objective outcomes. When considering the risk of bias in a study, it is important to consider it in the context of the impact the bias might have on the outcomes of interest. This may lead to different judgements of risk of bias across different outcomes in the same study.

Risk of bias may also be assessed for qualitative studies and ethical analyses, if these studies are likely to be important for decision making.

Table 26 Common risk of bias tools used for different study designs.

Study type	Applicable risk of bias assessment tools	Link / reference
Systematic reviews	ROBIS (2016)	www.bristol.ac.uk/population-health-sciences/projects/robis/
	AMSTAR-2 (2017)	amstar.ca
	NHLBI systematic review checklist	https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
Randomised controlled trials	Cochrane Risk of Bias 2.0 Tool (2019)	https://www.riskofbias.info/welcome/rob-2-0-tool/current-version-of-rob-2
	SIGN checklist for RCTs (2014)	https://www.sign.ac.uk/checklists-and-notes.html
	NHLBI controlled intervention checklist	https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
Non-randomised studies	ROBINS-I (non-randomised studies of interventions) (2016)	www.riskofbias.info
	Newcastle-Ottawa Scale (NOS) (1999)	http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp

Study type	Applicable risk of bias assessment tools	Link / reference
	Downs and Black checklist (1998)	Downs, S. H. and N. Black (1998). "The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions." Journal of Epidemiology and Community Health 52(6): 377-384.
	Cohort studies	
	SIGN checklist for cohort studies (2014)	https://www.sign.ac.uk/checklists-and-notes.html
	NHLBI cohort and cross-sectional checklist	https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
	Case-control studies	
	SIGN checklist for case-control studies (2014)	https://www.sign.ac.uk/checklists-and-notes.html
	NHLBI case-control studies checklist	https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
Test performance / diagnostic accuracy	QUADAS-2 (2011)	www.bristol.ac.uk/population-health-sciences/projects/quadas/quadas-2
Prognostic studies	QUIPS tool (2013)	https://bmjopen.bmj.com › content › embed › inline-supplementary-material-3
Prognostic and predictive prediction models	CHARMS checklist (2014) – prediction models	Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, et al. (2014) Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist. PLoS Med 11(10): e1001744. doi:10.1371/journal.pmed.1001744
Case series	IHE tool (2016, by Moga et al.)	IHE (2016). Institute of Health Economics: Quality Appraisal of Case Series Studies Checklist. Edmonton (AB). Moga, C., et al. (2012). Development of a quality appraisal tool for case series studies using a modified Delphi technique. Alberta, Canada, Institute of Health Economics.
	NHLBI Case series checklist (2014)	NHLBI (March 2014). "Quality Assessment Tool for Case Series Studies." Study Quality Assessment Tools. from https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools. https://www.nhlbi.nih.gov/health-topics/study-quality-assessment-tools
	NHS-CRD Case series quality assessment scale (2001)	Khan, K. S., et al. (2001). Undertaking systematic reviews of research on effectiveness. CRD's guidance for those carrying out or commissioning reviews. York, NHS Centre for Reviews and Dissemination, University of York.
Qualitative studies?	CASP checklist (2018)	https://casp-uk.net/casp-tools-checklists/
	JBI Checklist for Qualitative Research (2017)	http://joannabriggs.org/research/critical-appraisal-tools.html
Ethical analysis	Q-SEA (2017)	Scott, A. M., et al. (2017). "Q-SEA - a tool for quality assessment of ethics analyses conducted as part of health technology assessments." GMS Health Technology Assessment 13(Doc02): 1-9.

AMSTAR = A MeaSurement Tool to Assess systematic Reviews, ROBIS = Risk Of Bias In Systematic reviews, SIGN = Scottish Intercollegiate Guidelines Network, ROBINS-I = Risk Of Bias In Non-Randomised Studies – of Interventions, NHLBI = National Heart, Lung and Blood Institute, QUADAS = Quality Assessment of Diagnostic Accuary Studies, QUIPS = Quality in Prognostic Studies, IHE = Insitute of Health Economics, NHS-CRD = National Health Service – Centre for Reviews and Dissemination, CASP = Critical Appraisal Skills Program, JBI = Joanna Briggs Institute, Q-SEA = Quality Standards for Ethics Analyses.

Regardless of the study design, the following aspects are important to consider when assessing risk of bias:

- Possible bias in selecting the study population. The population included in the study should be appropriate and consist of a representative spectrum of participants. If the selection of participants is inappropriate, this could lead to (among other things) spectrum bias or selection bias.
- The outcomes should be measured and reported in a valid way. If a measure does not produce consistent findings, you cannot rely on the results. Some outcome measures are consistent, yet do not provide accurate/valid results. When outcomes are measured subjectively, and/or by surveys or observations, these results can be susceptible to recall bias, response bias, detection bias, verification bias, clinical review bias, diagnostic review bias and/or test review bias. Furthermore, selective reporting of outcomes/results will also introduce bias to the body of evidence.
- The applicability of intervention and study setting. The intervention and the comparator used in the study should be representative to the target intervention and the way this intervention is proposed to be used in Australia. If the intervention is used in a different setting (e.g. primary care instead of secondary care) the generalisability of the results of the study is questionable.

The best approach to assessing the risk of bias of the studies will depend on the design of the study. Justify the approach (or modifications to the approaches below) taken to capture the key limitations of the study design.

Systematic reviews and meta-analyses

The approach for assessing risk of bias of systematic reviews depends on how the included systematic review is used in the assessment report:

- If the systematic review is included in its entirety, and the assessment report relies on a pooled result from the published systematic review, assess the quality of the systematic review using a validated tool for systematic reviews. Report the methods used by the authors of the systematic review to assess risk of bias for the included studies.
- If individual studies in the systematic review are retrieved and used, or the systematic review is "broken up" such that some studies are excluded, it is preferable to assess the risk of bias for individual studies using a tool relevant to the study design. Using the risk of bias tables provided with a published systematic review is acceptable. Report when this approach is taken.

Only include systematic reviews (rather than individual studies within systematic reviews) if the review is of adequate quality and applicability to the assessment question. Justify the judgements regarding the quality and applicability.

Important aspects specifically for assessing risk of bias in systematic reviews are described in a publication by (Shea et al. 2017) and include whether the systematic review authors (1) have *a priori* agreed on review methods in a protocol, (2) have used a satisfactory technique for assessing the risk of bias of individual studies in the systematic review, (3) have reported any funding sources and potential conflicts of interest, (4) have investigated possible causes of heterogeneity, and (5) have carried out an adequate investigation of publication bias and discussed its impact.

Randomised controlled trials

The assessment of the risk of bias of RCTs is based on factual information about the design and conduct of the study – such as if and how the participants were allocated to groups, or whether or not participants or assessors were blinded. The five domains included in the Cochrane risk of bias tool are (1) bias arising from the randomisation process, (2) bias due to deviations from intended interventions, (3) bias due to missing outcome data, (4) bias in measurement of the outcome, and (5) bias in selection of the reported result.

It is important to consider the flow of participants through the included RCT. Consider the impact on the observed (treatment) effect of patients who are lost or discontinued at any point in the study. Tabulating the points at which patients discontinued or were lost to follow up may assist in identifying potential biases associated with attrition.

As a minimum, data extraction should include the analysed patients as a proportion of the patients enrolled into the study *by study arm.* Differential losses to follow up should be noted and incorporated in the interpretation of the synthesis of the data.

If a randomised trial is available for assessing an investigative technology, the effectiveness of a test usually depends on the patient and/or clinician knowing the result of the test. In many cases, blinding of allocation to different test arms is not possible. This may be acceptable, and in some circumstances preferable, as subsequent management decisions will be made in clinical practice with knowledge of the test that has been used to derive the results. The study should therefore not be rated down for risk of bias due to lack of blinding.

Nonrandomised studies

Nonrandomised studies have a higher risk of bias than randomised studies. Non-randomised studies would usually be included in several steps of the linked evidence approach of an investigative assessment (e.g. to determine the change in management due to an intervention). Methods for mitigating the risks associated with the differential distribution of known confounders because of non-random assignment (such as matching and controlling for confounders in the analysis) cannot adjust for the differential distribution of unknown confounders. If high quality randomised studies are available and form the basis of the assessment report, it may not be necessary to consider the risk of bias of the identified nonrandomised studies.

The internal validity of a non-randomised study can be elicited by reference to how the study design or conduct differs from that of a well-designed, double-blind randomised controlled trial. Bias in a non-randomised study is defined as the systematic difference between the results of the nonrandomised study and the results expected from the ideal double-blind randomised controlled trial. Potential sources of bias that are considered in the ROBINS-I checklist include (Sterne et al. 2016): (1) bias due to confounding, (2) bias in selecting the study population, (3) bias in the classification of interventions, (4) bias due to deviations from intended interventions, (5) bias due to missing data, (6) bias in measurement of outcomes, and (7) bias in selection of the reported results.

Studies on test accuracy / diagnostic accuracy

Some quality assessment tools have been specifically designed to assess the risk of bias in test accuracy studies, studies that are included in the diagnostic accuracy step of the linked evidence approach in an investigative assessment (e.g. QUADAS-2, see Table 26). Important aspects for assessing risk of bias of accuracy studies as per QUADAS-2 are (Whiting et al. 2011): (1) participant characteristics and recruitment (including spectrum bias), (2) applicability of the index test, (3) validity of the reference standard (e.g. misclassification of the target condition), (4) blinding (includes test

review bias, diagnostic review bias and clinical review bias), and (5) patient flow (includes verification and attrition bias).

Studies on prognosis

At least one risk of bias tool is available specifically for prognostic studies (the QUIPS tool). For the assessment of risk of bias of prognostic studies, six domains are included in the QUIPS tool (Hayden et al. 2013): (1) study participation, (2) study attrition, (3) prognostic factor measurement, (4) outcome measurement, (5) study confounding, and (6) statistical analysis and reporting.

Studies without a control group / case series

Case series are uncontrolled studies, and are therefore considered one of the weaker study designs from which to obtain evidence on the effectiveness of an intervention. However, case series have been increasingly included in HTAs due to absence of higher quality evidence, especially in investigative assessments. Studies that do not have a separate control group may provide information (e.g. about how patients benefit from testing or treatment, or about intervention safety), through the use of before-and-after data.

When assessing therapeutic interventions, if patients are selected for inclusion in the study based on the severity of symptoms, there is the risk of regression to the mean. The second measurement will be closer to the population mean than the first measurement, and could be misinterpreted as being attributable to the intervention (Morton & Torgerson 2005). Outcomes which have a high degree of random variability (such as blood pressure) are most susceptible to regression to the mean phenomenon. The problem is often made worse when there are substantial ceiling and floor effects, which is the case in many common quality-of-life scoring instruments (Morton & Torgerson 2005).

Case series may not perform a before and after comparison. Evidence of a comparative effect that is derived through a naïve comparison with another intervention is very low quality. Much of the uncertainty of this approach relates to the potential confounding in the naïve comparison. However, there are some key considerations regarding the methodological quality of a case series that may influence the confidence of the findings: (1) patient selection, (2) (in)adequate ascertainment of exposure/outcome, (3) causality, and (4) reporting[z].

While the "internal validity" of a case series may be reasonable, an estimate of the incremental treatment effect of a therapeutic intervention in a case series is only possible using a naïve comparison with a study of the main comparator(s), or of the natural history of the disease. Comparisons of this type are highly susceptible to confounding.

In randomised studies, confounders are usually balanced across arms. However, known and unknown confounders of intervention performance are likely to be imbalanced across separate case series. Potential confounders are discussed in Appendix 2.

A clear discussion of the potential confounding associated with a naïve comparison should be provided during the interpretation of any results that are derived from a naïve comparison of case series, or a comparison of a case series with the natural history of the disease.

[z] NHLBI (March 2014). "Quality Assessment Tool for Case Series Studies." Study Quality Assessment Tools. from https://www.nhlbi.nih.gov/health-pro/guidelines/in-develop/cardiovascular-risk-reduction/tools.

Other study designs: qualitative studies, prognostic and predictive prediction models, and ethical analyses)

For quality assessment tools specifically designed to assess risk of bias for specific study designs, see Table 26.

Tools are available to assess the risk of bias for other study designs which do not provide a quantitative estimate of the clinical utility of the health technology, e.g. qualitative studies and ethical analyses. Assessment groups should determine whether a bias assessment for these studies will assist MSAC in their decision making (based on perceived importance of this evidence).

Appendix 4 Certainty of the evidence (GRADE)

Each assessment should consider the overall quality of the evidence base (in addition to a separate assessment for each included study discussed in Appendix 3). This enables the conclusions for each efficacy and safety outcome to be weighed in terms of the strength of evidence across all the studies that reported the outcome in question. This feeds directly into the evidence synthesis, such that it is clear that the given outcome having a risk ratio of 'x' was based on k number of included studies, with N number of patients and characterised by (for example) reasonable consistency between trials or other key features. An overall measure of confidence in the result (high, moderate, low and so on indicated by the traffic light value) represents the quality of the evidence for that outcome, and the certainty that MSAC may have that the evidence represents the "true" effect of the health technology.

