HEALTH TECHNOLOGY ASSESSMENT METHODS FOR EVALUATING MEDICAL TESTS: DEVELOPING A NOVEL APPLICATION OF THE LINKED EVIDENCE APPROACH

TRACY MERLIN

BA(Hons), MPH, AdvDip PM

School of Public Health

Faculty of Health Sciences

University of Adelaide

Submitted for the Degree of Doctor of Philosophy in Medicine

May 2015

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

BACKGROUND

The health consequences of medical testing are often not apparent or easily measured. To address this, the ‘linked evidence approach’ (LEA) was developed to estimate the clinical utility of a test so that policy makers can make informed public funding decisions. Australia has the largest international experience with the application of LEA.

RESEARCH AIM 1

The first aim of the presented research was to investigate the feasibility, utility and policy impact of LEA.

To enable the use of LEA in test evaluation there needed to be a more rigorous approach taken to determine the risk of bias in test accuracy studies. An existing evidence hierarchy recommended by the Australian Government for use in health technology assessment (HTA) was consequently revised between 2005 and 2009 to consider design-related biases in test accuracy studies. The hierarchy underwent a national public consultation and pilot process and became widely used.

A study was conducted to model the overall impact of LEA on health policy; data were extracted from HTA reports commissioned before-and-after the use of LEA was mandated by the Australian Government in 2005. Logistic regression analyses and regression diagnostics were performed to estimate model fit, model specification and to inform model selection. There was no discernible impact of LEA on the direction of public funding decisions (OR=1.36, 95%CI 0.62, 3.01) but the use of LEA *did* strongly predict that a medical test would *not* receive interim funding (X^2=12.63, df=1, p=0.0004). This suggests that the method enables greater certainty in decision-making.

RESEARCH AIM 2

The second aim was to develop guidance on how LEA should be *applied* during the evaluation of medical tests. A systematic literature review was performed on the methods used in HTAs evaluating medical tests so that a decision framework could be constructed to guide the application of LEA and to address potential methodological problems with the approach.

The framework systematises the application of LEA by categorising medical tests into three possible scenarios, namely optimisation, trade-off and disease-spectrum change. The evidence collation and linkage practices need to be tailored to each of these scenarios.

RESEARCH AIM 3

The final aim of the presented research was to adapt LEA to the evaluation of a drug and its companion diagnostic test (‘personalised medicine’).

An analysis of guidance documents and a review of case studies was undertaken to identify key information to guide decisions concerning the reimbursement of personalised medicines. An evaluation framework, incorporating LEA, was created to determine the safety, effectiveness and cost-effectiveness of personalised medicines. 79 evaluation items were proposed and examples provided to demonstrate the linkage of different types of evidence to reduce decision-maker uncertainty. The framework underwent a public consultation and pilot process.

The impact of the evaluation framework on public funding decisions was critically reviewed in the three years’ after the framework was implemented nationally.

CONCLUSIONS

This thesis by publication resulted in three theoretical methods papers (published), one analytical paper (under review) and one published review paper (invited).

The methods developed for these publications were aimed at improving how medical tests are considered and valued by our health systems. LEA enables the clinical utility of medical tests to be estimated, leading to greater certainty for policy makers and reducing the need for ‘interim’ funding decisions. Methods for standardising the application of LEA have allowed consistent information to be provided to policy makers. The adaptation of LEA to the evaluation of personalised medicines has enabled previously siloed funding decisions on companion tests and therapeutics to be integrated.

The research outputs from this thesis have directly affected technology evaluation practice, with consequent impacts on health policy and test subsidy decisions.

DECLARATION

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that the copyright of published works contained within this thesis (as listed on page xiii) resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library catalogue, and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Signed: ______________________________ *Dated:* ______________________________

Tracy Merlin (Candidate)

ACKNOWLEDGEMENTS

Health technology assessment is a multi-disciplinary field characterised by collaboration and team work. As such, I would like to acknowledge each of the co-authors on my papers – this thesis would not be half as developed without having had the opportunity to discuss ideas and arguments with each of you; namely, Adele Weston, Rebecca Tooher, Claude Farah, Camille Schubert, Andrew Mitchell, Sam Lehman, Janet Hiller and Phil Ryan.

My sincere gratitude to those people in the Australian Government Department of Health who continually consider different and better ways to incorporate health technology assessment into policy. These public servants, and government advisory Committees of clinical experts and consumers, have a commitment to the public good that usually goes unrecognised but which has enormous health and economic implications for the Australian public at large.

My thanks to all of the wonderful people at Adelaide Health Technology Assessment (AHTA) for your understanding and support while I have been trying to get this PhD done – particular thanks go to Lashan Clifton, Skye Newton, Camille Schubert, Jacqueline Parsons and Shuhong Wang for holding the fort while I took study leave to ‘write up’. Thank you also to Jackie Street and Taryn Bessen for digging me out of my AHTA hidey-hole every now and then.

To Dr Brian Stein who inadvertently set me on the path of evidence-based medicine and to A/Prof John Moss, who saw my potential as an academic, my sincere appreciation.

In 2007 I chose Prof Janet Hiller and Prof Phil Ryan as my supervisors because I respect them both immensely. The biggest motivation to complete the PhD was to not let them down. Thank you so much for your support and honesty.

Especial thanks to my very good friend, Nikki McCaffrey, for being there and sharing PhD horror stories, along with many other things; and to Greg, for always listening and showing me you care.

To mum, for instilling a love of books and learning; and to dad, for understanding what it means to strive.

Finally, and most importantly, this thesis is dedicated to the two young people who have lived through this PhD with me - **To my children, Caitlin and Alex**.

MANUSCRIPTS CONTRIBUTING TO THESIS

PUBLISHED

Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology*, 2009, 9:34 doi:10.1186/1471-2288-9-34. Available at: http://www.biomedcentral.com/1471-2288/9/34

[Highly accessed designation by BMC Medical Research Methodology – 21,970 BioMed Central accesses; 103 citations in Google Scholar; 52 citations in ISI Web of Science (April 2015)].

Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing codependent technologies. *Medical Decision Making*, April 2013; 33(3):333-342.doi:10.1177/0272989X12452341. Available at: http://mdm.sagepub.com/content/33/3/333

[12 citations in Google Scholar; 6 citations in ISI Web of Science; 13 in Altmetric (measure of attention) which indicates the article is highly scored in this journal (ranked #22 of 369) and is in the top 25% of all articles measured by attention (April 2015)]

Merlin T, Lehman S, Ryan P, Hiller JE. The 'linked evidence approach' to assess medical tests: a critical analysis. *International Journal of Technology Assessment in Health Care*, July 2013; 29(3):343-350, doi: 10.1017/S0266462313000287. Available at:

http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8955598

[4 citations in Google Scholar; 1 citation in ISI Web of Science (April 2015)]

Merlin T. The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines. *Personalized Medicine*, July 2014; 11(4): 435-448. Available at: http://www.futuremedicine.com/doi/abs/10.2217/pme.14.28

[Not yet cited]

UNDER REVIEW

Merlin T, Ryan P, Hiller JE. Impact of the ‘linked evidence approach’ method on policies to publicly fund diagnostic, staging and screening medical tests. *Medical Decision Making* [submitted May 2015].

CONFERENCE PRESENTATIONS ARISING FROM THESIS

INTERNATIONAL CONFERENCES

Methods

1. Merlin T, Mujoomdar M, Kisser A, Wurcel V. *Living in testing times: translating evidence on medical tests and companion diagnostics into reimbursement decisions.* Panel presentation. XII Annual Meeting HTAi 2015, Oslo, Norway, 15-17 June 2015.

2. Merlin T, Schubert C. *Evaluating medical tests for coverage decisions using the linked evidence approach.* Pre-conference full day workshop. XI Annual Meeting Health Technology Assessment international 2014, Washington DC, June 15 2014.

3. Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. *How to assess personalised medicines for reimbursement decisions? Developing a framework for Australia.* Oral presentation. IX Annual Meeting Health Technology Assessment international 2012, Bilbao, 25-27 June 2012.

4. Merlin T, Lehman S. *The benefits and flaws of the Linked Evidence Approach (LEA) to assess diagnostic and screening tests.* Poster presentation. VII Annual Meeting Health Technology Assessment international 2010, Dublin, 6-9 June 2010.

5. Merlin TL, Moss J, Hiller J. *Location, location, location – the impact of health care setting of technology use on both HTA results and funding policy.* Poster presentation. V Annual Meeting Health Technology Assessment international 2008, Montreal, 6-9 July, 2008.

6. Merlin T, Brooks A, Lord S, Hiller JE. *How to assess the effectiveness of a triage test in the context of direct versus linked evidence?* Poster presentation. 3rd Annual Health Technology Assessment international. July 3-5, 2006. Adelaide, South Australia.

7. Merlin T, Middleton P, Salisbury J, Weston A. *Ways to ensure evidence-based clinical practice guidelines are of high quality. Discussion workshop.* W52, p196. XIII Cochrane Colloquium, Melbourne, Australia. October 22-26, 2005.

8. Merlin T, Weston A, Tooher R (NHMRC "Levels" Working Party). *Re-assessing and revising "levels of evidence" in the critical appraisal process.* Oral presentation. O32, p.49. XIII Cochrane Colloquium, Melbourne, Australia. October 22-26, 2005.

9. Merlin T, Weston A, Tooher R. *Revising a national standard: redevelopment of the Australian NHMRC evidence hierarchy.* Italian Journal of Public Health (Supplement 1), Summer 2005, Year 3, 2(2): 156. [Oral presentation. Bringing HTA into practice. 2nd Annual Meeting, Rome. June 20-22, 2005]

Case studies

10. Vogan A, Schubert C, Parsons J, Morona J, Merlin T. *Assessing the cost-effectiveness of HBA1c testing in the diagnosis of type II diabetes and the impact of an imperfect diagnostic reference standard.* Poster presentation. Society for Medical Decision Making 36th Annual North American Meeting, Miami, Florida, 18-22 October 2014.
11. Vogan A, Schubert C, Parsons J, Morona J, Merlin T. *The impact of an imperfect diagnostic reference standard on the cost-effectiveness of HbA1c testing in the diagnosis of Type II diabetes.* Oral presentation. XI Annual Meeting HTAi 2014, Washington DC, 16-18 June 2014.
12. Kessels S, Schubert C, Newton S, Merlin T. *Assessment of catheter-free (wireless) ambulatory oesophageal PH monitoring for Gastro- Oesophageal Reflux Disease (GORD).* Poster presentation. XI Annual Meeting HTAi 2014, Washington DC, 16-18 June 2014.
13. Milverton J, Ellery B, Newton S, Kessels S, Merlin T. *CT colonography for those at high risk or symptomatic for colorectal cancer.* Poster presentation. XI Annual Meeting HTAi 2014, Washington DC, 16-18 June 2014.
14. Newton S, Wang S, Schubert C, Merlin T. *One small step for an individual, one giant leap for their family: considerations required for assessing the cost-effectiveness of genetic tests.* Oral presentation. X Annual Meeting HTAi 2013, Seoul, 17-19 June 2013.
15. Morona J, Newton S, Merlin T. *Genetic testing for VHL disease: limited benefit to the individual, but reduced surveillance for family members.* Poster presentation. X Annual Meeting HTAi 2013, Seoul, 17-19 June 2013.
16. Newton S, Fitzgerald P, Merlin T. *Mutation testing of the RET gene for MEN2.* Poster presentation. X Annual Meeting HTAi 2013, Seoul, 17-19 June 2013.
17. Buckley E, Merlin T, Mundy L, Hiller JE. *Molecular testing for myeloproliferative disease: The conundrum of an imperfect reference standard.* Poster presentation. VII

Annual Meeting Health Technology Assessment international 2010, Dublin, 6-9 June 2010.

18. Newton S, Street J, Merlin T, Hiller JE. *Is MRI effective for staging newly diagnosed rectal carcinoma? That depends on whether it changes management.* Oral presentation. VI Annual Meeting Health Technology Assessment international 2009, Singapore, 22-24 June, 2009.

19. Merlin TL, Newton S, Wang S, Hiller J. *Technology assessment in the context of workforce shortages and changing supply: digital mammography.* Oral presentation. V Annual Meeting Health Technology Assessment international 2008, Montreal, 6-9 July, 2008.

NATIONAL CONFERENCES

20. Merlin T. *Examining the challenges to health technology assessment from personalised medicine: the need for innovative approaches.* Australian Health Technology Assessment Conference, Sydney. November 26-27, 2012.

INVITED ADDRESSES ARISING FROM THESIS

INTERNATIONAL

1. Invited presentation. *Best practice in the evaluation of companion diagnostics (workshop)*. European Diagnostics Manufacturers Association. June 11, 2015, Brussels, Belgium.

NATIONAL

2. Training workshop. *Evaluating medical tests and co-dependent technologies for coverage decisions using the linked evidence approach*. Australian Government Department of Health. September 12, 2014, Canberra.
3. Invited Plenary Presentation. *Examining the challenges to health technology assessment from personalised medicine: the need for innovative approaches*. Australian Health Technology Assessment Conference, Sydney. November 26-27, 2012.
4. Training workshop. *Evaluating co-dependent technology submissions to inform PBAC decision-making*. Australian Government Department of Health and Ageing. October 26, 2012, Canberra.
5. Invited Panel Presentation. *Challenges for independent evaluation: MSAC assessments – past, present and future*. ARCS Scientific Congress, National Convention Centre, Canberra. August 2, 2011.
6. Invited Panel Presentation. *The FORM grading method: advantages and challenges*. National Health and Medical Research Council Guideline Development Symposium, Melbourne, June 29, 2011.
7. Invited Seminar Presentation. *Feasibility of the linked evidence approach when assessing diagnostic tests for public funding*. Screening and Test Evaluation Program (STEP) Seminar, University of Sydney, Sydney. March 16, 2011.
8. Invited Panel Presentation. *Rationale for proposed description of evidence needs*. ARCS Scientific Congress, National Convention Centre, Canberra. September 14, 2010.
9. Invited Plenary Presentation. *Developing a framework for co-dependent technologies for reimbursement.* Test Evaluation Symposium. NHMRC Clinical Trials Centre and

Screening and Test Evaluation Program (STEP), University of Sydney. Sydney. September 8, 2010.