Overview of the GRADE approach for therapeutic technologies

Authors should use a Grading of Recommendations Assessment, Development, and Evaluation (GRADE) approach to present an overall assessment of the quality of the evidence for each critical outcome. GRADE requires an assessment of the following domains to rate the *quality* or *certainty* of the body of evidence.

- Study design (Balshem et al. 2011)
- Risk of bias or study limitations (Guyatt, Oxman, Vist, et al. 2011)
- Imprecision (Guyatt, Oxman, Kunz, Brozek, et al. 2011)
- Inconsistency of results (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Glasziou, et al. 2011)
- Indirectness of evidence (applicability of the population, intervention, comparator and outcomes) (Guyatt, Oxman, Kunz, Woodcock, Brozek, Helfand, Alonso-Coello, Falck-Ytter, et al. 2011)
- Publication bias (Guyatt, Oxman, Montori, et al. 2011)

Assessment authors initially describe the study design (RCTs start with a high rating, observational studies with a low rating), and then rate the evidence down for weakness in any of the above domains. Alternatively, the evidence may be rated up when:

- When a large magnitude of effect exists,
- When there is a dose response gradient,
- When all plausible confounders or other biases increases confidence in the estimated effect. (Guyatt, Oxman, Sultan, et al. 2011)

Following the assessment of individual study results and meta-analysis (if appropriate), an evaluation of the imprecision, inconsistency, indirectness and risk of publication bias across the evidence base (per outcome) should be performed. For each critical efficacy and safety outcome identified in the PICO Confirmation, discuss the overall strength of the evidence base, noting the number of trials (k) that provide direct from test to health outcomes evidence and the corresponding number of participants. An example of a simplified GRADE table is shown in Table 27.

Table 27 GRADE table for critical and important outcomes

Quality assessment for [patient relevant outcome #1]

No. of studies (Design)	Limitations (ROB)	Inconsistency	Indirectness	Imprecision	Publication bias	Certainty
k [design] k2 [design]						High ⨁⨁⨁⨁ Moderate ⨁⨁⨁⊙ Low ⨁⨁⊙⊙ Very low ⨁⊙⊙⊙

k=no. of studies

⨁⨁⨁⨁ Very confident that the true effect lies close to that of the estimate of the effect.

⨁⨁⨁⊙ Moderate confidence in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different

⨁⨁⊙⊙ Confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect.

⨁⊙⊙⊙ Very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

Indirect or surrogate outcomes (not patient-relevant) should only be included in the GRADE tables where direct from test to health outcomes evidence is inadequate and the supporting evidence from a surrogate endpoint (e.g. antibody titre) provides valuable context for interpretation of the more limited patient-relevant outcome data (e.g. pathology-confirmed cases). When a surrogate outcome is included for a therapeutic technology, it should be rated down for indirectness of the outcome measure.

For certain outcomes, evidence from the included trials may be minimal or absent (long-term outcomes such as overall survival; certain low frequency severe adverse events). If such outcomes have been identified as critical, these should be included within the limits of the evidence. It may not be possible to provide an effect estimate and the traffic light value may be very low or nil (for the latter where no estimate or evidence was available). This covers the outcomes that are relevant, not just those with available evidence.

Multiple GRADE tables may be needed for an evaluation that addresses more than one population or different applications of an intervention (e.g. monitoring as well as diagnosis). More detail on how to apply the GRADE approach as well as alternative formats for different types of evidence can be found in in the GRADE series of papers introduced in Journal of Clinical Epidemiology volume 64 (2011) and references therein, and also information in the GRADE Handbook (gradepro.org).

According to the GRADE approach, the directness domain includes applicability.

A basic summary of findings table is given below. Examples from published systematic reviews can be found in the GRADE Handbook. Additional information for investigative technologies is in the next section.

Summary of Findings
Comparison of intervention X with intervention Y in patients with Z condition and ABC clinical features

Outcome (duration of follow-up)	**Number of studies (k), study design; patients (n)**	**Intervention absolute effect (RD) [95% CI]**	**Comparator absolute effect (RD) [95% CI]**	**Relative effect (RR) [95% CI]**	**Certainty**	**Comments**
Effectiveness Outcome 1	…	…	…	…	⨁⨁⨁⨁ ⨁⨁⨁◯ ⨁⨁◯◯ ⨁◯◯◯ ◯◯◯◯	…
Effectiveness Outcome 2						
Safety Outcome 1 etc						

Specific issues for investigative technologies

Direct from test to health outcomes evidence

Where direct from test to health outcomes evidence is available to assess an investigative technology, the standard GRADE approach (developed for therapeutic technologies) is able to be used, with consideration of how additional applicability issues are addressed (see Technical Guidance 10). Randomised trials, or systematic reviews of randomised trials, are the highest level of evidence, and any outcomes measures that are not directly patient relevant are rated down for indirectness.

Any harms associated with testing or downstream consequences should be assessed as per the standard GRADE approach.

Linked evidence of clinical utility

Singh et al 2012 recommend that, since most evidence for efficacy of investigative technologies is indirect, authors should consider the quality for each link in the evidence framework, or justify why this has not been considered (Singh et al. 2012). For a linked evidence approach, adaptations of the GRADE process vary depending on the component of the linkage.

Elements that could affect the quality of the evidence could include variability in characteristics of the test population, differing test result definitions that lead to a change in management, differing options for treatment based on the same test result, and so on.

Change in management and health outcomes

For evidence on change in management and therapeutic efficacy, the standard GRADE approach for therapeutic technologies can be used, with the amendment that the outcomes are not rated down for indirectness, despite not directly informing on the question of the clinical utility of the test. Note: The outcomes may be rated down for indirectness due to other reasons (e.g. indirectness of outcome if a change in diagnosis is reported, rather than a change in management; or outcomes reported are management recommendations, but not management received, or indirectness of population for therapeutic efficacy, if the spectrum of patients treated does not match those who are likely to be treated based on test results).

Cross-sectional accuracy

Using the GRADE approach for questions regarding the diagnostic accuracy of a test are well established, with GRADEpro software having an option for diagnostic questions. Evidence for diagnostics usually includes studies of widely variable design, typically with few or no RCTS that adequately address *both* the tested and the treated populations. The study designs considered 'high' quality are cross-sectional studies (cohort type accuracy studies), whereas case-control accuracy studies are considered 'low' quality.

Assessment should consider both the quality of evidence as it relates to test accuracy, as well as the proportion of patients with unevaluable results, or inconclusive results (who may need to have a second sample retrieved). The GRADE approach has been developed for the purpose of clinical practice guideline developers, and has therefore been targeted towards assisting clinicians to understand the strength of evidence for guiding treatment of individual patients. The emphasis has therefore been on the outcomes of true positives, true negatives, false positives and false negatives. For the purpose of population-based funding decisions, it is suggested that the more relevant outcome measures (which are less susceptible to pre-test probability differences) are sensitivity and specificity.

A summary of findings table adapted for 'diagnostic interventions' is described in the GRADE Handbook and adapted below in Table 28.

Table 28 Summary of findings for diagnostic accuracy

Summary of Findings – test important outcomes
Comparison of test X with test Y in patients suspected to have Z condition

Test Outcome	Number of studies (k); patients (n)	Intervention result	Comparator result	Absolute difference	Quality	Comments
Sensitivity	…	% (range)	% (range)	…	⨁⨁⨁⨁ ⨁⨁⨁⊙ ⨁⨁⊙⊙ ⨁⊙⊙⊙ ⊙⊙⊙⊙	…
Specificity	…					
Unevaluable	…					
Inconclusive results	…					

Longitudinal accuracy

Tests which are performed to determine future patient outcomes, rather than current status are considered to provide 'predictive accuracy' (see TG 11.4). This type of evidence requires adapting the standard GRADE process. Instead of considering RCTs the ideal study design (as per therapeutic technologies) or cross-sectional data (as considered best for diagnostic accuracy studies), predictive accuracy is best assessed using prospective longitudinal data to confirm actual versus predicted patient outcomes. Prospective cohort studies should therefore be given a 'high' rating, whereas other study designs should be given a 'low' rating. If the GRADEpro software is used, select the study design 'randomised trial' if the evidence identified is prospective cohort studies, and 'observational study' for all other study designs. The results columns will need to be adapted to suit the information received.

Qualitative evidence

The standard GRADE approach cannot be applied to qualitative evidence. If evidence synthesis is based on qualitative research (i.e. for demonstrating personal utility), quality assessment may be done using the GRADE CERQual[aa] approach (Lewin et al. 2018; Lewin et al. 2015), and as recommended by the Cochrane Collaboration. Instead of considering the confidence in an overall effect estimate, authors are asked to consider whether the review finding is representative of the phenomenon of interest. This is based on the following elements:

- The methodological limitations of the qualitative studies contributing to a review finding;
- The relevance to the review question of the studies contributing to a review finding;
- The coherence of the review finding; and,
- The adequacy of data supporting a review finding.

A fifth element to address publication (dissemination) bias is in development (Booth et al. 2018). Note that this approach does not address the limits of individual qualitative studies nor the methodology used for the evidence synthesis. Further information is on the website (www.cerqual.org).

[aa] CERQUAL = Confidence in the Evidence from Reviews of Qualitative research

Appendix 5 Study characteristics

This appendix discusses the consideration of the characteristics of the included studies. These characteristics include:

- The populations enrolled in the study;
- The health technologies and circumstances of use;
- The definitions and timing of outcomes measures;
- The statistical approaches used for reporting outcomes.

There are two key purposes for considering the characteristics of the included studies. Firstly, a comparison across arms and across studies will identify potential confounders or sources of heterogeneity, and inform the interpretation of the synthesis of the evidence. Secondly, the characteristics of the included evidence can be compared against the proposed use of the health technology in the Australian setting.

The presentation of characteristics of the included studies will vary depending on the type of evidence and the volume of studies. For assessments that require multiple steps (such as a linked evidence approach for investigative technologies), a separate comparison of characteristics are required for each step.

For evidence presenting a direct comparison of a health technology with an appropriate comparator (such as an RCT), it is preferable to present the characteristics of the studies in a tabulated format to enable comparison across arms and across studies. A comparison of the following characteristics may assist in the identification of possible sources of heterogeneity and/or confounding:

1. Study design or conduct
2. Patient and disease characteristics
3. Eligibility criteria
4. Description of the interventions
5. Outcome definitions

The level of detail that is appropriate and achievable may depend on the number of studies included in the assessment report. For assessment reports based on large numbers of included studies, this approach may not be feasible to undertake for a large number of characteristics. Focus on key participant, treatment and setting characteristics that are informative to decision making. In cases where there are a large number of included studies, the following approach may be appropriate:

1. Identify key characteristics (study, patient baseline or disease characteristics) that may have an impact on the comparative performance of the health technology (see Appendix 2).
2. Report key characteristics for each study in the study data extraction tables
3. Tabulate the key characteristics of the included studies to enable a comparison across studies.
4. Highlight important differences in the key characteristics (where the differences have a likely or possible impact on the treatment effect).
5. Focus on key characteristics that will are informative to decision making.

Note: For assessment reports that have included a systematic review, it may not be necessary to present the study characteristics for individual studies. Present key characteristics of the populations, treatments and outcomes included in the systematic reviews. If individual studies are extracted from the systematic reviews to answer one or more assessment question, the characteristics of the individual studies are required.

Participants

For each of the included studies, consider the following details about the study participants:

- eligibility criteria for participants considered for recruitment into the study
- baseline demographic and clinical characteristics for each group or study arm
- median duration (and range) of follow-up for each group and for the entire study (also indicate whether the study is ongoing).

Identify characteristics that may have an impact on the target outcomes. These may be identified from a background search of the literature, or by reviewing subgroup analyses of the included studies.

Important characteristics should be extracted from included studies and reported in study tables. Do not report extensive characteristics of included studies that are unlikely to affect the interpretation of the evidence.

For studies with high losses to follow up, or differences across arms in terms of the extent or timing of losses to follow up (discontinuations, withdrawals, or other causes of censoring), compare the characteristics of the patients who were censored from the analysis with those who remained in the study. Whether information about this subgroup is or is not presented, consider the impact of censoring or loss to follow up, particularly when it is differential across study arms.

Reporting the results of the comparison of participant and disease characteristics should include:

- A summary of the key characteristics that may impact on the treatment effect (regardless of whether there are differences between studies)
- Key differences in these characteristics across arms within studies
- Key difference in these characteristics across studies
- Key differences in the characteristics presented in the included studies and the target population (particularly if there is an important subgroup that is not represented in the included studies that will access the proposed medical service in the Australian setting)

Where the assessment report is based on a subgroup of an included study, it is important to compare the baseline characteristics for the relevant subgroup as well as the whole study population. Discuss whether the selection of the subgroup has increased the risk of bias associated with the comparison of the health technology and the main comparator.