10. Panel Discussion. *Better defining evidence requirements for medical tests.* Test Evaluation Symposium. NHMRC Clinical Trials Centre and Screening and Test Evaluation Program (STEP), University of Sydney. Sydney. September 8, 2010

11. Invited Plenary Presentation. *Personalised medicine initiatives*. Evidence-based Pathology Seminar. The Royal College of Pathologists of Australasia. Coogee Beach, Sydney. September 6, 2010.

LOCAL

12. PhD Progress Seminar. *Health technology assessment methods for determining the clinical effectiveness of diagnostic tests: an evaluation of the utility of the Linked Evidence Approach*. School of Population Health and Clinical Practice, University of Adelaide, May 16, 2012.

13. Invited Seminar. *Developing a national framework for evaluating personalised medicines*. Research Conversations, School of Population Health and Clinical Practice, University of Adelaide, November 24, 2011.

14. PhD Progress Seminar. *Feasibility of the Linked Evidence Approach (LEA) when assessing diagnostic tests for public funding*. School of Population Health and Clinical Practice, University of Adelaide, March 10, 2011.

15. Invited Seminar. *Investment and disinvestment in health technologies by policy-makers: An unusual case-study*. School of Population Health and Clinical Practice, University of Adelaide, February 6, 2009.

16. PhD Progress Seminar. *Methods for assessing diagnostic tests in a Health Technology Assessment (HTA) framework: Are they appropriate for triage tests II?* Discipline of Public Health, University of Adelaide, August 22, 2008.

17. PhD Progress Seminar. *Methods for assessing diagnostic tests in a Health Technology Assessment (HTA) framework: Are they appropriate for triage tests?* Discipline of Public Health, University of Adelaide, November 21, 2007.

ABBREVIATIONS

ACCE	Analytic validity, Clinical validity, Clinical utility, and Ethical, legal, social implications
AHRQ	Agency for Healthcare Research and Quality
AHTA	Adelaide Health Technology Assessment
AIC	Akaike information criterion
AUC	Area under the curve
CDC	Centers for Disease Control and Prevention
CEBM	Centre for Evidence Based Medicine
CE-mark	Conformité Européenne - mark
CED	Coverage with evidence development
CER	Comparative effectiveness research
CI	Confidence interval
CNV	Copy number variation
DMAC	Data Management and Analysis Centre
DNA	Deoxyribonucleic acid
EBM	Evidence-based medicine
EGAPP	Evaluation of Genomic Applications in Practice and Prevention
EGFR	Epidermal growth factor receptor
EMA	European Medicines Agency
FDA	Food and Drug Administration
FFPE	Formalin-fixed paraffin embedded
FISH	Fluorescent *in situ* hybridisation
FN	False negative
FP	False positive

G-I-N	Guidelines International Network
GRADE	Grading of Recommendations Assessment, Development and Evaluation
HBV DNA	Hepatitis B Virus Deoxyribonucleic acid
HER2	Human Epidermal Growth Factor Receptor 2
HIV	Human Immunodeficiency Virus
HRM	High resolution melt method
HTA	Health Technology Assessment
HTAAP	Health Technology Assessment Access Point
HTAi	Health Technology Assessment international
ICER	Incremental Cost Effectiveness Ratio
IHC	Immunohistochemistry
INAHTA	International Network of Agencies for Health Technology Assessment
ITFOM	Information Technology Future Of Medicine
KIT D816V	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
K-RAS	Kirsten rat sarcoma viral oncogene homolog
LEA	Linked evidence approach
MBS	Medicare Benefits Schedule
MLPA	Multiplex ligation-dependent probe amplification
MSAC	Medical Services Advisory Committee
NATA	National Association of Testing Authorities
NHMRC	National Health and Medical Research Council
NHS CRD	National Health Service Centre for Reviews and Dissemination
NICE	National Institute for Health and Care Excellence

NPV	Negative predictive value
NSCLC	Non-small cell lung cancer
OECD	Organisation for Economic Cooperation and Development
OR	Odds Ratio
PBAC	Pharmaceutical Benefits Advisory Committee
PBS	Pharmaceutical Benefits Schedule
PCR	Polymerase chain reaction
PCT	Pragmatic clinical trial
PDGFR rearrangements	Platelet-derived growth factor receptor
PhD	Doctor of Philosophy
PLAC	Prostheses List Advisory Committee
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
RCT	Randomised controlled trial
RNA	Ribonucleic acid
ROC	Receiver operating characteristic
RR	Relative Risk
SE	Standard error
SIGN	Scottish Intercollegiate Guidelines Network
SNPs	Single nucleotide polymorphisms
SRDT	Systematic Reviews of Diagnostic Tests
SRE	Systematic review of evidence
SRT	Systematic review of trials
TGA	Therapeutic Goods Administration

TN	True negative
TP	True positive
UK	United Kingdom
USA	United States of America
USPSTF	United States Preventive Services Task Force

CHAPTER 1

INTRODUCTION

Health Technology Assessment (HTA) involves the systematic examination of the technical performance, safety, clinical efficacy and effectiveness, cost, cost-effectiveness, organisational implications, and social, legal and ethical considerations associated with the introduction of 'health technologies' (Busse et al. 2002). Its aim is to provide a tailored evaluation of a health technology in order to provide a rational basis for the development of health policy. 'Health technologies' is a term that encompasses various health interventions including pharmaceuticals, medical or surgical procedures, devices, medical tests, public health programs, and the health services used to deliver these interventions (Busse et al. 2002). Evidence-based information is provided to policy makers so that they can make appropriate regulatory and public funding decisions regarding these health technologies.

As the focus of HTA is usually at the health system level, the final policy decision on whether to introduce, fund, or continue to provide a health technology can affect public access to the intervention, health resource allocation, health service delivery, clinical practice and ultimately the health of the population (Andradas et al. 2008).

HTA takes a *global* approach to assessing evidence – usually a systematic review of the international literature – and then applies that assessment to the *local* situation. Thus, consumer, social, legal, ethical, modelled economic and political impacts of a 'health technology' in the local setting may be given consideration in an HTA, depending on the purpose of the HTA and the 'technology' being assessed (Draborg et al. 2005).

One class of health technology that has proliferated in recent years is medical testing. There has been both an increase in the use of currently available tests (Alter, Stukel & Newman 2006; Jackson, JB & Balfour 1988) as well as an increase in the development of new tests. For example, molecular diagnostic testing in the areas of infectious diseases, pharmacogenomics, genetics and oncology has seen recent, rapid and escalating growth (Suthers 2008; Wolcott, Schwartz & Goodman 2008).

In the past, medical tests were often only of concern to policy makers when they involved expensive, high cost equipment (Velasco-Garrido & Busse 2005), such as the various imaging modalities like positron emission tomography or computed tomography. However, there has been a growing realisation within the health technology assessment community that the consequences of medical testing (even of a simple blood test), although often not immediately apparent, can have a significant health and cost impact on society (NICE 2011).

The rationale for using a test is that it will allow better targeting of a treatment which will, in turn, result in better health outcomes for the patient. The consequences of medical testing are often not apparent because it is rare to find *direct* research evidence demonstrating the effect on patient health outcomes of using a particular medical test in a particular way (di Ruffano et al. 2012).

Because the evidence base on medical tests is often not ideal for answering the questions being asked by decision-makers (ie the impact of testing on patient health outcomes), there are methodological subtleties that need to be addressed when assessing medical tests to inform health policy. It is an over-simplification to assess a medical test in terms of its accuracy alone. Tests are used for various purposes but in general three aspects of testing require evaluation –

- test accuracy – accuracy of either a single test or of a test strategy [1], compared to a reference standard test, in making a 'correct' diagnosis or classification of disease or stage of a disease;
- the influence of the test results on the clinicians' choice of management or treatment options for the patient; and
- the impact of that treatment choice on patient health outcomes.

Information is needed on all of these aspects of the testing pathway. This information is the basis of the *linked evidence approach*, and is used when direct research evidence is insufficient or not available to inform decisions regarding the effectiveness of a medical test.

In 2005 the committee responsible for advising the Australian Government on whether Medicare funding of medical services is warranted, the Medical Services Advisory Committee (MSAC), instituted a guideline for the assessment and evaluation of diagnostic tests within a health technology assessment framework (MSAC 2005a). These guidelines provided for the *linkage* of different types of evidence in instances where *direct* trial evidence of test effectiveness was not available.

[1] Several tests are used whereby the previous test informs the next test – or in some cases several tests are performed concurrently until a diagnosis is made.

The use of *linked evidence* in the HTA of medical tests has not been formally investigated. It is unclear whether this linked evidence approach (LEA) is a "one size fits all" methodology that is relevant or applicable to all types of medical tests and test purposes, including diagnostic, staging, screening and companion diagnostics used in personalised medicine, or whether used as additional, replacement or triage tests to current testing strategies. It is also unclear whether the methodology is interpreted and applied in the same way by different health technology assessors.

However, even apart from these technical aspects, there are two broader reasons why the LEA method needs to be evaluated:

1. Although Australia has long experience with LEA, the method has recently been recognised as the preferred approach by international HTA agencies including the National Institute of Health and Clinical Excellence (NICE) (NICE 2011), the US Agency for Healthcare Research and Quality (AHRQ) (AHRQ 2010) and the European Network of Health Technology Assessment Agencies (EuNetHTA) (EUnetHTA 2008). With the method becoming common practice in the HTA of medical tests, a critical review of the strengths and weaknesses of LEA is needed; and
2. As the aim of HTA is to provide consistent, robust advice to policy makers, it is necessary to determine what effect, if any, the use of LEA has on policy decisions. The method used to assess medical tests could impact on the validity of the results of an HTA, as well as have serious implications – and the concomitant public health impact - for policy decisions made on the basis of that HTA. This is particularly relevant in Australia where there is such a close link between HTA and the policy-making process (Hailey 2009).

Aims of the PhD

Accurate HTAs that capture all the probable and relevant impacts of a medical test when delivered as part of a health service are essential to ensure that the formulation of health policy to introduce and/or fund the test is predicated on the right assumptions and information.

The aims of this thesis are to:

- Investigate the feasibility, utility and impact on policy of the use of the linked evidence approach in health technology assessment;
- To develop guidance on how the linked evidence approach can be applied; and
- To adapt the linked evidence approach to the evaluation of personalised medicines - that is, the use of a genetic test to target a pharmaceutical treatment.

Research questions

This PhD thesis will address the following questions:

1. Is the use of the linked evidence approach feasible when assessing medical tests for public funding decisions?
2. What effect (if any) has the linked evidence approach methodology had on Australian policy makers' decisions to publicly fund medical tests?
3. Are there any specific situations where the use of the linked evidence approach is inadequate? If so, are there ways that the approach can be improved?
4. Can the linked evidence approach be feasibly adapted to the evaluation of personalised medicines?

RESEARCH OBJECTIVES

The objectives of this thesis are to:

- Determine whether the use of the linked evidence approach in the evaluation of medical tests is appropriate for informing health policy;
- Assess the strengths and weaknesses of the methodology;
- Analyse whether the method can be applied to different tests and testing situations;
- Provide recommendations on how the methodology could be improved; and

- Provide guidance and tools to assist researchers and evaluators with the application of LEA and its adaptation to the HTA of personalised medicines.

RESEARCH PROGRAM

This is a thesis by publication. As such, the research questions posed above have been answered separately through five distinct publications. Each publication encapsulates a discrete body of research activity but with the common theme of developing methods for evaluating diagnostic tests. Like most theses by publication, there is unavoidable repetition of some content in the published papers, simply because for each paper the reader needs to be oriented as to the purpose and background behind the presented methodology. The totality of this published research allows a critical analysis of the linked evidence approach in the evaluation of medical tests for the purpose of informing health policy.

The thesis is structured as follows:

Chapter 1 – the rationale for undertaking the research.

Chapter 2 – an overview of HTA, the development of HTA in Australia, and background on the use of medical testing in Australia.

Chapters 3 and 4 – an introduction to the evaluation of test accuracy studies, followed by research addressing Research Question 1 (Paper 1 – published); namely, whether LEA is a methodology that is feasible for use in the evaluation of medical tests.

Chapter 4 – an introduction to the application of LEA methodology, followed by research addressing Research Question 2 (Paper 2 – under review); namely, determining the impact of LEA on health policy by comparing the funding decisions for medical tests both before and after the methodology was in widespread use in Australia.

Chapter 5 – an assessment of the strengths and weaknesses of LEA methodology through a systematic review of HTAs of medical tests that used the approach. On the basis of this research a decision framework is presented that provides guidance on how the LEA method should be applied, as indicated by Research Question 3 (Paper 3 – published).

Chapter 6 – a novel application of LEA to the evaluation of personalised medicines is presented to answer Research Question 4 (Paper 4 – published).

Chapter 7 – a synthesis of the material discussed in Chapters 4, 5 and 6 and a discussion of the ‘real world’ experience with application of the method to personalised medicines (Paper 5 – published).

Chapter 8 – overall conclusions arising from this research program are presented, along with a description of the significance of the methodological work and its contribution to the HTA field. Problems encountered while conducting the research are discussed and suggestions are offered for future research in the area.

CHAPTER 2

LITERATURE REVIEW

What is Health Technology Assessment?

As with evidence-based medicine and clinical practice guideline development, HTA belongs to a group of best practice and quality assurance activities in the healthcare sector (Busse et al. 2002). These kinds of activities are characterised by a systematic and structured way of answering questions by evaluating and synthesising available evidence. The primary audience of HTA consists of decision-makers in the health system at the policy level, while the other best practice activities are primarily aimed at the clinical level (Hailey 2003; Velasco-Garrido & Busse 2005).