Health technology details

Differences in the use of the proposed health technology or comparator may result in heterogeneity of the observed results across studies. Consider the following details about the health technologies provided in each study:

- How the health technology and the main comparator were defined and delivered. Important characteristics for interventions may involve dose, frequency / episodicity, duration, need for subsequent treatments, the line of therapy and concomitant treatments. Important characteristics for tests may involve the timing of the test, sample details and thresholds applied.
- Criteria for concomitant or subsequent intervention or confirmatory testing
- Settings in which the health technology and main comparator are used

Outcomes

Outcome definitions can differ across studies and may result in heterogeneity of the observed results. The following aspects relating to outcomes should be considered, and extracted, for each of the included studies:

- the primary outcome (or state if no primary outcome has been nominated)
- secondary outcomes that were identified in the PICO Confirmation

For each outcome:

- identify the definition of the outcome
- state the units of measurement and the method of statistical analysis
- describe the population in which the analysis is performed (ie intention to treat, per protocol)
- describe the timing of the outcome assessment and who performed the assessment
- describe the instrument used to measure the outcome (eg questionnaire, criteria such as RECIST, blood test), and state whether it has been validated
- state how missing data were dealt with (it is important to address both patients who remain in follow up who have not yet experienced an event, as well as those that were removed from the analysis).

Outcome measures may appear similar across studies; however, they may be influenced by covariates used in statistical approaches, and by censoring rules. State whether censoring applied in the study is appropriate or may be informative. When recording the method of statistical analysis, include the name of the statistical test and sufficient details to allow MSAC to ascertain how the analysis was performed.

Composite outcomes

A composite outcome is one in which multiple endpoints are combined. It is usually defined as having been experienced when the first of any of the component endpoints is experienced, even though subsequent component endpoints may occur.

For assessment reports that include composite outcomes, additional details relating to the definition of each composite outcome should be considered and reported. The assessment report should consider:

- The individual definitions of the components in the composite outcome
- The clinical importance of each of the components
- Whether the composite outcome was explicitly prespecified
- Whether the composite outcome can be disaggregated, or whether disaggregation is not possible due to censoring that occurs following the first event in a composite outcome.

The interpretation of a composite outcome should consider which of the components is driving the composite outcome, and whether this is similar across arms.

Patient-reported outcome measures

Patient-reported outcome measures include generic ('global') or condition-specific (eg for respiratory conditions, depression, arthritis) measures of quality of life, symptoms or function.

Patient-reported outcome measures may also include multiattribute utility instruments (MAUIs), in which the scoring method for the instrument is anchored on a quality-adjusted life year scale of 0 (death) to 1 (full health). Several commonly used MAUIs for which a detailed discussion of the

validity or reliability is not required are the Health Utilities Index (HUI2 or HUI3), the EQ5D-3L or -5L ('EuroQol'), the SF-6D (a subset of the Short Form 36, or SF-36), the Assessment of Quality of Life (AQoL) instruments, and the Child Health Utility 9D (CHU9D) index for children and adolescents.

An assessment report should describe the patient-reported outcome measurement, and state whether it is validated for use in the population, condition and interventions. Describe the timing of and the personnel who administered the assessment.

Missing data is an important consideration for all outcomes; however, it is common in patient reported outcome measures. The assessment report should consider compliance with the patient reported outcome measures, and whether compliance (particularly when differential across study arms) may have affected the comparison across arms. Describe any methods used to adjust for response bias (or methods for adjusting for missing data).

Minimal clinically important difference

The definition of a minimal clinically important difference (MCID) is varied across the literature. The central concept of an MCID is that it represents the smallest amount of difference in a score that would, in some way, be considered important. MCIDs may be reported in either relative or absolute measures.

When selecting an MCID, it is important that the source of the MCID is relevant for the population, disease and interventions included in the assessment report.

Likely sources for an MCID may be:

- study protocols (often for the purposes of powering the study)
- a previously accepted MCID by MSAC that is relevant to the study population and the proposed indication
- a commonly accepted MCID in the literature, relevant to the study population and the proposed indication
- a commonly accepted MCID in the literature for a similar indication that can reasonably be expected to be generalisable to the proposed indication.

The derivation of an MCID for a dichotomous outcome (eg haemorrhage or no haemorrhage) or time to event outcome (eg overall survival) is not straightforward and may not be available. The most common approach for determining a meaningful benefit to patients involves a consensus of clinical experts in the relevant fields.

The application of an MCID to a surrogate outcome should accompany a rigorous explanation. The MCID for the surrogate should reflect a minimal important difference in the target patient relevant outcome. The application of an MCID to a test accuracy outcome (eg, sensitivity or specificity) is not appropriate. The translation of a test result to change in management and the eventual impact on health outcomes (incorporating both test negative and test positive patients) is complex and not commonly quantified outside of decision analytics.

The interpretation of the results in the context of a nominated MCID can be difficult if only aggregated data are provided. Typically, an MCID reflects the average minimal difference in a score that is considered to be clinically important. Study results are most commonly aggregated to reflect the average estimate of change in a score for each study arm. If the average change experienced by a study arm is lower than the MCID does not mean that, for some patients, a clinically important change has not occurred. Equally, if one arm reports an average change above the MCID and the other arm reports an average change below the MCID, it may not be possible to infer that the difference between the arms is clinically meaningful.

A more meaningful method of examining response is to report the proportion of participants who experienced a change in a score that was greater than the MCID. A responder analysis can also be represented by a cumulative distribution function such that the proportion of responders can be viewed across multiple thresholds for an MCID.

Non-inferiority margin

A claim of non-inferiority means that, in terms of safety and effectiveness, the proposed therapeutic technology is no worse than the main comparator. However, a lack of a statistically significant difference between the proposed intervention and the comparator does not adequately establish non-inferiority. It is common practice to require that the confidence limits of the difference in treatment effect do not include an *a priori* stated clinically meaningful difference favouring the comparator.

If the proposed intervention is claimed to be non-inferior to the main comparator, state whether an acceptable non-inferiority margin has been identified in the PICO Confirmation, and for what outcomes. If a non-inferiority margin is not available, describe any non-inferiority margins identified in the literature, and state how they were derived.

The application of a non-inferiority margin that was not prespecified in a study is difficult to justify. If a non-inferiority margin is required to establish whether the proposed intervention is non-inferior to a comparator, and studies have not prespecified a non-inferiority margin, the selection of a conservative margin (eg narrow margin) is more appropriate.

A non-inferiority margin is not necessary for all outcomes, and is typically only applied to the primary outcomes of non-inferiority studies. Studies may be underpowered to support the use of non-inferiority margins for less common outcomes or outcomes that were not used to power the study.

Appendix 6 Sources of Heterogeneity

This appendix describes possible sources of heterogeneity between studies, or when comparing one jurisdiction with another. It is a useful reference for describing potential confounders when combining studies in a meta-analysis, performing indirect comparisons of randomised trials or network meta-analyses, or comparing variables from the clinical study setting with the target population.

Make comparisons across studies or jurisdictions based on the distributions or proportions of each characteristic rather than simply identifying whether there is a representation of each characteristic in each study or jurisdiction. For example, two trials may include patients aged 20–60 years, thus, the population may appear homogeneous. However, if one trial has a much lower mean age, or the proportion of patients younger than 40 is far higher than for the other trial, this may be a source of heterogeneity and violate the assumption of transitivity.

Table 29 provides a list of important factors to consider when exploring heterogeneity in the evidence or the applicability of the evidence to the target population.

Table 29 Example factors that might cause heterogeneity across studies or jurisdictions

Category	Factor
Study Quality	Adequate concealment of randomisation
	Blinding
	Duration of follow-up
	Loss to follow-up, methods for censoring or imputation of missing data
	Crossover or treatment switching
Participant characteristics	Age, sex, performance status, comorbidities, physiological reserve
	Severity of disease, stage or duration of disease, previous therapy, genetic variation
	Intensity of surveillance, Diagnostic workup
	Background therapy, advances in standard of care
	Values, expectations and adherence
Circumstances of use	Health systems, setting in hospital or ambulatory care
	Geography, urban or rural
	Date of studies (change in standard of care)
Management decisions	Regional / country variations in practice
	Different treatments available, accessible, reimbursed
Treatment characteristics	Dose, duration, timing
	Stopping or continuation criteria
Test characteristics	Assay platform, enzymes, reagents, protocols, primers
	Test thresholds
	Resolution of imaging
	Sampling method and handling
	Interpretation of results, interrater variation
Outcome measures	Definition of outcome(s)
	Rating instrument
	Frequency of measurement
	Start point of measurement against duration or progression of disease or treatment, especially in time-to-event analyses
	Statistical approach and covariates

Appendix 7 Test accuracy measures

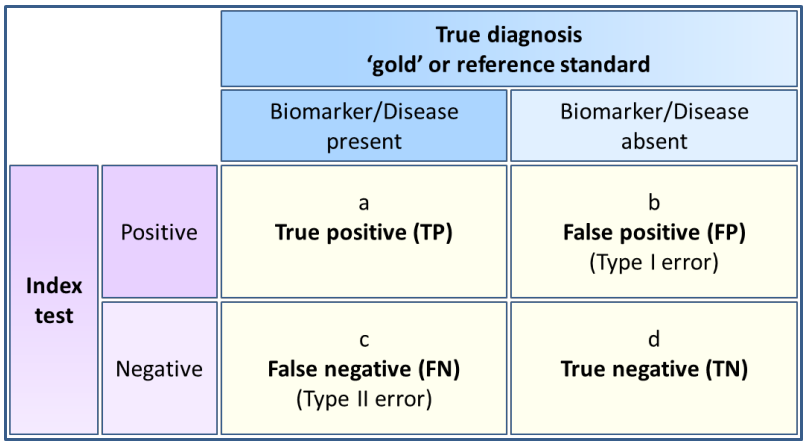


Figure 34 The 2-by-2 table

Table 30 Formulae for calculating test accuracy measures

Test accuracy measures	Calculations
Sensitivity	TP / disease positive = a / (a + c)
Specificity	TN / disease negative = d / (b + d)
Positive likelihood Ratio (LR+)	TP rate / FP rate or sensitivity / (1 – specificity) = [a / (a + c)] / [b / (b + d)]
Negative Likelihood Ratio (LR–)	FN rate / TN rate or (1 – sensitivity) / specificity = [c / (a + c)] / [d / (b + d)]
Diagnostic odds ratio (DOR)	(TP / FP) / (FN / TN) = LR+ / LR– = (a / b) / (c / d)
Positive Predictive Value (PPV)* Study-specific For disease or biomarker prevalence rate in PICO population	TP / test positive = a / (a + b) $\frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}$
Negative Predictive Value (NPV)* Study-specific For disease or biomarker prevalence rate in PICO population	TN / test negative = d / (c + d) $\frac{\text{specificity} \times (1 - \text{prevalence})}{(1 - \text{sensitivity}) \times \text{prevalence} + \text{specificity} \times (1 - \text{prevalence})}$
Number Needed to Diagnose (NND)	1 / Youden's index = 1 / (sensitivity +-specificity –1) 1 / ([a / (a + c)] + [d / (b + d)] – 1)
Number Needed to Misdiagnose (NNM)	1 / (1 – accuracy) = 1 / 1 – [(a + d) / (a + b + c + d)] 1 / 1 – specificity – [prevalence × (sensitivity – specificity)]
Concordance measures	**Calculations**
Positive percent agreement	100% × [a / (a + c)]
Negative percent agreement	100% × [d / (b + d)]
Overall percent agreement	100% × [(a + d) / (a + b + c + d)]

*PPV and NPV can be calculated from the 2-by-2 table for individual studies. However, this value is only valid for the prevalence of the disease or the biomarker in that study. For the PPV and NPV values to be applicable to the Australian population defined in the PICO, these values should be calculated using the pooled sensitivity and specificity of the test and the applicable prevalence rate for the PICO population.

Meta-analytical methods

A key difference between pooling interventional/therapeutic study data and test accuracy data is that test metrics tend to be correlated. The interpretation of calculated test metrics for each study can be assisted by presenting confidence interval plots (forest plots) of the sensitivity and specificity side-by-side. Ordering the results by ascending sensitivity or specificity can help with visualising the relationship between sensitivity and specificity (Figure 35).