There has been some confusion with regard to the different spheres of activity of evidence-based medicine, clinical practice guideline development, comparative effectiveness research (a term used primarily in the USA) and HTA. Luce et al. (2010) developed an organising framework to visually explain the relationship between these evidence-based spheres of activity (Figure 1).

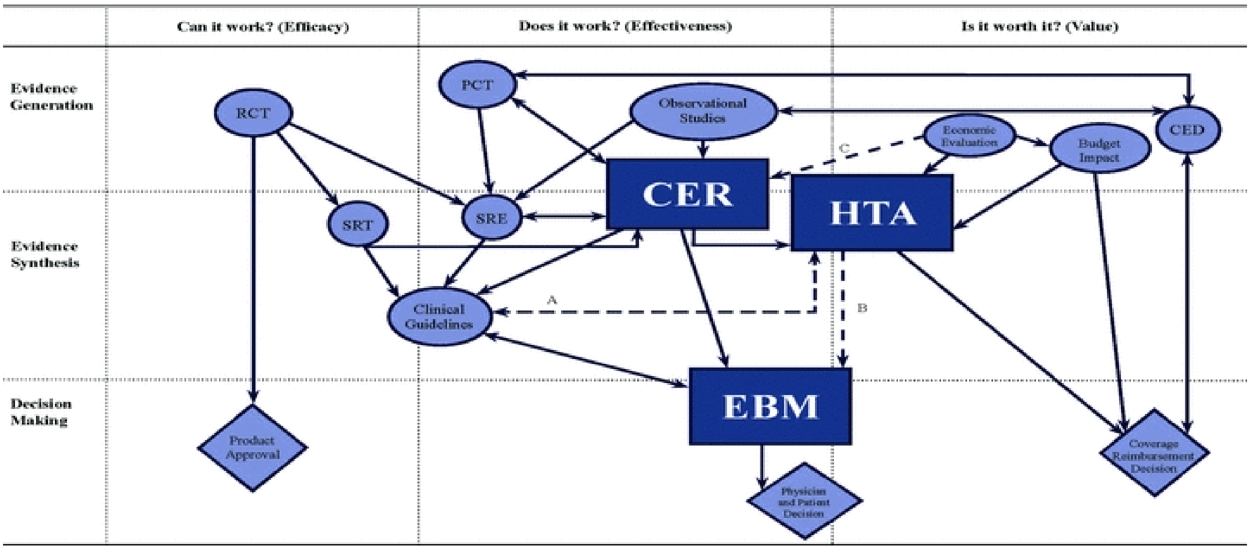
In this diagrammatic representation it is clear that HTA is viewed as being largely concerned with the value of a health intervention. 'Value' in this context involves an assessment of the health benefits obtained from the intervention; that is, the effectiveness of the intervention. The extent of the health benefit obtained from the intervention is then considered and a decision made as to whether this warrants public, or subsidised, funding of the intervention. Public funding of the intervention increases the likelihood that the health benefits will be accessible to the broader community. HTA has, therefore, been defined as

> "*a form of policy research that systematically examines the short- and long-term consequences, in terms of health and resource use, of the application of a health technology, a set of related technologies or a technology related issue*" (Henshall et al. 1997).

The International Network of Agencies for Health Technology Assessment defines a health technology as

> *"prevention and rehabilitation, vaccines, pharmaceuticals and devices, medical and surgical procedures, and the systems within which health is protected and maintained".*[2]

FIGURE 1 SPHERES OF EVIDENCE-BASED ACTIVITY



Source: (Luce et al. 2010) © 2010 Milbank Memorial Fund.

RCT= randomised controlled trial, CER= comparative effectiveness research, PCT= pragmatic clinical trial HTA= health technology assessment, SRT= systematic review of trials, EBM= evidence-based medicine, SRE= systematic review of evidence, CED= coverage with evidence development.

Solid lines indicate clear relationships, and dotted lines indicate disputed relationships. Diamonds represent decision processes, and circles and ovals represent all other evidence activities, except for the rectangles, which are reserved for EBM, HTA, and CER.

Like evidence-based medicine, HTA assesses the safety and effectiveness of a health intervention. However, it is also concerned with the context of a health technology's introduction, and this includes factors affecting the diffusion of the technology and the social, ethical, organisational, professional and economic consequences of technology implementation (Velasco-Garrido & Busse 2005).

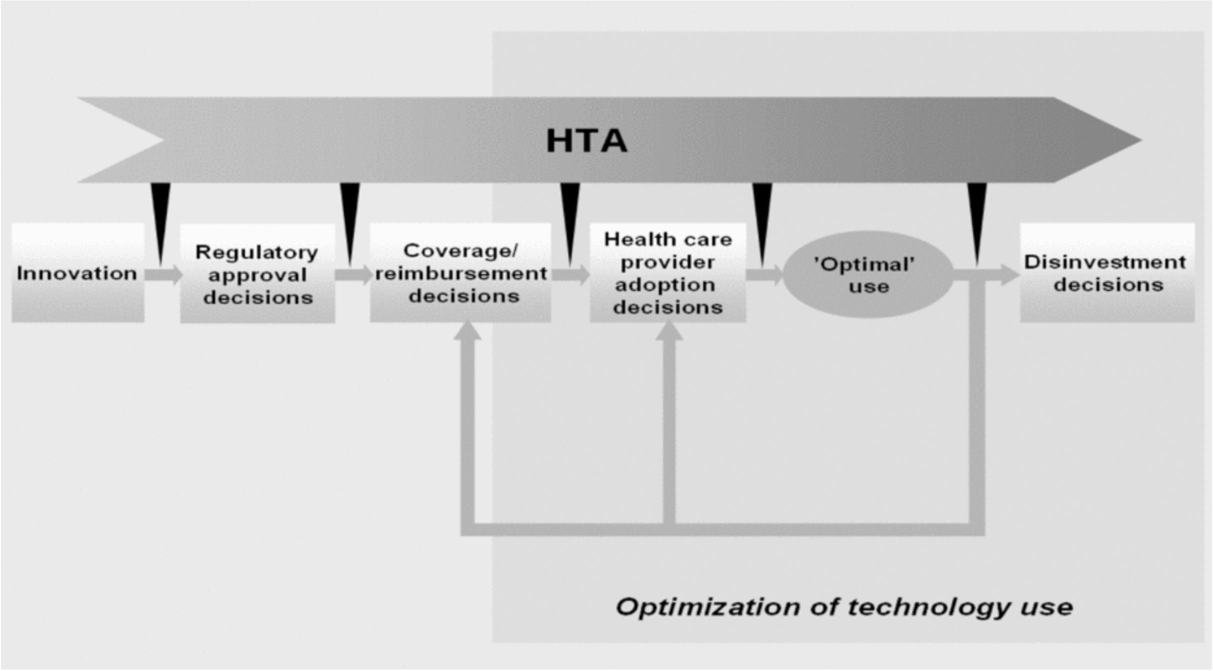
The plethora of primary research on particular health topics, often with conflicting results, means that it is difficult for health policy makers, planners, clinicians and consumers to

[2] http://www.inahta.org/HTA/

determine whether a particular health ‘technology’ or intervention is safe, effective and cost-effective – and thus whether or not it should be introduced, practiced or funded. HTA “*supports the process of decision-making in health care at the policy level by providing reliable information*” (Velasco-Garrido & Busse 2005).

Figure 2 provides a schematic indicating trigger points where HTA can be used to inform policy decisions on the introduction, use and removal of health technologies.

FIGURE 2 POTENTIAL USES OF HTA DURING THE TECHNOLOGY LIFE CYCLE


HTA
Innovation
Regulatory approval decisions
Coverage/ reimbursement decisions
Health care provider adoption decisions
'Optimal' use
Disinvestment decisions
Optimization of technology use

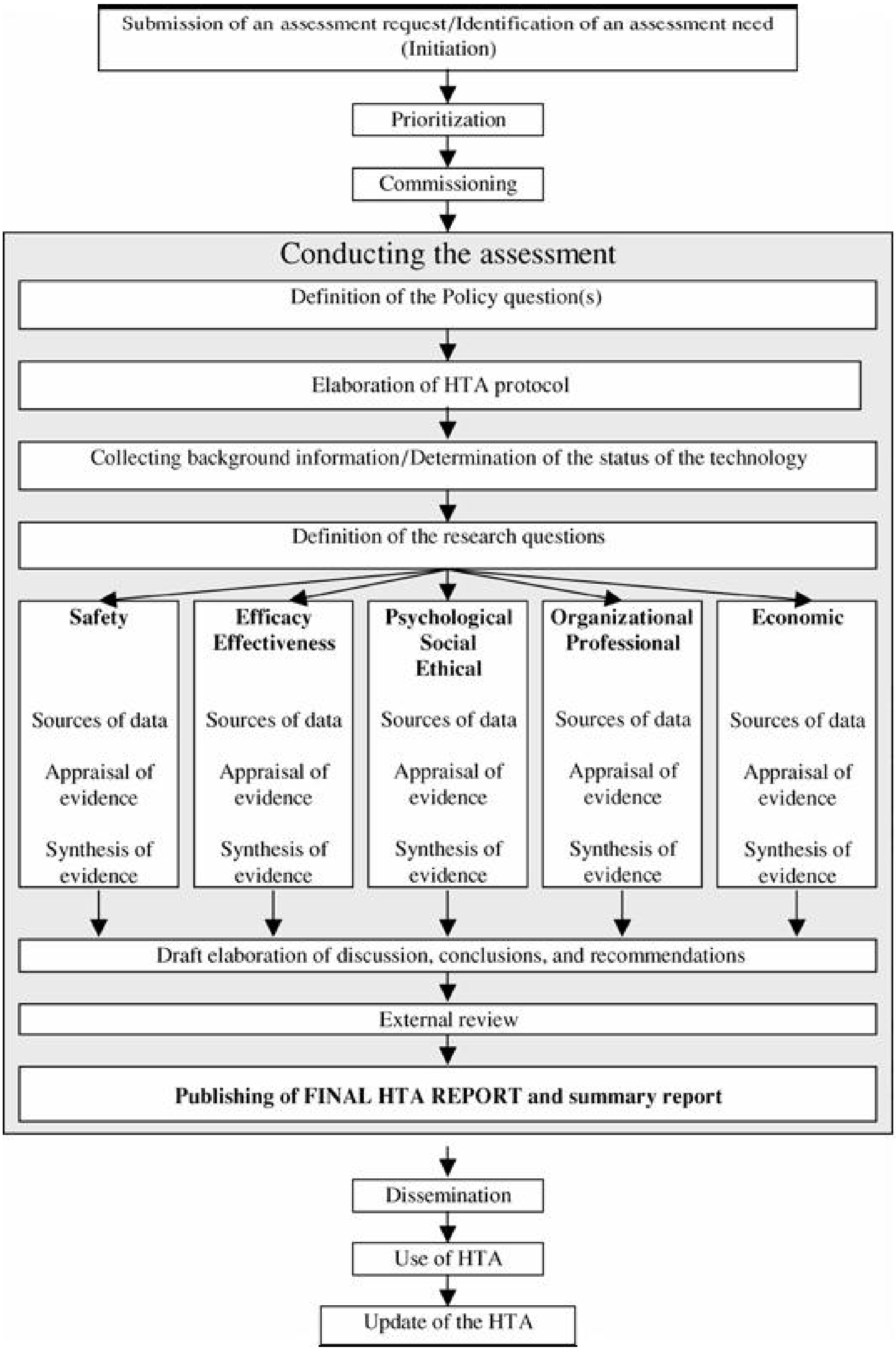
Source: (Fronsdal et al. 2010)

HTA methods

HTA, in its original form, was often restricted to 'providing reliable information' on new technologies that consisted primarily of very expensive medical equipment. Over the years the focus of HTA has expanded to address all levels of decision making in health care. There are various frameworks and methodological guidelines for conducting HTA but there is no standard approach (Draborg et al. 2005). HTA programs rely on different methods that are often tailored to the scope of the policy question being addressed, the financial resources available, time constraints, and other factors.

The process by which HTA can inform policy is given in Figure 3 but the specific characteristics usually depend upon the health system in which the HTA is being delivered. A survey in 2005 determined that 85% (n=69) of OECD countries used HTA in formal decision-making regarding health technologies (OECD 2005). However, the way in which health systems incorporate HTA in to their decision-making differs. In some health systems the HTA program has a direct impact on policy (eg Australia), while in others the impact is indirect (eg Canada) and information may have to be 'brokered' to policy makers (Hailey 2003; Hivon et al. 2005).

FIGURE 3 PROCESS OF HEALTH TECHNOLOGY ASSESSMENT


Submission of an assessment request/Identification of an assessment need (Initiation)
Prioritization
Commissioning
Conducting the assessment
Definition of the Policy question(s)
Elaboration of HTA protocol
Collecting background information/Determination of the status of the technology
Definition of the research questions
Safety
Efficacy Effectiveness
Psychological Social Ethical
Organizational Professional
Economic
Sources of data
Appraisal of evidence
Synthesis of evidence
Draft elaboration of discussion, conclusions, and recommendations
External review
Publishing of FINAL HTA REPORT and summary report
Dissemination
Use of HTA
Update of the HTA

Source: (Busse et al 2002)

In general HTA aims to 'globalise the evidence and localise the decision'. 'Globalisation' relates to the systematic synthesis of the international (globalised) evidence regarding the intervention; this is commonly undertaken with the aid of a systematic literature review. 'Localisation' consists of contextualising the results of the collated evidence to local health care practices and systems. It involves engaging with the local experts, clinicians, consumers and decision makers who would play important roles in the dissemination and utilisation of the health technology (HTAi and INAHTA 2007; Velasco-Garrido & Busse 2005).