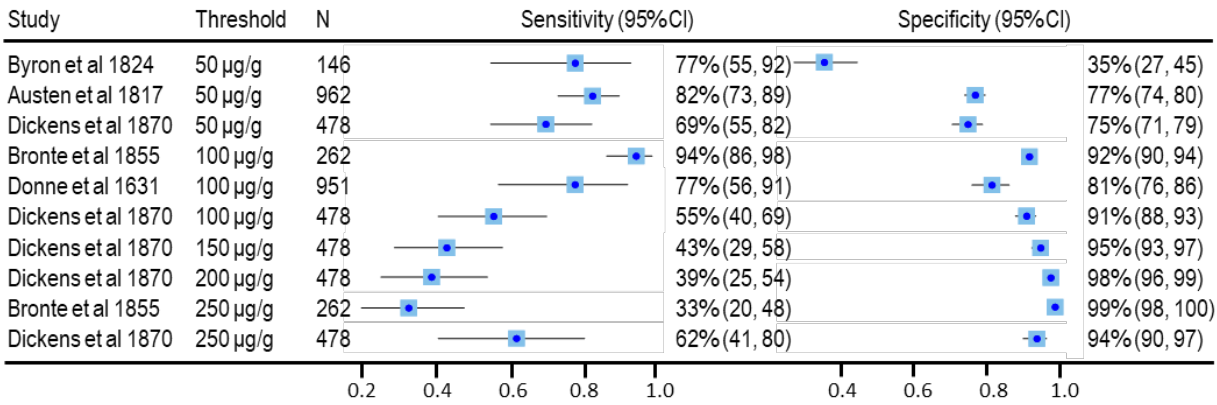


Figure 35 Forest plots showing the relationship between sensitivity and specificity of a test compared with the reference standard for different thresholds

The forest plot shows the sensitivity and specificity for a test compared with the reference standard for each threshold reported for each study. Overall, it looks like there is a trend that as the threshold increases, the sensitivity decreases and the specificity increases.
Note: no overall pooled sensitivity or specificity values were calculated, as this was not appropriate. The forest plot includes duplicated data: values: the population was the same for each threshold reported in the same study.
CI = confidence interval; N = number of patients

Typically, there is an inverse correlation between sensitivity and specificity (Cleophas & Zwinderman 2009). This may be due to different thresholds used to determine a positive sample, as in Figure 35. If the threshold to determine a positive test is decreased, this will permit more test positives, but will also increase the number of false positives. In this circumstance, sensitivity would increase and specificity would decrease. This inverse relationship between sensitivity and specificity, when related to differences in thresholds, is called a threshold effect.

For this reason, pooling of sensitivity and specificity using a bivariate meta-analysis method is preferred.

Bivariate meta-analysis

The bivariate models assume that the logit of sensitivity and specificity have a bivariate normal distribution between studies (Reitsma et al. 2005). There are other options for performing a random effects bivariate meta-analysis that assume beta-binomial distributions (Hoyer & Kuss 2015), and extensions of the bivariate model that may permit the inclusion of thresholds as a covariate (Hoyer & Kuss 2018). These other methods are more complex and should be clearly described when applied.

A minimum of four studies are required for bivariate meta-analysis to obtain summary point estimates of sensitivity and specificity. The decision to estimate a summary point should be based on the characteristics of the included study data. If all studies report the same or very similar test thresholds and population characteristics it is likely reasonable to use the bivariate model approach for overall estimates. Where the population subgroups and/or test thresholds differ meaningfully, or the sensitivity and specificity vary over a large range, an overall pooled estimate is difficult to interpret and is unlikely to be appropriate under these circumstances (Trikalinos et al. 2012). Applying the

bivariate model to each of the distinct subgroups is likely to be more appropriate, as shown in Figure 36.

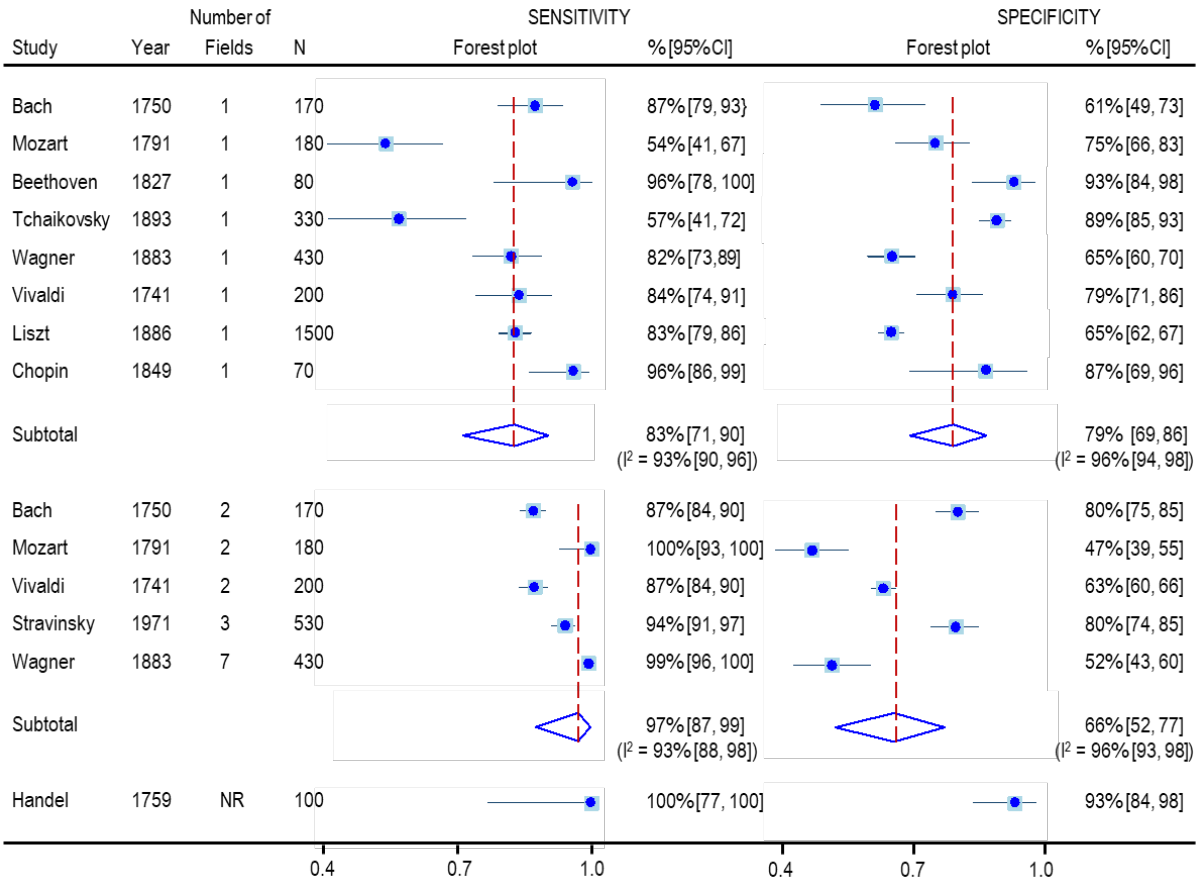


Figure 36 Forest plot of sensitivity and specificity for an imaging test compared with a reference standard according to the number of fields included

Subgroup analysis showed that the sensitivity increased and the specificity decreased when more than one field was included in the imaging test, although the 95%CIs were overlapping suggesting the difference may not be statistically significant.
CI = confidence interval; N = number of patients; I^2 = statistic describing the percentage of variation across studies that is due to heterogeneity rather than chance

Common statistical software packages that can be used to perform meta-analysis of test accuracy studies using bivariate or hierarchical models include:

- STATA using the midas or metandi commands
- SAS using the metaDAS macro
- R using the mada package.

Univariate meta-analysis

As a general rule, a bivariate model approach or a hierarchical model approach is preferred, unless the model does not converge, in which case separate univariate binomial meta-analyses can be used with justification and a discussion of the uncertainties in the approach. Non-convergence of models may occur when there are few studies or sparse data, particularly if there are several zero cells in the 2-by-2 table (Takwoingi et al. 2017).

Random effects univariate meta-analysis will produce "average" estimates of sensitivity and specificity. This approach does not account for heterogeneity in sensitivity and specificity related to the threshold effect, and statistical measures of heterogeneity may be difficult to interpret. In the

presence of a threshold effect, the individual pooled estimates of sensitivity and specificity may be incompatible (Reitsma et al. 2005).

In the absence of visual heterogeneity across studies, separate random-effects univariate meta-analyses of sensitivity and specificity will approximate the use of more complex model fitting methods. For univariate models, the most appropriate method for the pooling of sensitivity and specificity is to perform a binomial meta-analysis (Nyaga, Arbyn & Aerts 2014).

The diagnostic odds ratio (DOR), defined as the ratio of the odds of positivity in those with the biomarker or condition relative to the odds of positivity of those without the biomarker or condition (Glas et al. 2003), is a single parameter of test accuracy. Hence, DOR can be pooled using univariate models. However, DOR summary measures do not distinguish between the ability to detect true positive cases (sensitivity) and the ability to detect true negative cases (specificity) and, are therefore, more difficult to interpret in a clinically relevant way (Lee et al. 2015). In other words, the same DOR may be achieved with different sensitivity and specificity values. The use of DORs can overcome the issues with the negative correlation between sensitivity and specificity (Cleophas & Zwinderman 2009), and may help to ascertain the best "performing" test. However, if a clinical situation requires greater test sensitivity and the trade off in specificity is permissible (but not the other way around), then the highest DOR might not be the best "performing" test. DORs can be applied in meta-regression to explore heterogeneity.

Multiple thresholds from single studies

Test accuracy data for any one patient should only be included in a meta-analysis once. If two studies have the same or overlapping patient cohorts, only one can be included in the meta-analysis. Similarly, if individual studies report test accuracy data for the same patients or samples for different thresholds, only one threshold can be included in a single meta-analysis if summary estimates are to be reported. If a meta-analysis of different thresholds were undertaken, it would be appropriate to include the same study in the separate meta-analyses for the different thresholds, as shown in Figure 36.

Hierarchical summary receiver operating characteristic curve

For some tests, there is no universally agreed threshold for determining a positive result and some studies may use several different thresholds. If there are a mixture of thresholds used across and/or within studies, and there is no clear reason to limit the analysis to a single threshold, it may be appropriate to present a hierarchical summary receiver operating characteristic (HSROC) curve. HSROC curves can be generated using either HSROC models that directly estimate HSROC parameters or bivariate models by transforming the estimated parameters of the bivariate model so that a HSROC curve can be fitted. The two models are mathematically equivalent and provide equivalent estimates of expected sensitivity and specificity (Lee et al. 2015).

HSROC curves characterise the relationship between sensitivity and specificity across the included thresholds and accounts for within- and between-study heterogeneity. The HSROC curve plots the true positive rate (or sensitivity) against the false positive rate (or 1 – specificity), and this graphical representation of the included studies provides an easy way to examine both the threshold effect and between-study heterogeneity. The 95% confidence region is a measure of the precision of the test accuracy estimate and the 95% prediction region is a measure of between-study variability or heterogeneity and defines the area in the HSROC space where a future study would lie (Bossuyt & Leeflang 2008). However, as test accuracy studies tend to be highly variable the 95% prediction regions often cover large areas of the HSROC space. A 50% prediction region is equivalent to the interquartile range.

Figure 37 shows HSROC curves generated in STATA using the metandi command (HSROC model) and the midas command (bivariate model) using the same studies, with respecified thresholds, shown in

the forest plot in Figure 35. Although the point estimates derived for each HSROC curve are the same, there are some differences between the two HSROC curves with respect to the 95% confidence and predictive regions. Both curves enable visualisation of the threshold effect, where the studies with thresholds of 50 µg/g were more sensitive but less specific than those with thresholds at or above 100 µg/g, which was more difficult to discern in the forest plot. The large 95% prediction region suggests that there is some heterogeneity between the studies in Figure 37. In contrast, in Figure 38 the 95% prediction region fits much tighter to the HSROC curve suggesting there is less heterogeneity between the included studies.

When HSROC curves include multiple thresholds from the same study, reporting of summary measures such as the summary point sensitivity and specificity values as well as the area under the ROC curve (AUROC) are not appropriate.

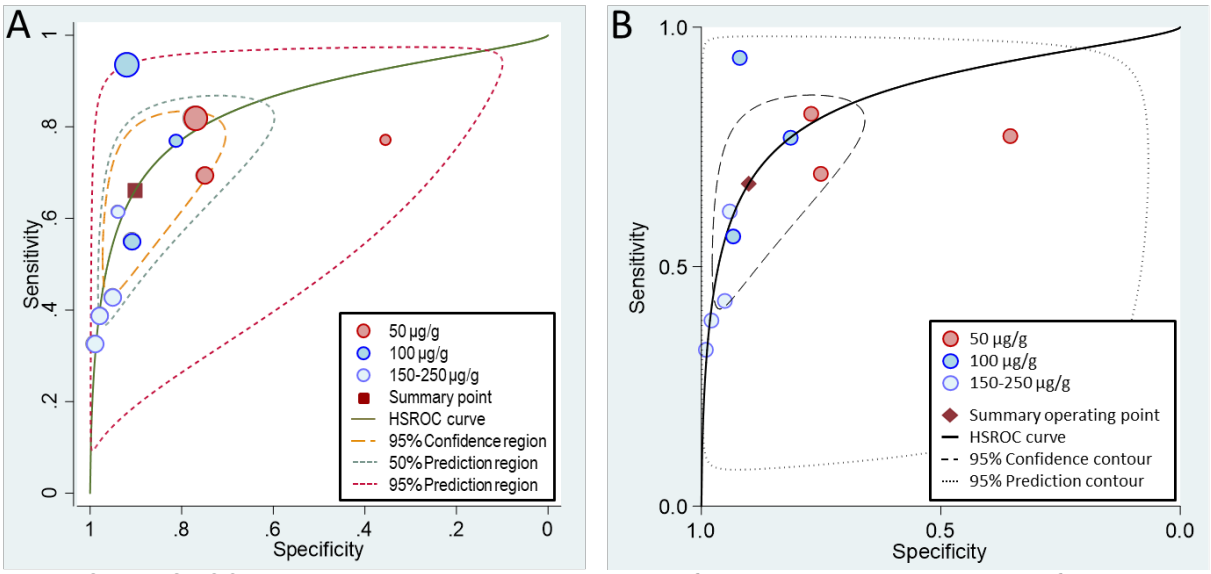


Figure 37 HSROC curves summarising the accuracy of a test compared with the reference standard for different test thresholds

The HSROC curves was generated in STATA using the metandi command (A) and the midas command (B), using the same studies shown in the forest plot in Figure 35. The curves show a trend where the studies with thresholds at or below 50 µg/g were more sensitive but less specific than those with thresholds at or above 100 µg/g. The 95% prediction region, defining where a future study would lie, is much larger when using the midas command.
HSROC = hierarchical summary receiver–operator characteristic

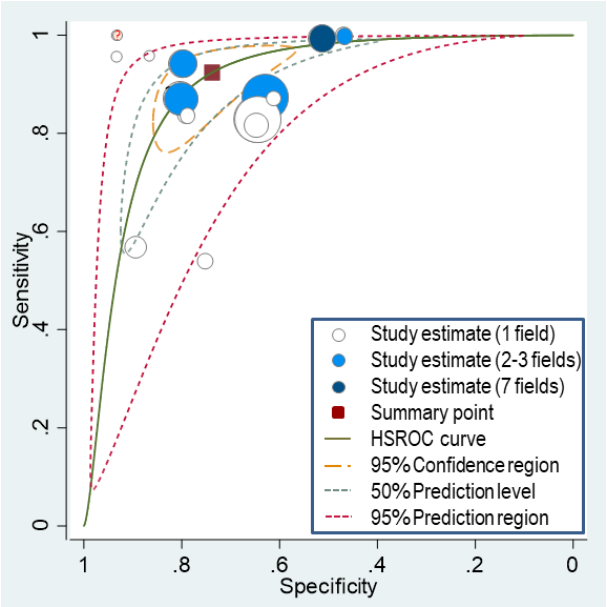


Figure 38 HSROC curve summarising the accuracy of a test compared with the reference standard for different test thresholds where there is no reporting of multiple thresholds in the same study

The HSROC curve was generated in STATA using the metandi command and shows a threshold effect; the sensitivity increases and the specificity decreases as the number of fields included in the test increases.
HSROC = hierarchical summary receiver–operator characteristic

In some cases, a HSROC curve may show a threshold effect even if there are no apparent differences in the test or the population characteristics between studies (Figure 39). This may be due to differences between studies (or laboratories) of test characteristics that may or may not have been reported. For example, differences in the test protocol (timing of processing steps, concentration of the solutes used etc.) or laboratory equipment used, variations in the interpretation of the results (e.g. scoring algorithms, inter-rater variability, etc.) may indicate systematic detection differences between studies, resulting in a "threshold effect". However, care should be taken in concluding a threshold effect in the absence of evidence that different thresholds do apply across studies.

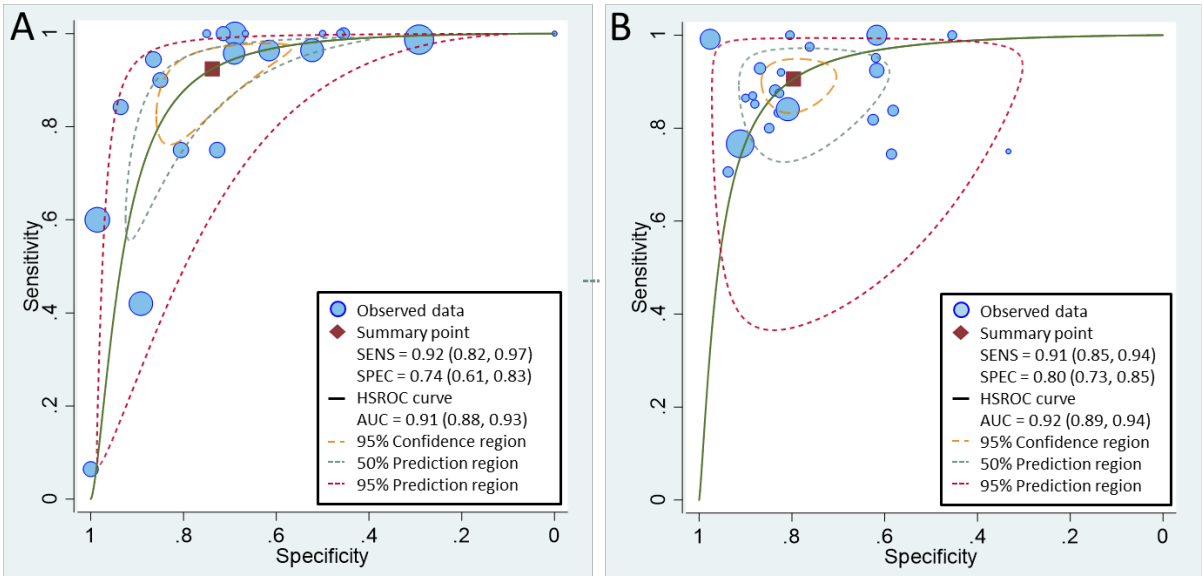


Figure 39 HSROC curves summarising the accuracy of two different tests compared with the relevant reference standard where there were no obvious key differences between studies

The HSROC curves were generated in STATA using the metandi command. The HSROC curve for the test in panel A shows a possible threshold effect, whereas the HSROC curve for the test in panel B does not.
AUC = area under the ROC curve; HSROC = hierarchical summary receiver–operator characteristic; SENS = sensitivity; SPEC = specificity

The AUROC is the average of the true-positive rate over the entire range of false-positive rate values and serves as a global measure of test accuracy that can be interpreted as follows. Given a randomly selected patient with the condition, and a randomly selected patient without the condition, the probability that the patient with the condition would be ranked more highly than the patient without the condition is higher than the AUROC value (Millard, Flach & Higgins 2016). The following guidelines have been suggested for interpretation of AUROC values: for values above 0.9 test accuracy is high, moderate for values between 0.7 and 0.9, and low for values below 0.7 (Swets 1988). An AUROC value of ≤0.50 indicates that the test cannot discriminate between true positives and true negatives, with the curve lying on or below the major diagonal.

Assessing heterogeneity between studies included in a meta-analysis

A test for heterogeneity examines the null hypothesis that all studies are evaluating the same effect. As test accuracy is usually calculated with two correlated estimates (sensitivity and specificity) from the same study, analysing the variability or heterogeneity in these estimates between studies is challenging.

Heterogeneity is usually measured using two different measures, the Cochran's Q statistic or the inconsistency measure, I-squared (I^2). The Cochran's Q statistic is the sum of the squared deviations of each study's estimate from the overall pooled estimate, according to the study weighting in the meta-analysis (Cochran 1954). P values are obtained by comparing the statistic with a Chi-squared ($\chi 2$) distribution with k–1 degrees of freedom, where k is the number of studies. However, the power of the test is low when only a small numbers of studies are included in the meta-analysis, and consequently the test is poor at detecting true heterogeneity as significant.

The I^2 statistic, which does not depend on the number of studies, measures the degree of inconsistency, or the percentage of total variation across studies, that is due to heterogeneity rather than chance. I^2 can be readily calculated from the Cochran's Q statistic as: $100\% \times (Q - df)/Q$, where df equals the degrees of freedom (Higgins et al. 2003). Negative values of I^2 are considered to be equal to zero so that I^2 lies between 0% (no heterogeneity) and 100%. It should be noted that separate I^2 statistics for sensitivity and specificity fails to account for variation explained by the correlation between sensitivity and specificity, as well as for threshold effects, and will overestimate the degree of heterogeneity observed.

Publication bias

Publication bias is usually evaluated by visual inspection of a funnel plot. if there is no publication bias, studies are evenly distributed within the inverted funnel. In the presence of publication bias, the distribution of studies in the funnel plot will be asymmetric. The standard methods for generating a funnel plot to determine publication bias were developed for therapeutic intervention studies by Egger et al (1997) and Begg & Mazumdar (1994) and can be inaccurate for test accuracy studies (van Enst et al. 2014).

The method by Deeks, Macaskill & Irwig (2005) has been developed for use with test accuracy studies. It plots the diagnostic log odds ratio against the effective sample size ($1/ESS^{1/2}$), where the effective sample size is a simple function of the number of diseased and non-diseased individuals. This method

is recommended for test accuracy studies in the 'Cochrane handbook for systematic reviews of diagnostic test accuracy' (Macaskill et al. 2010).

Additional test accuracy measures

When describing the accuracy of a test, there may be additional measures that will provide relevant information for decision makers. When presenting any test accuracy measure, provide an explanation of the measure, and an interpretation of the results. If test accuracy measures are not likely to influence decision making, do not present them.

The number needed to diagnose or misdiagnose

The number needed to diagnose (NND) is the number of patients who need to be examined to correctly identify one person with the disease in the PICO population. The number needed to misdiagnose (NNM) is the number of patients who need to be tested in order for one to be misdiagnosed by the test.

The number needed to diagnose or misdiagnose provides some information about the usefulness of the test in the clinical setting. The fewer patients needed to test to identify someone with the disease and the more patients that are tested before one is misdiagnosed, the more useful the test is to clinicians.

The post-test probability of having the disease with a positive or negative test result

The post-test probability of a test correctly identifying patients with and without disease provides a measure of the usefulness of the test in a clinical setting. The post-test probability can be calculated using either the LR+ and LR– values or the PPV and NPV values.

The pre-test probability of having the biomarker or condition is equivalent to its prevalence rate in the testing population.

Meta-analysis of PPV and NPV from individual studies is not recommended because these values are affected by the prevalence of the biomarker or condition in the testing populations and are not directly comparable when the prevalence rate varies between studies. However, PPV and NPV can be calculated from the pooled sensitivity and specificity estimates by applying the estimated prevalence of the disease or biomarker in the target testing population (formula for this calculation is provided in Table 30). Both PPV and NPV are valuable metrics for the interpretation of test accuracy in the clinical setting. PPV is the percentage of patients with a positive test who actually have the biomarker or condition, and is equivalent to the post-test probability of a positive test result being true. NPV is the percentage of patients with a negative test who do not have the biomarker or condition and its inverse (1–NPV) is equivalent to the post-test probability of a negative test being false. As the prevalence rate increases, for any given pair of sensitivity and specificity values, the PPV will increase and the 1–NPV will decrease, as shown in Figure 40.

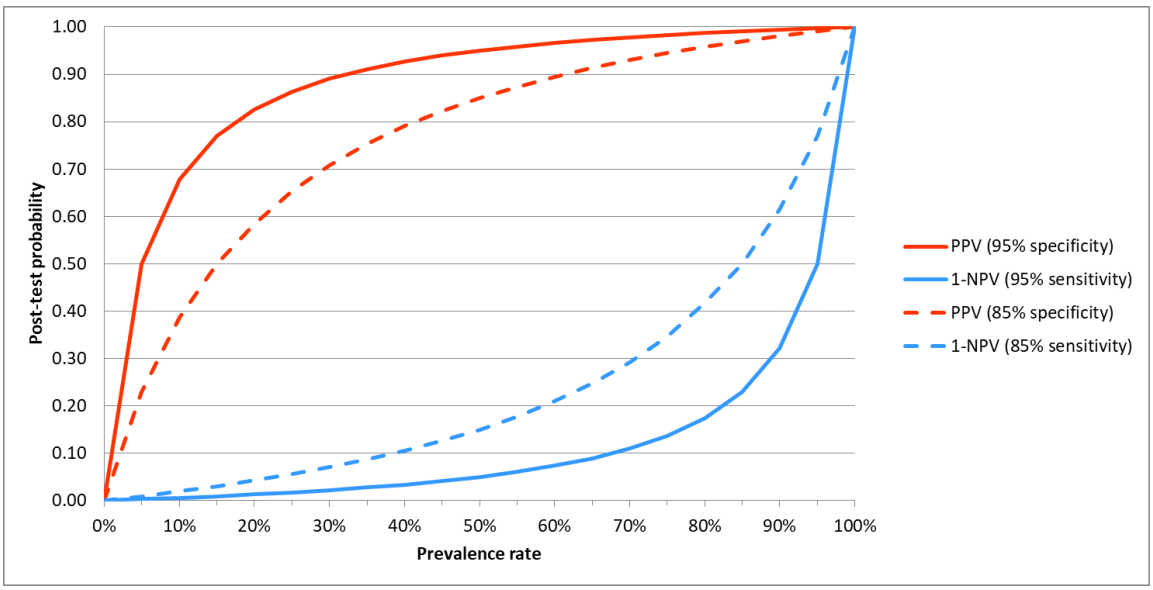


Figure 40 Graph showing the relationship between the prevalence rate and the post-test probability of a positive test being truly positive (PPV) and a negative test being falsely negative (1–NPV)

The solid lines show the post-test probability of being truly positive for a sensitivity and specificity of 95%, the dotted lines show how the post-test probability changes when the sensitivity and specificity are reduced to 85%.
NPV = negative predictive value; PPV = positive predictive value

The summary LRs of a test plus the prevalence rate (pre-test probability) of the biomarker or condition also enables an estimate of the post-test probability of having the biomarker or condition by plotting these values on a Fagan's nomogram.