"GLOBALISE THE EVIDENCE"

For 'globalisation' of the evidence, methods have been developed to conduct the synthesis of primary international research in a transparent, objective, systematic, and evidence-based manner to ensure the impact of bias is limited and to enable confidence in the conclusions (Draborg & Gyrd-Hansen 2005; Higgins & Green 2005; Khan et al. 2001a; MSAC 2000). The primary method is a systematic literature review, although there is variability in the scope and comprehensiveness of these evidence syntheses.

HTAs of therapeutic interventions generally include systematic literature reviews that use methodology similar to those proposed by the Cochrane Collaboration[3] for assessing the effectiveness of therapies (Higgins & Green 2005) but usually also containing pragmatic (sometimes considered 'weaker') types of evidence (Draborg & Gyrd-Hansen 2005). These types of evidence are incorporated because often a decision must be made before randomised controlled trials (or good quality evidence) are available; and particularly when randomised controlled trial evidence is unlikely to ever be available. HTAs of diagnostic interventions require slightly different systematic review methods (see page 19) but the principles are the same. Key components of systematic reviews in HTAs usually include (Busse et al. 2002; Clarke & Oxman 2003; Cooper & Hedges 1994; Mulrow, Cook & Davidoff 1997):

- the development of a specific research question (which relates to the policy question);
- a transparent methodical process defined *a priori* (ie a review protocol). This comprises the methodology for conducting the review and describes the criteria for

[3] http://www.cochrane.org/

deciding what sort of literature would be the most appropriate to include in the review to answer the question. These eligibility criteria relate to the appropriate target population, the intervention being assessed, the comparator against which the health technology's effectiveness will be measured, and the types of outcomes that will be used to answer the question (eg safety, effectiveness and cost-effectiveness outcomes). Literature sources and search strategies are delineated, and methods for critical appraisal of the literature, data extraction and data synthesis are made explicit. The review protocol is integral to the conduct of a systematic literature review as it is the "recipe" by which the review is systematically conducted;

- an exhaustive search for relevant primary research on the topic through databases cataloguing the literature, the internet, as well as potentially repositories of grey literature (literature that is difficult to find, including published government reports, theses, technical reports, non-peer-reviewed literature etc.);
- extraction of data from the studies and critical appraisal of this research to determine whether the study results are likely to be correct or have been influenced by chance, confounding or bias;
- an attempt to answer the research question and to resolve conflicts in the literature through a narrative and/or quantitative (meta-analysis) synthesis of the data; and
- derive conclusions from the synthesised evidence to inform the policy question, identify research gaps and suggest ways of producing future research on the topic.

"LOCALISE THE DECISION"

To 'localise' the impact of the health technology, several factors may be addressed in the HTA. These can include:

- selection of the relevant comparator – likely use of the technology locally; what would the technology be replacing? Or be used in addition to?;
- cost-effectiveness of the technology in the local health system – usually involving economic modelling;
- ethical issues;
- local access issues;

- consumer preferences;
- workforce planning; and
- training/credentialing of users of the technology.

HTA in Australia

REGULATION

In Australia all therapeutic goods, such as medicines and blood products, medical devices, biologicals, prostheses and laboratory tests, are required to be approved by the Therapeutic Goods Administration (TGA) before they can be marketed or exported. The TGA is Australia's principal regulator of therapeutic goods (McEwen 2007; Therapeutic Goods Administration 2015).

The TGA has three primary functions (Therapeutic Goods Administration 2015):

1. pre-market assessment and approval of healthcare products intended for supply in Australia. Products are required to be registered or listed on the Australian Register of Therapeutic Goods (ARTG) for import, export, or sale in Australia.
2. the licensing of Australian manufacturers and certifying of overseas manufacturers to Australian standards.
3. the post-market surveillance of therapeutic goods (i.e. adverse event monitoring), and therapeutic goods advertising.

Devices and medicines are classified into risk categories, based on their potential to cause harm and different risk-categories require different levels of assessment in order to gain ARTG approval. Higher risk medicines require registration on the ARTG and are evaluated for their quality, safety and efficacy. Lower risk medicines that contain pre-approved, low risk ingredients are simply listed on the ARTG but there cannot be any implication that they will be useful in the treatment or prevention of serious illnesses. For medical devices, higher risk devices are evaluated for quality, safety and performance, while lower risk devices are not evaluated for performance (Therapeutic Goods Administration 2015).

The 'value' of these health technologies is, however, assessed through a different mechanism.

REIMBURSEMENT

In Australia there is no separation of HTA 'advice' and coverage or public funding decisions at the federal level (Jackson, T 2007). HTA is undertaken to directly inform the decision-making of three committees who have distinct responsibility for making funding decisions about different technologies. These committees are the:

- Medical Services Advisory Committee (MSAC) for decisions about medical services (ie procedures, devices, tests, consultations) to be funded by Medicare, as well as programs that may be funded through other arrangements or agreements between the federal government and the jurisdictions (eg a neonatal hearing screening program);
- Pharmaceutical Benefits Advisory Committee (PBAC) for decisions about public funding of pharmaceuticals (PBS); and the
- Prostheses List Advisory Committee (PLAC) for decisions about listing of prostheses and human tissue prostheses on the Prostheses List and the benefit to be paid by private health insurers.

These committees take a population health perspective to assess the safety, effectiveness and cost-effectiveness of health interventions and then provide advice to government as to whether public funding is warranted, as well as what restrictions need to be put in place to access the technology. Policies made on the basis of these evidence-based assessments affect the health of the nation to a greater or lesser extent, depending on the technology or intervention being assessed.

Australia has formally invested in HTA since the 1990s to assess the safety, effectiveness and cost-effectiveness of new 'technologies'. The latter requirement of cost-effectiveness initially occurred with pharmaceuticals in 1993 (through the PBAC) and then medical services in 1998 (through MSAC) (Jackson, T 2007). This process built upon existing TGA regulatory processes for assessing the safety and efficacy of therapeutic goods; beginning in the 1960s for pharmaceuticals (McEwen 2007; Productivity Commission 2005), and in the 1980s for medical devices or technologies, especially those medical technologies of national import and/or high cost (McEwen 2007; MSAC 2005b).

The introduction of a 'fourth hurdle' in the Australian regulatory process in 1993 - of economic evaluation or determining the cost-effectiveness of a technology – was a groundbreaking and controversial policy in the international context (Dickson, Hurst & Jacobzone 2003).

This policy was initially only applied to the assessment of pharmaceuticals because the evaluation of pharmaceuticals is a relatively straight forward process. The pharmaceutical evidence base is often of good quality because (1) treatment trials are generally uncomplicated to mount; and (2) the return on money invested in pharmaceuticals by large multinational companies is substantial, and thus there are often resources available for good clinical trial design and analysis.

The HTA model in Australia with respect to pharmaceuticals involves a comprehensive evidence-based submission (systematic review and economic analysis) to the PBAC by the pharmaceutical industry (or contractors to Industry) which is independently appraised and critiqued by experts or evaluators (usually from independent academic institutions) contracted to the Australian Government (Pharmaceutical Benefits Advisory Committee 2008; Productivity Commission 2005).

HTA for medical services in Australia follows a different model. The medical device or service market in Australia is smaller than that for pharmaceuticals. Submissions for a new or subsidised medical service are equally as likely to arise as referrals from Government or from submissions from health professional organisations or non-profit organisations, as from manufacturers or industry. As a consequence, the submission process allows two options for evaluation:

(1) a submission-based assessment, whereby evidence is supplied by the applicant that is critiqued by an independent evaluator (similar to the PBAC evaluation process), or

(2) a contracted assessment – an independent evaluation usually conducted by an academic group – is commissioned by the Australian Government. In both cases the submission or independent evaluation must follow a protocol agreed by Government.

The process is informed by clinical advice from relevant craft groups and consumers. The final assessment report and/or critique is submitted to the Medical Services Advisory Committee (MSAC) and recommendations for public funding are then made to the Minister for Health. This includes listing of the medical service on the Medicare Benefits Schedule (MBS) (Australian Government Department of Health and Ageing 2012b). Subsidisation through the MBS facilitates equitable access to medical services by Australian patients. The current process of medical service evaluation is the result of several public reviews of HTA in

Australia (Australian Government Department of Health and Ageing 2009a; MSAC 2005b; Productivity Commission 2005).

Not only has the *process* of medical service assessment in Australia received scrutiny but so has the *quality* of the HTAs. A review of treatment or intervention HTAs conducted in response to an application (rather than a referral) to MSAC has shown variable quality and consistency in evaluations, with gradual improvements over time (Petherick et al. 2007).

Medical tests

'Health technologies' is a term that encompasses various health interventions including pharmaceuticals, medical or surgical procedures, devices, public health programs, medical tests, and the health services used to deliver these interventions (Busse et al. 2002).

A medical test is:

> *"any measurement (including an examination or investigation) used to identify individuals who may benefit from therapeutic intervention."* (Muir Gray 2001)

Measurements are done in various ways– through assessing:

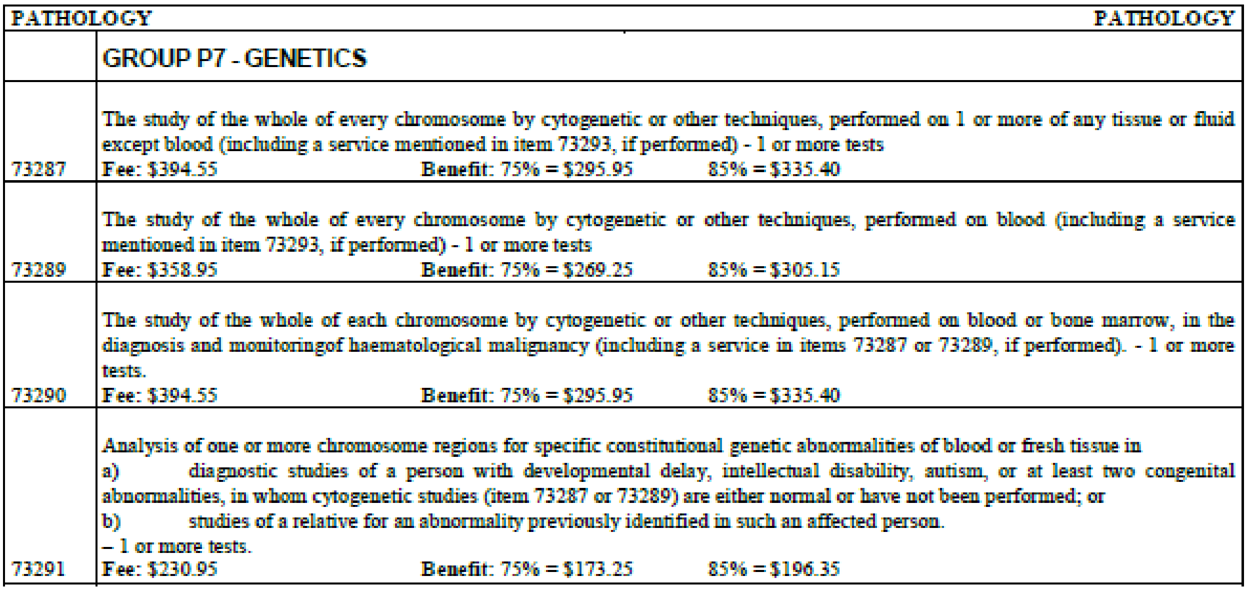
- Symptoms – something a patient feels;
- Signs – something a health professional can observe;
- Laboratory results – expressed numerically;
- Radiological images –interpreted visually; and
- Pathological specimens – interpreted visually (Muir Gray 2001).

Tests are performed for different clinical reasons. These include:

- **Diagnosis** – to identify *symptomatic* individuals who will or will not benefit from a therapy;
- **Screening** – to identify *asymptomatic* individuals who might benefit from a therapy;
- **Monitoring** – to assess the effect of treatment and determine whether treatment should be continued, modified or ceased;
- **Prognosis** – to provide an indication of the future course of the disease; and
- **Risk assessment** – to indicate presence or absence of risk of a condition.

Approximately 50% of medical services on the Australian Medicare Benefits Schedule (MBS) involve a medical test (Australian Government Department of Health and Ageing 2012a). Figure 4 provides a snapshot of a medical test once it is listed on the MBS, using a genetic test as an example. The MBS indicates the subsidy (benefit) that the Australian Government will pay towards the cost of particular medical services provided by health professionals. Generally this subsidy is 85% of the Schedule fee but for services provided in private hospitals the subsidy is 75% of the Schedule fee.