The red line plots the pre-test prevalence rate and the LR+ to show the post-test probability of having the condition if the test result is positive (Figure 41). The blue line plots the pre-test prevalence rate and the LR– to show the post-test probability of having the condition if the test result is negative.

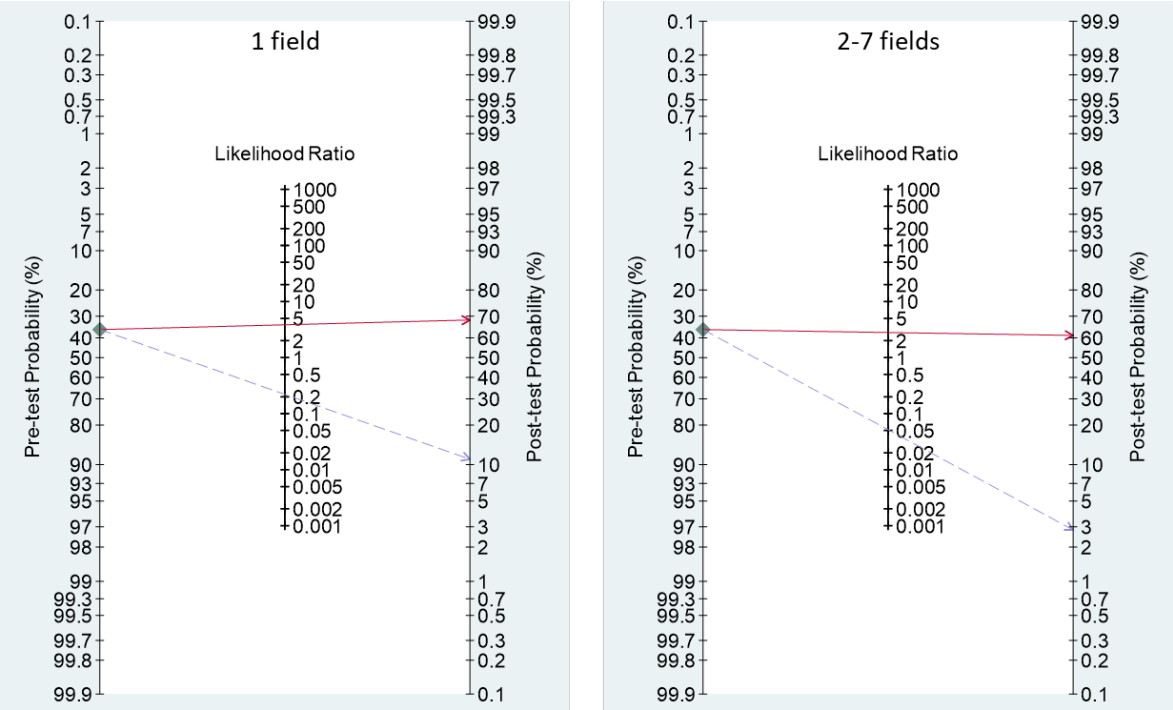


Figure 41 Fagan's nomogram showing the post-test probability of having the condition with a positive or negative imaging test result, depending on the number of fields photographed

The Fagan's nomogram was generated in STATA using the midas command
The prevalence rate (or pre-test probability of having the condition; 36%), LR+ (red line) and LR– (blue line) values were plotted to obtain the post-test probability of having the condition. The post-test probability of having the condition is almost double the pre-test probability with a positive test result, although the number of fields photographed has little effect on the diagnosis. However, the post-test probability of having the condition decreases from 11% to 3% if > 1 field is photographed. This increases the usefulness of the test as a triage test, where only patients with a positive test result are retested with other tests for confirmation of the presence of the condition.
CI = confidence interval; LR+ = positive likelihood ratio; LR– = negative likelihood ratio

Appendix 8 Co-dependent technologies

Placeholder - Co-dependent framework to be inserted.

Appendix 9 Expert opinion

Uses of expert opinion

Consider providing expert opinion to supplement or support the observed data from randomised trials or nonrandomised.

Determining an appropriate body of experts will depend on the nature of the information gap that requires filling. Experts may be panels of medical practitioners, a medical specialty group or consumers. Consumers may provide advice on factors such as the patient relevance of outcomes (particularly if elicited at the time of trial design) or how health technologies might be used. Expert opinion can be useful in several aspects of preparing a PICO Confirmation or an assessment report for consideration by MSAC – for example, to help:

- define the clinical need for the proposed health technology and inform the main indication (discussed in Technical Guidance 1)
- determine how the health technology is most likely to alter the clinical management algorithm (TG 2.6Technical Guidance 2) and support the choice of the main comparator (TG 2.3), noting that a comparator should not be determined by expert opinion alone
- interpret the clinical importance and patient relevance of the outcome measures reported in studies (Appendix 5)
- modify the patterns of health care resource use measured in studies conducted in different settings, such as in other countries (Technical Guidance 22 and Section 4)
- predict which health care resources would be used and how often each would be used to manage outcomes reported in the included studies (Technical Guidance 22 and Section 4)
- estimate the proportion of patients with the medical condition that would be eligible according to the requested listing, and predict uptake rates (Technical Guidance 19and Section 4)
- predict the impact on the utilisation of other health technologies (TG 27.3).

In several examples above, trial data, registry data or analyses of data from other countries, where available, would be used in preference to expert opinion, and it would be expected that the expert opinion supports the applicability of the observed data. An example is to support the representativeness of a utilisation evaluation conducted in another country. In this case, expert opinion reduces uncertainty.

Presenting expert opinion

Justify the use of expert opinion in the introduction of the appropriate section. Include a clear rationale for, and the aims of, eliciting the expert opinion. Where expert opinion is used to fill a gap in information, clearly describe the nature of this gap and indicate the other steps that have been taken to address the gap, such as a literature search.

Describing the collection and collation of expert opinion

Using a well-designed methodology to elicit expert opinion helps to reduce uncertainty. The methods used may vary from large, published questionnaires and surveys with statistical analysis to a summary of interviews with a panel of clinical experts. Present expert opinion as qualitative or quantitative (but not statistically analysed) information.

Include copies of administered surveys or hypothetical scenarios that were presented to experts.

When summarising expert opinions and their variability, interpret the findings, and discuss the limitations and biases of the method chosen. Qualitative studies and interviews should follow best practice for reporting and analysis. Indicate how the opinions have been used in PICO or the assessment report.

Where multiple sources of expert opinion are available to address a single assumption or estimate, compare the results, and assess their concordance or lack of it. Present a summary table that compares multiple sources or multiple variables. Table 31 provides guidance on the details that should be included. Where multiple estimates (or data) are generated to fill a gap in the information – either from multiple sources of expert opinion or a combination of expert opinion and observed data – compare the estimates (or data) and justify the choice of data used in the submission.

Where expert opinion is used in place of observed data, as may occur when observed data are generated from other health care systems or are historical, present both and clearly justify the use of expert opinion. State if expert opinion (compared with alternative sources of data) is likely to lead to a more favourable clinical, economic or financial assessment of the proposed health technology.

MSAC is concerned when information used within the clinical, economic or financial analysis of the proposed health technology is uncertain. Where expert opinion is sought for a disease or condition for which the number of medical practitioners is likely to be large, do not rely on surveys of small numbers of practitioners because this leads to highly uncertain results. In all cases where expert opinion is used to derive estimates for the assessment, use the final estimate, to minimise the risk for MSAC on relying on overestimation of effectiveness or cost-effectiveness, or underestimation of financial implications for the Australian Government or other funding body. To reduce uncertainty associated with expert opinion, provide sensitivity analyses around the derived estimates, or clearly state where results in the assessment are not sensitive to different estimates.

Table 31 Methods to collect and collate expert opinion

Information to be provided	Notes
Criteria for selecting experts	Prefer a random or comprehensive set of health practitioners likely to prescribe the proposed health technology, or the appropriate medical specialty group. In general, an advisory board or group of practitioners associated with the manufacturer or sponsor may not be representative of experts in Australian clinical practice. The generalisability of expert opinion derived from such boards is difficult to assess
Number of experts approached[a]	Where the likely number of health practitioners is large, it is less acceptable to provide expert opinion derived from a small number of practitioners
Number of experts who participated[a]	Assess whether the extent and characteristics of the nonresponders are likely to diminish the representativeness of the opinions provided, compared with the intended sample approached
Declaration of potential conflicts of interest from each expert or medical specialty group whose opinion was sought	Provide a signed statement from each expert and specialty group specifying any potential conflict of interest and stating the nature of any contractual arrangement, including how much payment was offered and accepted. Where the collection of expert opinion has been contracted out, the contractor should provide this statement, reporting on both the arrangements made between the applicant or evaluator and the contractor, and the arrangements made between the contractor and those whose opinions were sought
Background information provided and its consistency with the totality of the evidence provided in the assessment report	Include a copy of any background information provided in the technical document or attachment. If background information has been provided, ask the experts to define the comparative clinical place of the proposed health technology and the main comparator based on this background information. Including the experts' definitions in the technical document or attachment allows an assessment of the consistency of the background information with the evidence provided in the assessment
Method used to collect opinions	For example, were the experts approached individually or was a meeting convened? Was any incentive used to maximise responses?
Medium used to collect opinions	For example, was information gathered by direct interview, telephone interview or self-administered questionnaire?
Questions asked[b]	Explain the design of the tool (quantitative or qualitative). Describe its development. Indicate whether it was pilot tested and, if so, provide the results of that testing and explain how the results were used to improve the questions. On a question-by-question basis, assess the extent to which each question is neutral or biased, and the extent to which each question is open or closed. To allow an independent assessment, include the questionnaire or an outline of the interview questions in the technical document (or attach a copy)
Whether iteration was used in the collation of opinions and, if so, how it was used	The Delphi technique, for example, uses an iterative approach
Number of responses received for each question[a]	Assess whether the extent of any nonresponse is likely to diminish the representativeness of the opinions provided to particular questions, compared with the intended sample approached
Whether all experts agreed with each response	If not, specify (i) the approach used to finalise the estimates (eg the majority opinion or a Delphi technique could be applied; for quantitative results, point estimates [such as the mean, median or mode] could be presented), and (ii) the approach used to present the variability in the opinions (eg present the range of opinions expressed, including common and outlying views; for quantitative results, measures of variance [such as confidence intervals, range, centiles] could be presented)

a Tabulate these information items.

b The way the questions are asked is an important source of potential bias in obtaining expert opinion. A particularly influential extension question extends the respondent beyond 'what' the opinion is (eg what would be done, what extent of benefit would be clinically important) to ask 'why' (eg explain why would you do this, explain why this is important). Conveying these reasons alongside expert opinion–based estimates might help improve their acceptability. Including these explanations in the technical document or attachment would allow the opinions to be assessed based on the underlying reasoning rather than only depending on the authority of the experts.

Appendix 10 Including non-health outcomes in a supplementary analysis

Presenting non-health outcomes

Occasionally, listing a proposed health technology might generate worthwhile impacts that are not captured as health outcomes, such as the value of information to the patient generated by an additional diagnostic test that does not change management of a medical condition.

Supplementary methods to estimate the monetary (or other) value of the non-health benefit may include a conjoint analysis or a discrete choice experiment that includes a monetary attribute, an attribute reflecting a range of options for each of the non-health outcomes of interest, and/or other attributes.

Where there are no other substantive changes in health outcomes between the proposed health technology and its main comparator, this estimate (eg willingness to pay) can be included in a supplementary cost-benefit analysis. Where this cost-benefit analysis results in a consumer surplus, nominate a suitable basis for sharing this consumer surplus between the sponsor and the taxpayer.

Production changes

In the context of health economics analyses, a production change is a change in total output value across society of productive work in the economy. Productivity is a function of output units (eg days of work) multiplied by their value (eg an appropriate daily wage as a proxy for the value of each day of work).

Health interventions may claim to result in a change in production across society associated with patients gaining or losing working time as a result of changes in their health and consequent capacity to work. Less commonly, a health intervention may claim that worker efficiency will be affected, such that the value of work output is changed on a per-unit basis (ie it can be represented by a higher or lower wage).

Changes in production as an outcome of therapy may be included in supplementary analyses in submissions to MSAC, but do not include them in the base-case analysis. This separation allows MSAC to consider the impact of including production changes on the direction and extent of change on the base case. Including production gains favours interventions that improve the health of people who are able, and choose, to return to contributing to societal production and, hence, there are equity implications of including productivity changes in the base case.

If presenting productivity claims associated with a proposed health technology, there are several difficulties in estimating the net present value of production changes. From a societal perspective, the productivity of an individual worker cannot be considered in isolation, but should be considered in the context of a workplace, a workforce and society. The following three underpinning assumptions should be incorporated into all productivity analyses:

- For short-term absence, production will be made up on return to work.
- Employers usually have excess capacity in the labour force to cover absenteeism.
- For long-term absence, production will be made up by a replacement worker who would otherwise be unemployed.

When presenting estimates of the marginal increase in society's production because of the return of healthy workers:

- provide details of the method used and its assumptions
- discount appropriately any productivity changes anticipated beyond one year
- address each of the assumptions listed above when estimating production changes from the potential working time gained or lost (reported in time units).

For example, the claim that there has been a recovery of production lost because of returning to health from an episode of illness depends on demonstrating the following three factors:

- The worker returns to work and the worker is productive.
- The production lost is not made up elsewhere by others in the company or the same worker following return to work.
- No temporary replacement has been employed.

Address each of these three factors to provide robust evidence in support of estimates.

Ensure that estimates of the proportion of people who choose to return to work account for those who would choose not to return (and instead use their time gain on other activities that will have been captured by a gain in utility weights), as well as the influence of incentives provided through sickness benefits, which may operate differently across jurisdictions.

The approach above may be adapted to other contexts, such as a health technology that prevents future episodes of illness, or one that might improve production capacity in individuals who, without the proposed health technology, would otherwise stay at work, although unwell, and therefore function at less than full production capacity.

When the economic approach is a CUA, discuss how the method of estimating productivity changes avoids double-counting the estimates of health-related quality-of-life changes. The utility weights in this analysis already capture these health-related changes because they incorporate the utility impacts of productive capacity for the individual receiving the proposed medicine. These health-related changes are therefore already appropriately included in the denominator of the cost-utility ratio.

Strongly justify any production changes that are combined with surrogate outcome indicators in an economic evaluation, because this combination is generally associated with inappropriately high levels of uncertainty.

Appendix 11 Selection of studies for indirect comparison

Where the approach taken in the assessment report includes an indirect comparison, the studies included will typically be randomised trials of the intervention and randomised trials of the main comparator(s), which share a common reference. In some circumstances, included trials will have transitivity issues that would reduce the certainty of the results of the indirect comparison. If this is the case, justify the exclusion of trials that are unsuitable for use in the indirect comparison.

In general, a simple indirect comparison (ie a pairwise Bucher method) is adequate to estimate the treatment effect of a proposed medical service compared with a main comparator. In some circumstances, more complex methods (such as network meta-analysis) may be appropriate.

A general approach to identifying suitable trials for an indirect comparison is summarised here:

1. Perform appropriate searches to identify all studies of the intervention and of the comparator (Appendix 2).

2. Draw a network diagram to show all the possible links.

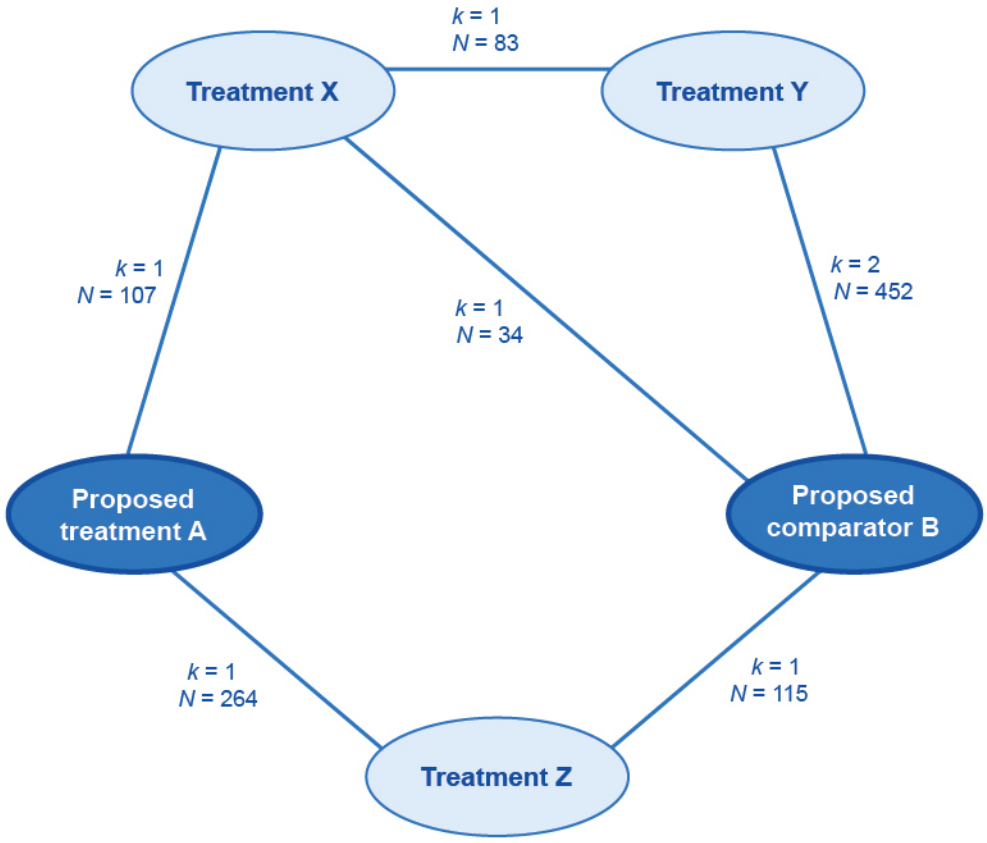


Figure 42 Example network diagram of the trials included to inform an indirect comparison of the proposed medical service with the main comparator

k = number of trials; *N* = number of patients enrolled

3. Where pairwise comparisons are possible, the assessment report may seek to exclude linkages requiring multiple steps, or include these as a supplementary analysis. In the absence of pairwise comparisons, presenting the smallest number of steps is usually preferred. Provide a justification for the choice of the base case.

4. Examine heterogeneity within trial sets and across trial sets, and justify the exclusion of trials with differences in factors that may affect the transitivity of the trials in the indirect comparison.

    - Do not exclude studies on the basis of differences in characteristics that are unlikely to influence the treatment effect for either the proposed medical service, or the proposed comparator. The exclusion of studies due to concerns regarding heterogeneity of characteristics is supported by a demonstrable impact on the treatment effect, and requires a clear explanation.
    - If studies are removed at this step, it is useful to include them in a sensitivity analysis.
    - Possible sources of heterogeneity are listed in Appendix 6.

5. Examine the event rates in the common reference arms. The indirect comparison will, by design, adjust for differences in the event rates in the common reference arms of the studies. However, the success of the indirect comparison is based on an assumption of a constant relative (or absolute) treatment effect in the studies across different baseline risks. This is not certain.

    - If there is evidence from subgroup analyses that there is a constant treatment effect across different risk groups, do not exclude studies on this basis. Select the most appropriate outcome (relative or absolute) which best fits the assumption of constant treatment effects.
    - If it is likely that there is not a constant treatment effect, consider and justify the exclusion of studies on the basis of differences in the event rates in the common reference arms.

6. Present a list of the studies included in the main analysis, the studies included in supplementary or sensitivity analyses, and the studies excluded from all analyses.

Appendix 12 Translating comparative treatment effects of proposed surrogate measures to target clinical outcomes

Introduction

MSAC prefers submissions that do not rely on proposed surrogate measures (PSMs) to inform effectiveness, in terms of patient-relevant or clinically relevant outcomes. Where possible, present evidence from direct randomised trials on the treatment effect of the proposed health technology on clinically relevant outcomes.

Where no such evidence is available, establish the likely comparative treatment effect on clinically relevant outcomes by transforming the comparative treatment effect of a surrogate measure.

A surrogate measure is a biomarker that is intended to substitute for one or more target clinical outcomes (TCOs). Although a surrogate measure may or may not have clinical relevance, it is not the key purpose for treatment, which is to affect the severity of, or the transition to, future TCOs.

Relevant to MSAC, the relationship between a PSM and a TCO is one that quantifies the change in the TCO as a consequence of a change in the PSM. Throughout this appendix, the transformation of the PSM to the TCO should be interpreted as the transformation of the comparative treatment effect on the PSM, to the comparative treatment effect on the TCO.

This appendix takes the following approach:

- 0– Define the PSM and the TCO.
- 0 – Establish the biological reasoning for the link between the PSM and the TCO, including how pivotal the PSM is to the causation pathway of the TCO, and present epidemiological evidence to support this.
- 0 – Present randomised trial evidence to support the nature of the PSM-TCO comparative treatment effect relationship.
- 0 – Translate the comparative treatment effect on the PSM from the studies included in Section 2, to an estimate of the comparative treatment effect for the TCO.

When interpreting the evidence to identify the relationship between the PSM and the TCO (0 of this appendix), and the relationship between the comparative treatment effect on the PSM and the comparative treatment effect on the TCO (0 of this appendix), present indications of causality. That is, the PSM (and the comparative treatment effect on the PSM) always precedes the TCO (and the comparative treatment effect on the TCO), and their associations are strong, measured with high precision, and maintained after adjustment for confounders (if there are sufficient numbers of trials with sufficient information to enable such adjustment).

Use the following types of evidence to analyse a PSM-TCO relationship (listed from strongest to weakest):

1. multitrial meta-regression
2. single trial or small number of randomised trials where individual patient data are available (including multicentre analysis where participants were randomised by centre)
3. one randomised trial – no individual patient data or not randomised by centre

4. no randomised trial data.

Given the uncertainty associated with transforming PSMs to TCOs, ensure that the treatment effect observed on the PSM is robust and unbiased. Bias may result from, for example, issues of study quality, imbalances in baseline characteristics, loss to follow-up, discontinuations, inappropriate dosing, subgroup analysis or adjustments for crossover. Where an unknown proportion of the comparative treatment effect on the PSM may be the result of bias, the estimate of the comparative treatment effect on the TCO will be uncertain. In the absence of a robust estimate of the comparative treatment effect on the PSM, transformation to a comparative treatment effect on the TCO is not informative.

The approach taken in this appendix has been informed by the Surrogate to Final Outcomes Working Group report, and this remains a useful resource when additional explanation is required.[bb]

Definition, selection and measurement

Proposed surrogate measure

Where an intervention may have multiple benefits (eg avoiding multiple strains of a virus or multiple forms of cardiovascular events), a PSM that captures the overall intended clinical outcome is more persuasive. Ensure that the PSM is responsive and able to be measured with reliability and validity.

Define and describe the PSM, with reference to the epidemiological and randomised trial evidence identified in this appendix, by including the following:

- the units of measurement
- the measurement tool(s) or criteria used
- the evidence of reliability from test to test
- the variability across observers or different measurement tools
- the measurement of the comparative treatment effect (eg odds ratio, standardised mean difference).

Ensure that the definition and method of measurement of the PSM are consistent across the evidence. Report and discuss any discrepancies when presenting evidence in this appendix.

Target clinical outcome

Ensure that the choice of TCO is patient-relevant and captures the key purposes for intervening in a disease process. The goal of treatment may be to improve quality of life, or prevent or slow a medical condition in the long term. Ensure that the TCO is consistent with the health states defined in the natural history of the disease or condition. In some cases, more than one TCO may be required to capture the effects of the proposed health technology on the disease or condition process. Justify if the nominated TCO does not capture an outcome of the disease or condition, or an adverse outcome of the treatment. There may be evidence that the proposed health technology has a positive treatment effect for one TCO (eg myocardial infarction) and a negative treatment effect for another TCO (eg haemorrhagic stroke).

[bb] www.pbs.gov.au/info/industry/useful-resources/pbac-feedback

With reference to the epidemiological and randomised trial evidence identified in this appendix, ensure to:

- justify the choice of the TCO and justify the exclusion of other potentially relevant TCOs (particularly those for which the proposed health technology may have a negative treatment effect)
- describe how the TCO is patient-relevant and nominate, with evidence, the extent of change that would be considered meaningful (see 0)
- state whether the TCO is reversible
- state whether the TCO is itself a substitute for a more clinically relevant outcome (multistep transformation to a subsequent TCO is discouraged)
- provide the units of measurement
- list the measurement tools or criteria used
- provide evidence of reliability from test to test
- explore variability across observers or different measurement tools
- describe the measurement of the comparative treatment effect (eg odds ratio, standardised mean difference).

Ensure that the definition and method of measurement of the TCO are consistent across the evidence. Report and discuss any discrepancies when presenting evidence in this appendix.