FIGURE 4 EXAMPLE OF MEDICAL TEST ITEMS ON THE MEDICARE BENEFITS SCHEDULE

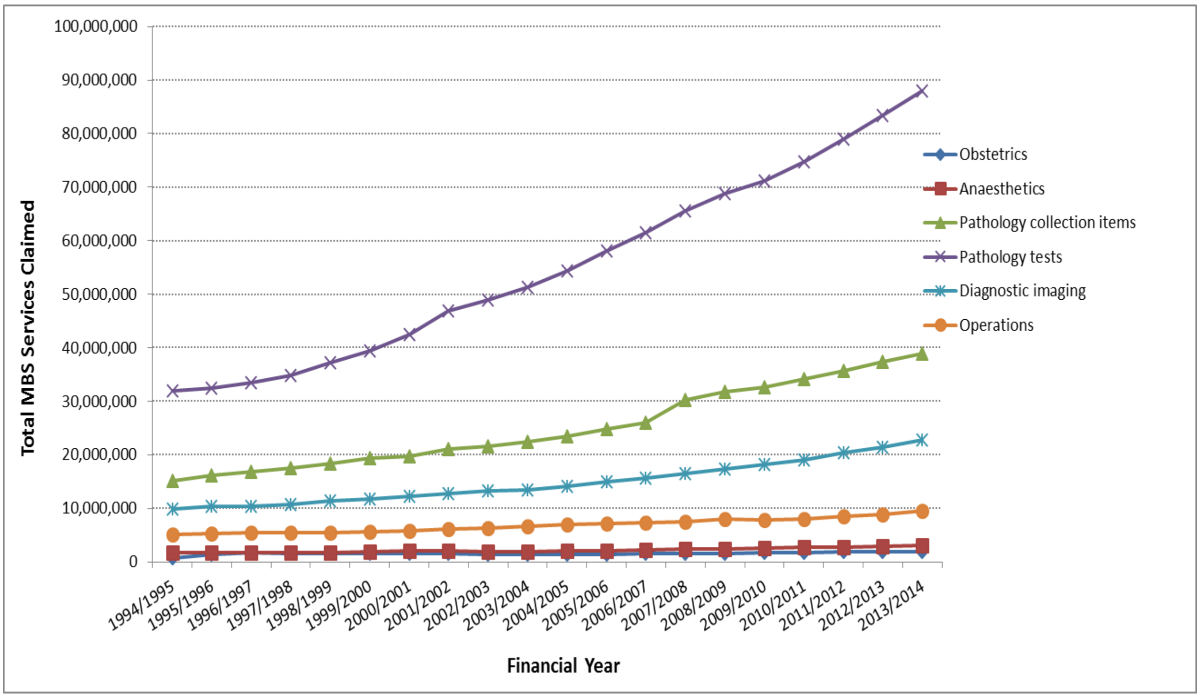


PATHOLOGY	PATHOLOGY
	GROUP P7 - GENETICS
73287	The study of the whole of every chromosome by cytogenetic or other techniques, performed on 1 or more of any tissue or fluid except blood (including a service mentioned in item 73293, if performed) - 1 or more tests Fee: $394.55 Benefit: 75% = $295.95 85% = $335.40
73289	The study of the whole of every chromosome by cytogenetic or other techniques, performed on blood (including a service mentioned in item 73293, if performed) - 1 or more tests Fee: $358.95 Benefit: 75% = $269.25 85% = $305.15
73290	The study of the whole of each chromosome by cytogenetic or other techniques, performed on blood or bone marrow, in the diagnosis and monitoringof haematological malignancy (including a service in items 73287 or 73289, if performed). - 1 or more tests. Fee: $394.55 Benefit: 75% = $295.95 85% = $335.40
73291	Analysis of one or more chromosome regions for specific constitutional genetic abnormalities of blood or fresh tissue in a) diagnostic studies of a person with developmental delay, intellectual disability, autism, or at least two congenital abnormalities, in whom cytogenetic studies (item 73287 or 73289) are either normal or have not been performed; or b) studies of a relative for an abnormality previously identified in such an affected person. – 1 or more tests. Fee: $230.95 Benefit: 75% = $173.25 85% = $196.35

Source: (Australian Government Department of Health and Ageing 2012a)

There has been a substantive increase in medical testing, most notably pathology tests, in recent years in Australia. Figure 5 provides an overview of MBS services claimed between 1994 and 2014, according to broad types of medical service. Results appear to indicate that diagnostic medical services such as diagnostic imaging, pathology collection services and pathology tests, have increased steeply in comparison to other types of services such as obstetrics, anaesthetics and, particular therapeutic services like surgical operations.

FIGURE 5 MEDICARE BENEFITS SCHEDULE SERVICES CLAIMED BETWEEN 1994 AND 2014, BY BROAD TYPE OF SERVICE

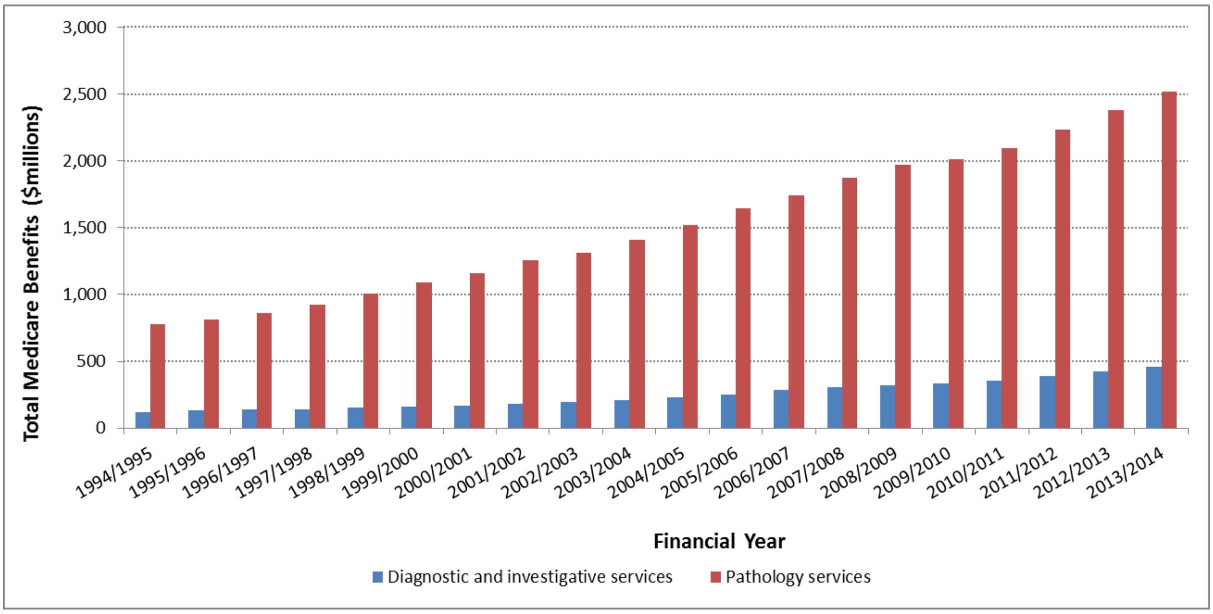


Source: *All Medicare by Broad Type of Service (BTOS) processed from July 1994 to July 2014.*

Note: Calculated using Medicare Australia data. http://www.medicareaustralia.gov.au

Pathology testing nearly trebled (increasing 2.7 times) between 1994 and 2014, with costs escalating at approximately the same rate – that is, increasing from $778 million in 1994 to $2.52 billion in 2014 (increasing 3.24 times). The rise in the volume of diagnostic and interpretive services increased five-fold between 1994 and 2014 and the total expenditure on diagnostic and investigative services by the Government increased 5.8 times, from $79 million in 1994 to $457 million in 2014 (Figure 6).

FIGURE 6 MEDICARE BENEFITS SCHEDULE BENEFITS PAID BETWEEN 1994 AND 2014 FOR DIAGNOSTIC SERVICES AND PATHOLOGY TESTS



Source: *All Medicare by Broad Type of Service (BTOS) processed from July 1994 to July 2014.* Calculated using Medicare Australia data. http://www.medicareaustralia.gov.au

The increase in medical testing may, in part, be driven by servicing the ageing “baby boomer” generation in Australia. However, it would be expected that a population-driven increase in demand for diagnostic services would be fairly evenly matched with increases in treatment and therapeutic services. Figure 5 indicates that this is not the case. In fact many people consider that there is an unprecedented proliferation of medical tests as a consequence of “over-diagnosis”; that is, treating ‘normal’ human conditions as pathologies or identifying diseases and treating them, when - without the use of tests - they would not ordinarily affect the survival of the individual (Bleyer A & Gilbert Welch H 2012; Moynihan 2013; Tikkinen et al. 2012).

In addition to the problem of “over-diagnosis” using existing tests, there also appears to be a general increase in the development of new tests (Alter, Stukel & Newman 2006; Jackson, JB & Balfour 1988). Many of these are ‘add on’ tests that provide further surety of a diagnosis, but which do not replace other tests – nearly half of the medical tests evaluated by MSAC between 1999-2014 were ‘add on’ tests (see Paper 2). As a consequence, the pool of medical tests is expanding.

Molecular testing, in particular, for infectious diseases and oncology has seen recent, rapid and escalating growth (Suthers 2008; Wolcott, Schwartz & Goodman 2008). Molecular testing commonly involves the identification of specific sequences of human DNA or RNA to identify errors (mutations) that may or may not be associated with disease (ie single nucleotide polymorphism, gene insertion, deletion or rearrangement). With the mapping of the human genome, the scientific understanding of genes and their functional roles has expanded. In oncology this understanding of the impact on genes and genetic expression (proteins) on the regulation of cell growth and reproduction has transformed the treatment of cancer[4]. Genetic tests are used to identify patients with tumours that have particular biomarkers that are predicted to respond to a targeted drug therapy (pharmacogenetics) or predicted to be resistant to specific therapies. This is commonly known as personalised medicine and the 'companion diagnostic tests' needed to identify these specific genetic biomarkers are leading the surge in new medical tests.

With the increase in the availability and marketing of new medical tests it is clear that there is a current (and likely future) demand from policy makers to make informed decisions as to whether the public should have access to these tests and whether the specimen collection and interpretation of tests should be subsidised by Government.

With this increased demand for evidence-based policy advice on new medical tests the HTA community needed to develop standardised and accurate HTA methods to capture all the probable and relevant impacts of the tests when delivered as part of a health service. It was essential that the formulation of health policy to introduce and/or fund these tests was predicated on the right assumptions and information – both in terms of protecting the public from the harms of inaccurate testing, which can lead to inappropriate or delayed treatment, as well as the cost inefficiencies that would result.

This thesis describes some of the research that was undertaken to fill this gap in the methodology of test evaluation.

[4] Cancer is characterised by unregulated cell growth and division.

CHAPTER 3

EVALUATING TEST ACCURACY STUDIES

Diagnosis is a process with multiple stages: the previous test informs the next test (or in some cases, several tests are performed concurrently), until a diagnosis is made; subsequently the test result may or may not influence a clinician's choice of management or treatment option for a patient; and the choice of treatment may or may not impact on the patient's health outcomes.

> *"To make sense of a diagnostic investigation a clinician needs to be able to make an inference regarding the probability that a patient has the disease in question according to the result obtained by the test.*
>
> *Tests rarely make a diagnosis 100% certain, but they provide enough information to rule-in or rule-out a diagnosis in a pragmatic manner. That is, they may make a diagnosis certain enough for the expected benefits of treating the patient to outweigh the expected consequences of not treating them."* (Deeks, J.J. 2001a)

There are multiple reasons for evaluating the performance of a medical test, whether to inform individual patient management decisions or to determine whether the test should be publicly funded by the health system. Specific considerations in test evaluation are:

- to determine the accuracy of a given patient diagnosis;
- to identify more cases (people with the disease) by adding the new test to the existing diagnostic work-up;
- to replace an existing test which is less safe, accurate or cost-effective;
- to decide the order in which tests should be undertaken;
- to decide whether further (invasive) testing for the suspected condition is required, or whether testing for a different condition should occur (known as triage testing);
- to identify more cases at an earlier stage or pre-clinical stage of their disease (known as screening);
- to determine the prognosis of a patient in order to inform treatment planning; or
- to monitor a patient's response to treatment/management (Bossuyt, P.M. et al. 2006).

Research into methods for the evaluation of medical tests has been burgeoning in the last decade (Biesheuvel, Grobbee & Moons 2006; Deeks, J.J. 2001b; Leeflang et al. 2006; Whiting

P 2003), largely initiated by the Cochrane Diagnostic and Screening Test Methods Working Group at the 1999 Cochrane Colloquium meeting. At this meeting in Rome there was a discussion regarding the low methodological quality and substandard reporting of diagnostic test evaluations and, as a consequence, the Standards for Reporting of Diagnostic Accuracy (STARD) initiative was born (Bossuyt, Patrick M. et al. 2003).

The Working Group aimed to develop a checklist of items that would be included in the report of a study of diagnostic accuracy in a peer-reviewed journal. The goal was to improve the accuracy and completeness of reporting of studies of diagnostic accuracy in order to allow readers to assess the internal and external validity of a study. This was analogous to the CONSORT initiative developed to standardise the reporting of randomised controlled trials (Moher, Schulz & Altman 2001).

In a similar endeavour, Whiting et al (2003) developed the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool to assess the methodological quality of diagnostic accuracy studies, and this tool was also revised approximately a decade later (QUADAS-2) (Whiting P 2003; Whiting et al. 2011). As a consequence of improvements in the reporting and appraisal of test accuracy studies, over time there has been a similar improvement in the quality and reporting of systematic reviews of diagnostic tests (AHRQ 2010). Part of this attention to the quality of systematic reviews of test accuracy could be attributed to the decision by the Cochrane Collaboration to undertake systematic reviews of test accuracy. The Cochrane Collaboration Diagnostic and Screening Tests (SRDT) Methods Group was formed and developed different statistical methods for synthesising (meta-analysing) the results of test accuracy studies. The handbook, software and training to support the formal evaluation of diagnostic tests in systematic reviews were released in 2010[5]. The number of meta-analyses of test accuracy studies has increased from fewer than 10 per year in the early 1990s to almost 100 publications per year in recent years (AHRQ 2012).

Test accuracy studies and systematic reviews of these studies do not, however, provide information on the clinical utility of a test; that is, how the test impacts on the health of the patient. The emphasis on test accuracy studies and systematic reviews of test accuracy studies has largely been because of the lack of high quality evidence (ie *direct* randomised controlled trial evidence) to assess the impact of medical tests on patient health outcomes.

[5] http://srdta.cochrane.org/handbook-dta-reviews

Using capture-recapture methodology, Di Ruffano et al (2012) estimated that, on average, only 37 randomised controlled trials are published each year that measure the impact of diagnostic testing on patient outcomes. This compares to approximately 21,949 publications per year of randomised controlled trials that evaluate a treatment. As a consequence,

> *"policy and decision makers frequently need to resort to lower grade evidence, such as decision models to provide guidance on test selection and use"* (di Ruffano et al. 2012).

HTAs of medical tests are generally more complex than HTAs of treatment interventions because the primary research evidence on tests is not straight forward.

Relevant research question

1. *Is the use of the linked evidence approach (LEA) feasible when assessing medical tests for public funding decisions?*

PROBLEM IDENTIFIED AND ADDRESSED IN THE PEER-REVIEWED PUBLICATION

How to assess the first evidence linkage in LEA methodology? That is, how should evidence on test performance or test accuracy be critically appraised?

When a systematic literature review is done, after the collation of all the relevant primary research addressing a technology, usually the first step is to critically appraise the research to determine the impact of bias, confounding and chance on the results obtained (ie internal validity). At the time the research for **Paper 1** (page 33) was undertaken the critical appraisal of evidence was usually done by determining the likelihood that bias had impacted on a study's results because of the way the study was designed.[6] A more comprehensive appraisal was then done by assessing the execution or conduct of the study (Whiting P 2003; Whiting et al. 2011).

In 2009, when Paper 1 was published, the methods for evaluating studies of test accuracy were immature. There was no standard method for assessing the risk of bias

[6] An assessment of the risk of bias is now generally done at the level of each patient-relevant outcome. See section starting on page 64 for a discussion of GRADE methodology.

inherent in the design of test accuracy studies. The most commonly used hierarchy of evidence in Australia was recommended by the National Health and Medical Research Council (NHMRC), in its guidance on conducting systematic reviews of the literature to evaluate medical interventions (NHMRC 1999a). This hierarchy was concerned with evidence demonstrating *therapeutic* effectiveness, rather than *test* effectiveness. Given the lack of published direct evidence to determine diagnostic effectiveness, the objective of Paper 1 was to revise the existing NHMRC hierarchy of evidence so that research studies relating to test accuracy, as well as evidence supporting other types of clinical question, could be critically appraised.

Statement of authorship

Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology*, 2009, 9:34 doi:10.1186/1471-2288-9-34. Available at: http://www.biomedcentral.com/1471-2288/9/34

AUTHORS' CONTRIBUTIONS

Tracy Merlin (Candidate)

Conceived and instigated the revision of the original NHMRC evidence hierarchy, drafted the revised evidence hierarchy and incorporated feedback from co-authors and public submissions, wrote the explanatory notes and glossary, drafted the manuscript and incorporated feedback on the manuscript from co-authors and peer-reviewers.

Adele Weston

Conducted the background review of international frameworks assessing non-randomised or non-interventional evidence (in conjunction with Dr Kristina Coleman and Dr Sarah Norris), contributed to the revised evidence hierarchy and critiqued the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Rebecca Tooher

Contributed to the revised evidence hierarchy and critiqued the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology*, 2009, 9:34 doi:10.1186/1471-2288-9-34. Available at: http://www.biomedcentral.com/1471-2288/9/34

Paper 1 is reproduced as follows -

Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'

Tracy Merlin[1], Adele Weston[2] and Rebecca Tooher[3]

[1]Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, University of Adelaide, Adelaide, South Australia, Australia

[2]Health Technology Analysts, Balmain, New South Wales, Australia

[3]Discipline of Obstetrics and Gynaecology, University of Adelaide, Adelaide, South Australia, Australia

TM* - tracy.merlin@adelaide.edu.au;

AW - aweston@htanalysts.com;

RT - rebecca.tooher@adelaide.edu.au

* Corresponding author

ABSTRACT

Background: In 1999 a four-level hierarchy of evidence was promoted by the National Health and Medical Research Council in Australia. The primary purpose of this hierarchy was to assist with clinical practice guideline development, although it was co-opted for use in systematic literature reviews and health technology assessments. In this hierarchy *interventional* study designs were ranked according to the likelihood that bias had been eliminated and thus it was not ideal to assess studies that addressed other types of clinical questions. This paper reports on the revision and extension of this evidence hierarchy to enable broader use within existing evidence assessment systems.

Methods: A working party identified and assessed empirical evidence, and used a commissioned review of existing evidence assessment schema, to support decision-making regarding revision of the hierarchy. The aim was to retain the existing evidence levels I-IV but increase their relevance for assessing the quality of individual diagnostic accuracy, prognostic, aetiologic and screening studies. Comprehensive public consultation was undertaken and the revised hierarchy was piloted by individual health technology assessment agencies and clinical practice guideline developers. After two and a half years, the hierarchy was again revised and commenced a further 18 month pilot period.

Results: A suitable framework was identified upon which to model the revision. Consistency was maintained in the hierarchy of “levels of evidence” across all types of clinical questions; empirical evidence was used to support the relationship between study design and ranking in the hierarchy wherever possible; and systematic reviews of lower level studies were themselves ascribed a ranking. The impact of ethics on the hierarchy of study designs was acknowledged in the framework, along with a consideration of how harms should be assessed.

Conclusions: The revised evidence hierarchy is now widely used and provides a common standard against which to initially judge the likelihood of bias in individual studies evaluating interventional, diagnostic accuracy, prognostic, aetiologic or screening topics. Detailed quality appraisal of these individual studies, as well as grading of the body of evidence to answer each clinical, research or policy question, can then be undertaken as required.

BACKGROUND

The corner-stone of evidence-based healthcare and health technology assessment is critical appraisal of the evidence underpinning a finding. Different methods are available for assessing the quality of the evidence, including ranking the body of evidence according to a hierarchy which indicates the level of bias associated with the different study designs that have contributed to the evidence-base. In Australia, the standard evidence hierarchy in use since 1999 has been the National Health and Medical Research Council (NHMRC) Designation of Levels of Evidence [1]. This hierarchy ranks the body of evidence into four levels - from systematic reviews of randomised trials at the top of the hierarchy, to case series and case reports at the bottom of the hierarchy (Table 1). Its intended purpose was to summarise the body of evidence for interventions (eg treatment effectiveness). Through widespread use in clinical practice guideline development and health technology assessment, it became increasingly clear that: i) the hierarchy was being used to address research questions that did not relate to interventions; ii) the hierarchy – which is primarily concerned with the association between bias and study design characteristics – was being relied upon for the entire evidence appraisal rather than there being a standardised appraisal of study quality as suggested [2]; and iii) that although the aim was to use the hierarchy to summarise the entire body of evidence – this was occurring rather haphazardly in practice.

TABLE 1 DESIGNATIONS OF LEVELS OF EVIDENCE [1]

Level of evidence	Study design
I	Evidence obtained from a systematic review of all relevant randomised controlled trials
II	Evidence obtained from at least one properly-designed randomised controlled trial
III-1	Evidence obtained from well-designed pseudorandomised controlled trials (alternate allocation or some other method)
III-2	Evidence obtained from comparative studies (including systematic reviews of such studies) with concurrent controls and allocation not randomised, cohort studies, case-control studies, or interrupted time series with a control group
III-3	Evidence obtained from comparative studies with historical control, two or more single arm studies, or interrupted time series without a parallel control group
IV	Evidence obtained from case series, either post-test or pre-test/post-test

This paper describes the first stage of developing a hierarchy to rank the quality of *individual* study designs to address different types of questions. The second stage of developing or adapting a simple, intuitive system to grade the entire *body of evidence* is discussed elsewhere [3, 4], and will be the subject of a forthcoming publication.

The existing hierarchy

The existing NHMRC hierarchy of evidence was developed as part of a comprehensive series of handbooks which outlined the methods for evaluating evidence and developing and disseminating clinical practice guidelines [1, 2, 5-9].

These handbooks recommended that the body of evidence should be assessed along three dimensions: strength, size of effect and clinical relevance. In this schema the strength of evidence was determined by the level of evidence, the quality of the evidence and its statistical precision. It was further assumed that the results from a 'body of evidence' could be distilled down to a single size of effect, with associated statistical precision and that the clinical relevance of this result could be determined eg a pooled relative risk and confidence interval obtained through meta-analysis. The *evidence level*, designated according to the hierarchy (Table 1), assessed the likelihood that the 'body of evidence' producing this single size of effect was affected by bias.

It became clear on applying this schema that the available evidence-base for clinical practice guidelines and health technology assessments was often not amenable to meta-analysis. Thus *statistical synthesis* for each of the outcomes of interest into one estimate of effect, with associated statistical precision and determination of clinical relevance, was often not possible. As a consequence, in practice, the dimensions of evidence were often applied to *individual* studies and were complemented with a *narrative synthesis* of the overall findings from the body of evidence. The difficulty with this approach was that the original hierarchy of evidence was not designed, nor worded, to refer to the strength of the evidence obtained from individual studies.

Further, the hierarchy was designed to assess evidence from intervention studies that evaluated therapeutic effectiveness. It was therefore not appropriate for assessing studies addressing diagnostic accuracy, aetiology, prognosis or screening interventions. The study designs best suited to answer these types of questions are not always the same, or presented in the same order, as that given in the original NHMRC hierarchy of evidence. It was clear that an alternative approach to appraising evidence was needed.

The NHMRC therefore created a working party of clinical practice guideline developers, health technology assessment producers and methodologists (the Working Party) to develop a revised hierarchy of evidence for individual studies (first stage) which addressed these issues, as well as a method for appraising the body of evidence (second stage) that could be used by guideline developers and others.

The objective of the first stage was to create a framework that aligned as closely as possible with the original evidence hierarchy – to minimise confusion for current users and maintain consistency with previous use of the hierarchy – but which could also rank individual studies addressing questions other than therapeutic effectiveness. Due consideration was to be given to methods used by other organisations to develop "levels of evidence", in order to minimise duplication of effort.

METHODS

Recognising the need for an updated hierarchy of evidence, a review was conducted of existing frameworks for assessing non-randomised and non-interventional evidence that are used by Health Technology Assessment (HTA) agencies and guideline developers world-wide [10]. This internal report commissioned by the NHMRC, and conducted by HTAnalysts, provided a resource for the NHMRC and the Working Party to enable revision of the current hierarchy of evidence. The aim was to adapt, if possible, an existing evidence hierarchy or hierarchies.

The report searched for comprehensive evidence frameworks that incorporated non-intervention evidence via HTA and Guideline group websites that were identified through the membership of the International Network of Agencies for Health Technology Assessment (INAHTA) and the Guidelines International Network (GIN) (see Appendix). Bibliographies of identified publications were examined and targeted Medline/ EMBASE searches were conducted. Frameworks were included if they were published in English, were developed by a reputable HTA or guideline agency, and contained guidance on at least one of the methodological processes involved in undertaking an evidence-based assessment (Guideline, HTA or systematic review).

The identified frameworks were then used to inform the revision of the NHMRC evidence hierarchy. Six key factors were considered integral to this revision process, specifically that:

- the hierarchy addressed all types of questions and was not limited to treatment effectiveness alone;
- the levels I-IV were maintained and aligned as closely as possible with the current NHMRC (treatment effectiveness) hierarchy;
- the hierarchy related to individual studies rather than a body of evidence (given a multi-factorial method of “grading” the body of evidence was being developed/adapted concurrently via the NHMRC Working Party);
- the hierarchy remained broadly consistent across types of question;
- empirical evidence supported the placement of a particular study design in the evidence hierarchy wherever possible - that is, the relationship between study design and bias for each clinical or research question had been assessed empirically; or if not, there were good theoretical grounds for such placement in the hierarchy; and
- subjective terms regarding the “quality” of studies eg “well designed”, “properly designed” would be removed. The level of evidence would be assessed on the basis of study design characteristics alone. Determination of the overall “quality” of the study would be independently determined using appropriate - and validated, where possible - checklists suitable for each study design and question.

The “Levels” subgroup of the Working Party addressed each of these criteria while drafting a revision of the evidence hierarchy. This first iteration of the hierarchy was slightly modified after consultation with other methodological experts within the wider Working Party. A second iteration of the hierarchy was presented to Australian and New Zealand evaluators undertaking health technology assessments for the Australian Medical Services Advisory Committee (MSAC). Other international experts on evidence appraisal were contacted and provided feedback on the hierarchy. These suggestions were discussed and some substantial revisions – particularly concerning the diagnostic accuracy evidence hierarchy - were incorporated into a version of the hierarchy that was suitable for piloting.

The hierarchy was piloted by NHMRC clinical practice guideline developers and health technology assessment evaluation groups in Australia and New Zealand from November 2004 until June 2007. Public consultation throughout this period was invited through the medium of international conferences and workshops – specifically the Cochrane Colloquium and the Health Technology Assessment international (HTAi) conference [11-13] - and through the NHMRC website. With the website, a feedback form allowing free text responses to a series of questions regarding the utility and adaptability of the revised hierarchy was provided, along with a section for suggested methods for improving the hierarchy. The hierarchy was amended and a further pilot stage was then conducted from February 2008 to February 2009. In total, approximately a dozen responses were submitted through the website, predominantly by individuals or organisations that had trialled the new evidence hierarchy.

RESULTS

Identifying possible frameworks for adaptation

The 2004 report commissioned by the NHMRC identified 18 evidence frameworks that were relevant for clinical evaluation of non-interventional evidence at that time [10]. Three of the evidence evaluation frameworks were found to use a hierarchy that related to questions other than treatment or intervention effectiveness. The National Institute for Health and Clinical Excellence (NICE) provided a hierarchy that used levels of evidence for assessment of therapeutic effectiveness (based on those developed by the Scottish Intercollegiate Guidelines Network - SIGN) as well as for diagnostic accuracy [14]. The National Health Service Centre for Reviews and Dissemination (NHS CRD) used a framework that included levels of evidence for assessing questions of effectiveness, diagnostic accuracy, and efficiency [15]. Finally, the Centre for Evidence Based Medicine (CEBM) hierarchy included levels of evidence for assessing questions of therapy/prevention and aetiology/harm, prognosis, diagnosis, differential diagnosis/symptom prevalence, and economic and decision analyses [16].

In terms of addressing different types of questions, the CEBM framework was found to be the most comprehensive and a suitable evidence hierarchy upon which to model the revised NHMRC hierarchy of evidence, although all three provided useful information.

The revised NHMRC hierarchy

Each of the six key factors considered integral to a revised NHMRC evidence hierarchy were adopted. Five separate research areas were addressed – interventions, diagnostic accuracy, prognosis, aetiology and screening.