Relationship between the proposed surrogate measure and the target clinical outcome

When exploring the nature of the PSM-TCO relationship in subsequent parts of this appendix, comment on the following:

- Is the nature of the PSM-TCO relationship still current?
- Have there been changes to treatments or health care systems over time that may have affected the PSM-TCO relationship?
- Is there any evidence of resistance or tolerance to a health technology, or a waning treatment effect over time? Consider and explain any waning treatment effects, and any effects of having no long-term randomised trials that capture the PSM and the TCO.

Derive the PSM-TCO comparative treatment effect relationship from randomised trials that measure both the PSM and the TCO. If this type of evidence is unavailable, it is difficult to quantify the link between changes in the PSM and changes in the TCO. Ensure that the epidemiological evidence in Section 0 of this appendix is unequivocal and robust.

Biological reasoning and epidemiological evidence

Biological reasoning

The information request for biological reasoning concerns the disease pathogenesis and disease or condition pathways, and how the PSM and the TCO relate to them, independent of health technology actions. (Mechanisms of action are presented in Section 0 of this appendix.) To provide confidence that altering the PSM provides clinical benefit, clearly explain the biological relationship between the PSM and the TCO.

Present and discuss the disease or condition pathway, clearly linking the PSM to the TCO. State whether the PSM is a necessary step in the development of the TCO, and discuss how close the development of the PSM is, in both temporal and pathological terms, to the development of the TCO.

Epidemiological evidence

Epidemiological or observational studies support a claimed biological plausibility of the PSM-TCO relationship. Reasons for examining any association are also relevant for investigating the association between the PSM and the TCO.

Describe in detail the epidemiological evidence identified, which may include in vitro studies, animal studies, case reports, cross-sectional observational studies, ecological association studies, retrospective observational cohort studies, non–population based prospective observational cohort studies, or population-based prospective observational cohort studies.

Describe the limitations of the evidence with reference to the study design (eg individual-based associations from observational studies are more convincing than ecological associations).

Present the statistical associations, including the nature or shape of the association, the strength of the association and the precision (95% confidence interval [CI]). Report all relevant statistical outputs, such as regression coefficients and R-squared.

Describe and explain any contradictory findings, primarily where the direction of effect changes, or there is a large difference in the magnitude of effect.

Randomised trial data for all health technologies

Identifying relevant trials

Review the literature systematically to find randomised trials that explore the relationship between the PSM and the TCO, irrespective of the health technology examined. Present the search terms, inclusion criteria and the PRISMA flowchart, clearly showing the exclusion of trials. List the excluded trials and reasons for exclusion in an attachment.

From the list of included trials, compile a list of the health technologies, categorised by mechanism of action or class, that act on the PSM (see Table 32). Present the extension studies associated with the identified trials.

For each mechanism of action, discuss the biological reasoning for the effect of the health technology on the PSM. Discuss whether the mechanism of action of the health technology is the same as, or similar to, the pathological mechanism of the disease or condition. Rationalise any lag in onset of the treatment effect and the implications for the PSM or the TCO, or both.

Table 32 Biological reasoning for the effect of the health technology on the proposed surrogate measure

Class of health technology	Mechanism of action	Biological reasoning for the effect of the health technology on the proposed surrogate measure	Trials available, citations (health technology included in each trial)
[add]	[add]	[add]	[add]
[add]	[add]	[add]	[add]

Trial characteristics

For each of the included trials or meta-analyses, discuss the following factors that may affect the estimate of the relationship between the comparative treatment effect on the PSM and the comparative treatment effect on the TCO:

- The quality of the included trials or meta-analyses (present an assessment of the internal validity of the included trials according to the guidance provided in Appendix 3, in an attachment).
- Whether relevant trials have been excluded from any meta-analyses or meta-regressions.
- Whether the analysis of the PSM was designed prospectively or retrospectively.

Present the characteristics of each of the trials as per Table 33.

Table 33 Characteristics of trials included in the assessment of the relationship between the proposed surrogate measure and the target clinical outcome

Trial and date	Patient characteristics	Disease or condition characteristics	Treatment settings	Measurement of proposed surrogate measure and target clinical outcome
[add]	[add]	[add]	[add]	[add]
[add]	[add]	[add]	[add]	[add]

Trial results

Present the results of the randomised trials and the proposed relationship between the comparative treatment effect on the PSM and the comparative treatment effect on the TCO (Table 34). Where multiple trials exist for a class of health technologies, clearly show the results of a meta-analysis for individual studies. Present the results of any meta-regressions, including the intercept and coefficient (and their 95% CIs), the R-squared for trials and for individuals (if individual patient data are available), and the surrogate threshold effect as determined by prediction bands. Justify where a meta-regression has not been presented.

Discuss the PSM-TCO comparative treatment effect relationship. Include details of the shape of the relationship (eg linear, exponential) and whether there is any evidence of a floor or ceiling effect, below or above which the comparative treatment effect on the PSM no longer predicts a comparative treatment effect on the TCO.

Table 34 Results of randomised trials

Trials/meta-analyses (grouped and meta-analysed by class or mechanism of action)	Baseline value of PSM / final value of PSM[a]	Comparative treatment effect on PSM	Comparative treatment effect on TCO[b]	Proposed relationship (and measure of uncertainty)
[add]	[add]	[add]	[add]	[add]
[add]	[add]	[add]	[add]	[add]

PSM = proposed surrogate measure; TCO = target clinical outcome

a Where the PSM is a continuous variable, present the mean baseline and mean final value for the PSM, separated by treatment arm. Where the PSM is a dichotomous variable, such as progression-free survival, this column may be adapted to show the proportion in each arm achieving the PSM.

b Where the trial has included a placebo, no treatment or best supportive care arm, report the absolute number of TCO events in that arm to give an indication of the baseline risk. A long-standing comparator may also be used as an adequate reference for baseline risk.

Where available, present results of the relationship between the comparative treatment effect for the PSM and the TCO across different trial dates, disease or condition stages, treatment settings and patient populations. State which particular subpopulations (or subpopulations are not included in the overall trial populations) do not have trial evidence available. Where these subpopulations would have access to the health technology through the proposed funding arrangements (Technical Guidance 3), strongly justify the extrapolation of the PSM-TCO relationship to this subpopulation in Section 0 of this appendix.

Discuss where the relationship of the comparative treatment effect for the PSM and the TCO differs across trials, health technologies or mechanisms of action. Discuss possible causes of the heterogeneity – for example:

- mechanism of action of the health technology
- population characteristics
- disease or condition characteristics, or severity
- treatment settings
- definition or measurement of the PSM
- definition or measurement of the TCO
- quality of the trial
- nature of the proposed relationship (eg linear, asymptotic, floor or ceiling effects).

Multiplicity of pathways

Although unexplained heterogeneity is difficult to interpret, heterogeneity that can be linked to a characteristic will require further consideration, particularly if the cause of the difference in the relationship between the PSM and the TCO differs according to mechanism of action of a health technology, population characteristics, or disease or condition characteristics. Where differences in the relationship between the PSM and the TCO are present, it is likely that the TCO can be affected by an alternative pathological pathway that is more or less prevalent across differences in the included trials. Where the PSM-TCO comparative treatment effect relationships differ according to the:

- mechanism of action, explain why different health technologies with similar effects on the PSM may result in different effects on the TCO
- patient characteristics, or disease or condition characteristics, explain why similar changes in the PSM in these subpopulations may result in different effects on the TCO.

Alternative pathological pathways that do not involve the PSM undermine the validity of the PSM. Therefore, where appropriate, exclude trials with health technologies or populations in which the alternative pathway is present if:

- there is compelling evidence of the existence of the alternative pathway (such evidence may be randomised trial evidence linking an alternative PSM with the TCO)

and

- the alternative pathway is not present for the proposed health technology (and the main comparator) or the population in which listing is being sought.

Present evidence to support these claims.

Where trials are removed that have health technologies of different mechanisms of action or populations that do not reflect the proposed listing, present the estimate of the PSM-TCO comparative treatment effect relationship with all trials included as the base case. Remove less-relevant trials through a sensitivity analysis.

Validity of results

For each of the trials, meta-analyses and meta-regressions, compare the observed TCO comparative treatment effect with the predicted effect on the TCO if calculated according to the epidemiological evidence presented in Section 0 of this appendix (Table 35).

Table 35 Comparing randomised trial evidence and epidemiological evidence

Trial, meta-analysis or meta-regression	Comparative treatment effect on PSM	Observed comparative treatment effect on TCO	Predicted comparative treatment effect on TCO after applying the relationship observed in epidemiological studies
[add]	[add]	[add]	[add]
[add]	[add]	[add]	[add]

PSM = proposed surrogate measure; TCO = target clinical outcome

Discuss differences between the observed and predicted comparative treatment effect on the TCO.

Summarising the evidence

Several parameters of the evidence presented are critical to understanding and interpreting the translation of the PSM for the proposed health technology to an estimate of the TCO (Table 36). These are general conditions, outside of which the translation of the PSM to the TCO becomes less certain.

Table 36 Summary of conditions under which the relationship has been determined

Parameter of evidence	Results	Cross-reference
Median baseline value of PSM (IQR)	[add]	[add]
Median final value of PSM (IQR)	[add]	[add]
Median change in PSM (IQR)	[add]	[add]
Median change in PSM for the comparator (IQR)	[add]	[add]
Range of disease or condition severity	[add]	[add]
Range of patient characteristics (eg age, sex, race)	[add]	[add]
Range of trial dates	[add]	[add]
Range of TCO event rates (from placebo arms)[a]	[add]	[add]
Range of estimates of the PSM-TCO comparative treatment effect relationship	[add]	[add]

IQR = interquartile range; PSM = proposed surrogate measure; TCO = target clinical outcome
a Placebo, no treatment or best supportive care arms, or long-standing comparator

Where more than one estimate of the relationship between the comparative treatment effect on the PSM and the comparative treatment effect on the TCO has been established, justify the selection of one estimate for the base case, and present the remainder as sensitivity analyses.

Applying the relationship between comparative treatment effects to the proposed technology

Mechanism of action

When applying the PSM-TCO comparative treatment effect relationship to the trial evidence for the proposed health technology, it is critical that both the proposed health technology and the main

comparator have the same mechanism(s) of action as health technologies for which the PSM-TCO comparative treatment effect has been established in Section 0 of this appendix. When a health technology is not of a class of health technology presented in Section 0 of this appendix, it is not possible to determine to what extent the TCO is affected by changes in the PSM, and to what extent it is affected by alternative pathological pathways or by negative physiological effects. Therefore, where one or both of the proposed health technology and the main comparator are not represented by the mechanism(s) of action in Section A5.3 of this appendix, the comparative treatment effect on the PSM may have a very different relationship to the comparative treatment effect on the TCO. Where this is the case, the transformation of the PSM to the TCO will be uncertain.

Explain the mechanism(s) of action and the biological reasoning for the mechanism(s) of action of the proposed health technology and the main comparator on the PSM and the TCO. Identify differences between the mechanism(s) of action of the proposed health technology, and the main comparator and the health technologies identified in the trial evidence in Section 0 of this appendix. Clearly explain how any differences will not result in a different measurement of the PSM-TCO comparative treatment effect relationship.

Where the proposed health technology and the main comparator are within the same class of health technologies identified in Section 0 of this appendix, it is still important to identify differences in physiological effects, and discuss whether different effects can alter the disease or condition process and, hence, the PSM-TCO comparative treatment effect relationship.

Applicability of the evidence

As outlined in Section 0 of this appendix, the applicability of the results of the relationship between the treatment effect on the PSM and the treatment effect on the TCO to different populations and stages of disease is not guaranteed. However, evidence of consistency across different populations and stages of disease is supportive. Compare the patient population, disease or condition stages and circumstances of use for the proposed health technology and the studies identified in Section 0 of this appendix. If there are differences, justify why the relationship between the treatment effect on the PSM and the treatment effect on the TCO identified in Section 0 is applicable to the clinical trial(s) of the proposed health technology.

The PSM-TCO comparative treatment effect relationship is uncertain beyond the observed ranges for the PSM presented in Section 0 of this appendix. Compare the baseline values of the PSM and the comparative treatment effect on the PSM presented in Section 0 with that observed for the key trials of the proposed health technology, and discuss.

Estimate the comparative treatment effect for the proposed health technology

Present the proposed health technology's comparative treatment effect (with CIs) on the PSM for each trial and for a pooled analysis. Translate this using the relationship proposed in Section 0 of this appendix. The comparative treatment effect on the PSM and the estimate of the PSM-TCO relationship will have a degree of uncertainty; thus, capture this in the statistical approach and present as a 95% CI around the estimated comparative treatment effect on the TCO. Do not simply translate the upper and lower CIs of the comparative treatment effect for the PSM observed in the key trial by the point estimate of the relationship established in Section 0 of this appendix, because this does not adequately capture the uncertainty in the estimate of the comparative treatment effect on the TCO.

Discuss the implications of any surrogate threshold effect identified in Section 0 of this appendix.

State whether there are any concerns about the duration of the treatment effect.