A greatly expanded table was created, largely based on the design of the CEBM framework, which included five separate columns for each of the different research areas (see Additional file 1). However, even though the CEBM layout was very closely followed in the revised NHMRC hierarchy, the number of research questions addressed and description of studies did differ markedly from the CEBM framework. Empirical evidence of study design biases and epidemiological theory were used to rank the study designs within each research area. It was suggested that when referring to studies designated a level of evidence according to the revised NHMRC hierarchy, both the level and corresponding research area or question should be used eg. level II intervention evidence; level IV diagnostic evidence; level III-2 prognostic evidence.

ADDITIONAL FILE 1 NHMRC EVIDENCE HIERARCHY: DESIGNATIONS OF 'LEVELS OF EVIDENCE' ACCORDING TO TYPE OF RESEARCH QUESTION (INCLUDING EXPLANATORY NOTES)

Level	Intervention [1]	Diagnostic accuracy [2]	Prognosis	Aetiology [3]	Screening Intervention
I [4]	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies	A systematic review of level II studies
II	A randomised controlled trial	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among consecutive persons with a defined clinical presentation[6]	A prospective cohort study[7]	A prospective cohort study	A randomised controlled trial
III-1	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)	A study of test accuracy with: an independent, blinded comparison with a valid reference standard,[5] among non-consecutive persons with a defined clinical presentation[6]	All or none[8]	All or none[8]	A pseudorandomised controlled trial (i.e. alternate allocation or some other method)

III-2	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial[9] ▪ Cohort study ▪ Case-control study ▪ Interrupted time series with a control group	A comparison with reference standard that does not meet the criteria required for Level II and III-1 evidence	Analysis of prognostic factors amongst persons in a single arm of a randomised controlled trial	A retrospective cohort study	A comparative study with concurrent controls: ▪ Non-randomised, experimental trial ▪ Cohort study ▪ Case-control study
III-3	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study[10] ▪ Interrupted time series without a parallel control group	Diagnostic case-control study[6]	A retrospective cohort study	A case-control study	A comparative study without concurrent controls: ▪ Historical control study ▪ Two or more single arm study
IV	Case series with either post-test or pre-test/post-test outcomes	Study of diagnostic yield (no reference standard)[11]	Case series, or cohort study of persons at different stages of disease	A cross-sectional study or case series	Case series

EXPLANATORY NOTES

1 Definitions of these study designs are provided on pages 7-8 *How to use the evidence: assessment and application of scientific evidence* (NHMRC 2000b*)* and in the accompanying Glossary.

2 These levels of evidence apply only to studies of assessing the accuracy of diagnostic or screening tests. To assess the overall effectiveness of a diagnostic test there also needs to be a consideration of the impact of the test on patient management and health outcomes (Medical Services Advisory Committee 2005, Sackett and Haynes 2002). The evidence hierarchy given in the 'Intervention' column should be used when assessing the impact of a diagnostic test on health outcomes relative to an existing method of diagnosis/comparator test(s). The evidence hierarchy given in the 'Screening' column should be used when assessing the impact of a screening test on health outcomes relative to no screening or alternative screening methods.

3 If it is possible and/or ethical to determine a causal relationship using experimental evidence, then the 'Intervention' hierarchy of evidence should be utilised. If it is only possible and/or ethical to determine a causal relationship using observational evidence (eg. cannot allocate groups to a potential harmful exposure, such as nuclear radiation), then the 'Aetiology' hierarchy of evidence should be utilised.

4 A systematic review will only be assigned a level of evidence as high as the studies it contains, excepting where those studies are of level II evidence. Systematic reviews of level II evidence provide more data than the individual studies and any meta-analyses will increase the precision of the overall results, reducing the likelihood that the results are affected by chance. Systematic reviews of lower level evidence present results of likely poor internal validity and thus are rated on the likelihood that the results have been affected by bias, rather than whether the systematic review itself is of good quality. Systematic review *quality* should be assessed separately. A systematic review should consist of at least two studies. In systematic reviews that include different study designs, the overall level of evidence should relate to each individual outcome/result, as different studies (and study designs) might contribute to each different outcome.

5 The validity of the reference standard should be determined in the context of the disease under review. Criteria for determining the validity of the reference standard should be pre-specified. This can include the choice of the reference standard(s) and its timing in relation to the index test. The validity of the reference standard can be determined through quality appraisal of the study (Whiting et al 2003).

6 Well-designed population based case-control studies (eg. population based screening studies where test accuracy is assessed on all cases, with a random sample of controls) do capture a population with a representative spectrum of disease and thus fulfil the requirements for a valid assembly of patients. However, in some cases the population assembled is not representative of the use of the test in practice. In diagnostic case-control studies a selected sample of patients already known to have the disease are compared with a separate group of normal/healthy people known to be free of the disease. In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias or spectrum effect because the spectrum of study participants will not be representative of patients seen in practice (Mulherin and Miller 2002).

7 At study inception the cohort is either non-diseased or all at the same stage of the disease. A randomised controlled trial with persons either non-diseased or at the same stage of the disease in *both* arms of the trial would also meet the criterion for this level of evidence.

8 All or none of the people with the risk factor(s) experience the outcome; and the data arises from an unselected or representative case series which provides an unbiased representation of the prognostic effect. For example, no smallpox develops in the absence of the specific virus; and clear proof of the causal link has come from the disappearance of small pox after large-scale vaccination.

9 This also includes controlled before-and-after (pre-test/post-test) studies, as well as adjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C with statistical adjustment for B).

10 Comparing single arm studies ie. case series from two studies. This would also include unadjusted indirect comparisons (ie. utilise A vs B and B vs C, to determine A vs C but where there is no statistical adjustment for B).

11 Studies of diagnostic yield provide the yield of diagnosed patients, as determined by an index test, without confirmation of the accuracy of this diagnosis by a reference standard. These may be the only alternative when there is no reliable reference standard.

Note A: Assessment of comparative harms/safety should occur according to the hierarchy presented for each of the research questions, with the proviso that this assessment occurs within the context of the topic being assessed. Some harms (and other outcomes) are rare and cannot feasibly be captured within randomised controlled trials, in which case lower levels of evidence may be the only type of evidence that is practically achievable; physical harms and psychological harms may need to be addressed by different study designs; harms from diagnostic testing include the likelihood of false positive and false negative results; harms from screening include the likelihood of false alarm and false reassurance results.

Note B: When a level of evidence is attributed in the text of a document, it should also be framed according to its corresponding research question eg. level II intervention evidence; level IV diagnostic evidence; level III-2 prognostic evidence.

Note C: Each individual study that is attributed a “level of evidence” should be rigorously appraised using validated or commonly used checklists or appraisal tools to ensure that factors other than study design have not affected the validity of the results.

Source: Hierarchies adapted and modified from: NHMRC 1999; Bandolier 1999; Lijmer et al. 1999; Phillips et al. 2001 (see Additional File 2).

To support users of the revised NHMRC evidence hierarchy, explanatory notes (see Additional file 1) and a glossary of study designs and terminology (see Additional file 2) were developed and adapted from the NHMRC handbooks [1, 2, 5-9]. The explanatory notes provide the context for the evidence hierarchy, with guidance on how to apply and present the levels of evidence. The glossary provides a definition of each of the given study designs.

DISCUSSION

The revised NHMRC hierarchy of evidence largely addresses the issues which brought about its development. This hierarchy was developed using a combination of evidence, theory and consultation. The Working Party was able to successfully achieve its aim of providing a practical and usable tool for evidence-based healthcare practitioners and researchers. A number of special considerations were addressed in the development of this revised hierarchy, and some limitations were acknowledged when designing the hierarchy.

Limitations

The evidence-base underpinning the development of a hierarchy such as this is limited. For intervention research questions there were some studies and a systematic review showing the degree of bias associated with observational and non-randomised studies, in comparison to randomised controlled trials [17-19]. However, for diagnostic research questions, at the time of developing the hierarchy we were aware of only one study on design-related bias associated with diagnostic studies [20]. In instances where the evidence was lacking to determine placement of the study design in the hierarchy, the CEBM evidence framework was used, along with epidemiology texts [21] and consensus expert opinion.

An evidence hierarchy addressing individual studies, alone, cannot provide interpretation of the results of a ‘body of evidence’ and the various contextual factors that can impinge on the interpretation of results (eg external validity/applicability). The ‘Working Party’ believes that any assessment of evidence underpinning a question involves three steps:

1. determine the level of evidence of individual studies addressing that question and rank the evidence accordingly;
2. appraise the quality of the evidence within each ranking using basic clinical epidemiology and biostatistical principles outlined in widely available critical appraisal checklists and tools; and

3. synthesise the findings from steps 1 and 2 and give greatest weight to the highest quality/highest ranked evidence. After including consideration of contextual factors, make a clear and transparent decision or recommendation regarding the strength and applicability of the findings from the body of evidence, and grade that recommendation.

Steps 1 and 2 are addressed in this paper. Step 3 was undertaken by the NHMRC Working Party through creating a process and system for classifying and grading the body of evidence that takes into account dimensions other than the internal validity of the studies – an issue which has received similar attention in other countries [22, 23]. Progress on other grading systems to date has primarily centred on therapeutic safety and effectiveness research questions [24, 25], although there have been recent moves towards explicitly incorporating diagnostic evidence [26]. The NHMRC Working Party has developed a multi-dimensional system to grade the evidence and develop recommendations in a user-friendly manner but which also addresses various types of research question (through use of this revised NHMRC evidence hierarchy as an intermediary step). This "grading" process is reported elsewhere and will be the subject of a subsequent publication [3, 4].

While the revised hierarchy described in this paper has greatly expanded the types of studies that can be assigned a level of evidence, it does not cover qualitative research or economic analysis. There are existing hierarchies of evidence for economic analysis, although it is unclear if the methodological basis for the ranking within these hierarchies is supported by evidence and theory [15, 16]. Should there be an expressed need to expand the revised NHMRC hierarchy to include economic analysis, this can occur when the NHMRC handbooks are updated.

Methods for synthesising qualitative research evidence are still being developed by groups such as the Cochrane Collaboration [27] and others [28, 29]. In this context, critical appraisal guides and hierarchies of qualitative evidence have begun to appear in the literature [30]. A proper consideration of these issues was beyond the scope of this project and outside the methodological expertise of the Working Party. However, this should be addressed by investigators with appropriate expertise in qualitative research methods as part of the NHMRC handbook updates.

Special considerations

1. Systematic reviews of lower level evidence

In general, the Working Party took the view that systematic reviews should only be assigned a level of evidence as high as the studies contained therein. Even the best quality systematic reviews will still only be able to answer a research question on the basis of the evidence it has collated and synthesised. Thus any overall conclusions will be affected by the internal validity of the primary research evidence included. However, consistent with the original NHMRC hierarchy of evidence, Level I of the revised hierarchy was retained as a systematic review of all relevant level II studies, recognising that meta-analysis of Level II studies can increase the precision of the findings of individual Level II studies [31].

2. Studies of diagnostic test accuracy

The effectiveness of a diagnostic test or a screening test requires either direct evidence ie the impact of the test on patient health outcomes (outlined in the 'Intervention' and 'Screening' columns, respectively, in the revised hierarchy) [26] or, if certain conditions are fulfilled, the linking of evidence of diagnostic test accuracy (assessed using the 'Diagnostic accuracy' column in the hierarchy) with evidence of change in management and the likely effect of that change in management on patient health outcomes (assessed using the 'Intervention' column in the revised hierarchy) [32, 33].

The development of levels of evidence for studies of diagnostic accuracy proved to be more difficult than for the other types of research question. In studies of diagnostic accuracy the basic study design is cross-sectional, in which all participants receive both the index test and the reference standard. In order to rank the validity of each individual study's results it was found that a more specific discussion of study design was required. To aid with the interpretation and ranking of studies comprehensive explanatory notes were developed. To some extent the degree of bias introduced by a particular study design feature is dependent upon both the disease and the diagnostic test context under investigation. Well-developed critical appraisal skills of the reviewers of diagnostic test interventions are therefore essential. Methods for assessing diagnostic test accuracy by systematic review and meta-analysis have been progressing over a relatively short period of time (especially compared with studies of therapeutic effectiveness) [34-37]. As this methodology matures, the descriptive nature of the 'Diagnostic accuracy' levels in the revised hierarchy may no longer be required, as study designs in which bias is minimised are recognised (and possibly even named) as is currently the case with studies of therapeutic effectiveness.

3. Correct classification of the research question

One other difficulty has been noted with use of the evidence hierarchy. The difficulty is not with the study designs or the ranking of the study designs, but rather with distinguishing between an aetiological and prognostic research question – and thus correct use of the relevant hierarchy. Both aetiology and prognosis relate to an identification of risk factors for an outcome and so the relevant study designs are quite similar. The key when determining if a research question is aetiological or prognostic is to identify the population of interest. For prognostic questions, all the population has the condition/disease and the aim is to determine what factors will predict an outcome for that population (eg survival) [2]. For example, "What are the risk factors for suicide in adolescent depression?" These factors can be causal (eg a treatment modality), effect modifiers (eg age) or just associations or markers. For aetiology questions, the key is ensuring the population of interest do not or did not have the condition/disease at some point in time, so that causality of the risk factor can be determined [2]. For example, "What are the risk factors for adolescent depression?" The explanatory notes to the hierarchy cannot make this distinction between aetiology and prognosis completely clear because of the degree of overlap in the relevant study designs.

4. Assessment of study quality

The revised hierarchy of evidence is intended to be used as just one component in determining the strength of the evidence; that is, determining the likelihood of bias from the study design alone. This component is seen as a broad indicator of likely bias and can be used to roughly rank individual studies within a body of evidence. However, study quality within each of the levels of evidence needs to be assessed more rigorously. The Working Party believes that there are so many factors affecting the internal validity of study results (e.g. bias, confounding, results occurring by chance, impact of drop-outs), with different factors affecting different study designs, that a proper assessment of study quality can only occur with the use of an appropriate and/or validated checklist suitable for each study design or research question [2, 15, 25, 37, 38]. In the accompanying documentation to the revised evidence hierarchy, suggestions have been made as to the appropriate checklists for a formal critical appraisal of studies addressing the different types of research question [4].

5. Ethical considerations

The impact of ethics on the hierarchy of study designs was acknowledged in the revised evidence hierarchy. Separate columns for aetiology and intervention research questions were produced in order to address trial feasibility and ethical issues. Explanatory notes appended to the hierarchy indicate that if it is possible and/or ethical to determine a causal relationship using experimental evidence, then the 'Intervention' hierarchy of evidence should be used. However, if it is only possible and/or ethical to determine a causal relationship using observational evidence (for example if it is not ethical to allocate groups to a potentially harmful exposure such as nuclear radiation), then the 'Aetiology' hierarchy of evidence should be used [39, 40]. In the latter scenario, the highest level of evidence that could be used to address the question would be observational and not experimental.

6. Assessment of harms/safety

There is guidance in the explanatory notes about how to deal with the evaluation of comparative harms and safety in the research area of interest. Assessment of comparative harms/safety should occur according to the hierarchy presented for each of the research questions, with the proviso that this assessment occurs within the context of the topic being assessed. Some harms (as well as some effectiveness outcomes) are rare and cannot feasibly be captured within randomised controlled trials [41, 42] , in which case lower levels of evidence may be the only type of evidence that is practically achievable; physical harms and psychological harms may need to be addressed by different study designs [43]; harms from diagnostic testing include the likelihood of false positive and false negative results [44, 45]; harms from screening include the likelihood of false alarm and false reassurance results [46].

No single evidence-framework can address all of the safety and effectiveness issues associated with different research areas. The aim of the explanatory note was to explicitly recognise that these differences will occur and to adapt the hierarchy where necessary.

CONCLUSIONS

Given the extensive pilot process – four years – this new evidence hierarchy is now the standard for judging “levels of evidence” for the purposes of health technology assessment and clinical practice guideline development in Australia.

Although this broad ranking tool for assessing study quality is intended for use as an intermediary step within the new NHMRC system to grade the body of evidence addressing a clinical, research or policy question [4], it can be applied within existing grading systems eg GRADE [47], SIGN [25] with the benefit of allowing a ranking of evidence that addresses research questions or areas other than therapeutic effectiveness.

This tool is particularly advantageous for structuring a narrative meta-synthesis of results in an evidence report or health technology assessment. Studies and study results can initially be ranked by study design (evidence level) using the revised evidence hierarchy, and then be further ranked *within* each evidence level with the use of appropriate and validated quality appraisal checklists. A grading of the body of evidence can then be applied, if relevant.

COMPETING INTERESTS

Meeting attendance fees for this methodological work were paid to the authors by the National Health and Medical Research Council (NHMRC), a not-for-profit research organisation funded by the Australian Government Department of Health and Ageing. One of the functions of the NHMRC is to develop and disseminate health publications to health professionals and consumers in Australia. They produce health advisories, evidence-based clinical practice guidelines and methodology publications.

AUTHORS' CONTRIBUTIONS

TM instigated the revision of the original NHMRC evidence hierarchy, co-developed the revised evidence hierarchy, wrote the explanatory notes and glossary, drafted the manuscript, and incorporated the feedback received on both the hierarchy and the manuscript. AW conducted the review of international frameworks assessing non-randomised or non-interventional evidence (in conjunction with Dr Kristina Coleman and Dr Sarah Norris), co-developed the revised evidence hierarchy, and contributed to the development of the manuscript. RT co-developed the revised evidence hierarchy and contributed to the development of the manuscript. All authors read and approved the final manuscript.

APPENDIX

Searches were conducted in June 2004. Enquiries regarding the search strategies should be directed to the Evidence Translation Section, National Health and Medical Research Council, Canberra, ACT, Australia.

ACKNOWLEDGEMENTS

We would like to thank the other members of the NHMRC Guideline Assessment Working Party for their input – specifically, Kristina Coleman, Sarah Norris, Karen Grimmer-Somers, Susan Hillier, Philippa Middleton, and Janet Salisbury. We would also like to acknowledge the efforts of Janine Keough and Chris Gonzalez, formerly of the NHMRC Health Advisory Section. We appreciated the feedback provided to the NHMRC by Paul Glasziou, Brian Haynes, Andrew Oxman, Nicki Jackson and those who submitted suggestions via the website. The suggestions provided by Sally Lord and Les Irwig were particularly helpful. We would also like to thank Janet Hiller and Phil Ryan for providing internal peer review and Mike Clarke and Myfanwy Jones for their very useful suggestions during external peer review of the manuscript.

REFERENCES

1. NHMRC: A guide to the development, implementation and evaluation of clinical practice guidelines. Canberra, ACT: National Health and Medical Research Council, Commonwealth of Australia; 1999.

2. NHMRC: How to review the evidence: systematic identification and review of the scientific literature. Canberra: National Health and Medical Research Council; 2000.

3. Middleton P, Tooher R, Salisbury J, Coleman K, Norris S, Grimmer K, Hillier S: Assessing the body of evidence and grading recommendations in evidence-based clinical practice guidelines. In: *Corroboree: Melbourne. XIII Cochrane Colloquium, 22-26 October 2005;* Melbourne: Australasian Cochrane Centre; 2005.

4. NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. Stage 2 consultation. Early 2008 - end June 2009 [Available from: http://www.nhmrc.gov.au/guidelines/consult/consultations/add_levels_grades_dev_guidelines2.htm]

5. NHMRC: How to present the evidence for consumers: preparation of consumer publications. Canberra: National Health and Medical Research Council; 1999.

6. NHMRC: How to use the evidence: assessment and application of scientific evidence. Canberra: National Health and Medical Research Council; 2000.

7. NHMRC: How to put the evidence into practice: implementation and dissemination strategies. Canberra: National Health and Medical Research Council; 2000.

8. NHMRC: How to compare the costs and benefits: evaluation of the economic evidence. Canberra: National Health and Medical Research Council; 2001.

9. NHMRC: Using socioeconomic evidence in clinical practice guidelines. Canberra, ACT: Commonwealth of Australia; 2003.

10. Coleman K, Standfield L, Weston A: The utilisation of established frameworks in assessing and applying non-intervention/non-randomised evidence [Internal report]. Canberra, ACT: Health Advisory Committee, National Health and Medical Research Council (NHMRC); 2004.

11. Merlin T, Weston A, Tooher R: Re-assessing and revising “levels of evidence” in the critical appraisal process. In: *Corroboree: Melbourne. XIII Cochrane Colloquium, 22-26 October 2005;* Melbourne: Australasian Cochrane Centre; 2005.

12. Merlin T, Weston A, Tooher R: Revising a national standard: redevelopment of the Australian NHMRC evidence hierarchy. *Italian Journal of Public Health (Supplement 1)* 2005, 2(2):156.

13. Merlin T, Middleton P, Salisbury J, Weston A: Ways to ensure evidence-based clinical practice guidelines are of high quality. In: *Corroboree: Melbourne. XIII Cochrane Colloquium, 22-26 October 2005;* Melbourne: Australasian Cochrane Centre; 2005.

14. National Institute for Health and Clinical Excellence: The guidelines manual. London: National Institute for Health and Clinical Excellence; 2007.

15. Khan KS, Ter Riet G, Glanville JM, Sowden AJ, Kleijnen J: 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; 2001.

16. Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M: Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001). Oxford: Centre for Evidence-Based Medicine; 2001.

17. Benson K, Hartz AJ: A comparison of observational studies and randomized, controlled trials. *N Engl J Med* 2000, 342(25):1878-1886.

18. Kunz R, Oxman AD: The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. *British Medical Journal (Education and Debate)* 1998, 317(7167):1185-1190.

19. Concato J, Shah N, Horwitz RI: Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000, 342(25):1887-1892.

20. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, Bossuyt PM: Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association* 1999, 282(11):1061 - 1066.

21. Elwood JM: Critical appraisal of epidemiological studies and clinical trials, Second edn. Oxford: Oxford University Press; 1998.

22. The GRADE working group: Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches. *BMC Health Serv Res* 2004, 4(1):38.

23. Bellomo R, Bagshaw SM: Evidence-based medicine: classifying the evidence from clinical trials - the need to consider other dimensions. *Critical Care* 2006, 10:232.

24. The GRADE working group: Systems for grading the quality of evidence and the strength of recommendations II: A pilot study of a new system for grading the quality of evidence and the strength of recommendations. *BMC Health Serv Res* 2005, 5(1):25.

25. Scottish Intercollegiate Guidelines' Network (SIGN): SIGN 50: a guideline developer's handbook. Edinburgh: SIGN; 2008.

26. Schunemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, Williams JW, Kunz R, Craig J, Montori VM *et al*: Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ* 2008, 336:1106-1110.

27. Noyes J, Popay J, Pearson A, Hannes K, Booth A: Chapter 20: Qualitative research and Cochrane reviews. In: *Cochrane Handbook of Systematic Reviews of Interventions, Version 501.* Edited by Higgins J, Green S, Version 5.0.1 (updated September 2008) edn: The Cochrane Collaboration; 2008. [Available from: http://www.cochrane-handbook.org]

28. Popay J (ed.): Moving beyond effectiveness in evidence synthesis: methodological issues in the synthesis of diverse sources of evidence. London: National Institute for Health and Clinical Excellence; 2006.

29. Denny E, Khan KS: Systematic reviews of qualitative evidence: What are the experiences of women with endometriosis? *J Obstet Gynaecol* 2006, 26(6):501-506.

30. Daly J, Willis K, Small R, Green J, Welch N, Kealy M, Hughes E: A hierarchy of evidence for assessing qualitative health research. *J Clin Epidemiol* 2007, 60:43-49.

31. Egger M, Ebrahim J, Davey Smith G: Where now for metaanalysis? *Int J Epidemiol* 2002, 31:1-5.

32. Medical Services Advisory Committee: Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia; 2005.

33. Sackett DL, Haynes RB: The architecture of diagnostic research. *BMJ* 2002, 324:539-541.

34. Deeks JJ: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001, 323(21 July):157-162.

35. Harbord R, Bachmann L, Shang A, Whiting P, Deeks J, Egger M, Sterne J: An empirical comparison of methods for meta-analysis of studies of diagnostic accuracy. In: *Corroboree: Melbourne. XIII Cochrane Colloquium, 22-26 October 2005;* Melbourne: Australasian Cochrane Centre; 2005.

36. Mallett S, Deeks J, Halligan S, Hopewell S, Cornelius V, Altman D: Systematic review of diagnostic tests in cancer: review of methods and reporting. *BMJ* 2006, 333:413.

37. Whiting P RA, Reitsma JB, Bossuyt PM, Kleijnen J.: The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003, 3(1):25.

38. Downs SH, Black N: The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998, 52(6):377-384.

39. Edward SJ, Stevens AJ, Braunholtz DA, Lilford RJ, Swift T: The ethics of placebo-controlled trials: a comparison of inert and active placebo controls. *World J Surg* 2005, 29(5):610-614.

40. Black N: Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 1996, 312:1215-1218.

41. Eikelboom JW, Mehta SR, Pogue J, Yusuf S: Safety outcomes in meta-analyses of phase 2 vs phase 3 randomized trials: Intracranial hemorrhage in trials of bolus thrombolytic therapy. *JAMA* 2001, 285(4):444-450.

42. Lancet Editorial: Opren scandal. *Lancet* 1983, 1:219-220.

43. Scaf-Klomp W, Sanderman R, van de Wiel HB, Otter R, van den Heuvel WJ: Distressed or relieved? Psychological side effects of breast cancer screening in The Netherlands. *J Epidemiol Community Health* 1997, 51(6):705-710.

44. Jackson BR: The dangers of false-positive and false-negative test results: false-positive results as a function of pretest probability. *Clin Lab Med* 2008, 28(2):305-319, vii.

45. Leung GM, Woo PP, Cowling BJ, Tsang CS, Cheung AN, Ngan HY, Galbraith K, Lam TH: Who receives, benefits from and is harmed by cervical and breast cancer screening among Hong Kong Chinese? *J Public Health (Oxf)* 2008, 30(3):282-292. Epub 2008 May 2014.

46. Shickle D, Chadwick R: The ethics of screening: is 'screeningitis' an incurable disease? *J Med Ethics* 1994, 20(1):12-18.

47. The GRADE working group: GRADE: what is quality of evidence and why is it important to clinicians? *BMJ* 2008, 336:995-998.

ADDITIONAL FILE 2

STUDY DESIGN GLOSSARY (ALPHABETIC ORDER)

Adapted from NHMRC 2000ab, Glasziou et al 2001, Elwood 1998

Note: This is a specialised glossary that relates specifically to the study designs mentioned in the NHMRC Evidence Hierarchy. Glossaries of terms that relate to wider epidemiological concepts and evidence based medicine are also available – see *http://www.inahta.org/HTA/Glossary/*; *http://www.ebmny.org/glossary.html*

All or none – all or none of a series of people (case series) with the risk factor(s) experience the outcome. The data should relate to an unselected or representative case series which provides an unbiased representation of the prognostic effect. For example, no smallpox develops in the absence of the specific virus; and clear proof of the causal link has come from the disappearance of small pox after large scale vaccination. This is a rare situation.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among consecutive patients with a defined clinical presentation – a cross-sectional study where a consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

A study of test accuracy with: an independent, blinded comparison with a valid reference standard, among non-consecutive patients with a defined clinical presentation – a cross-sectional study where a non-consecutive group of people from an appropriate (relevant) population receive the test under study (index test) and the reference standard test. The index test result is not incorporated in (is independent of) the reference test result/final diagnosis. The assessor determining the results of the index test is blinded to the results of the reference standard test and vice versa.

Adjusted indirect comparisons – an adjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C when there is statistical adjustment for B. This is most commonly done in meta-analyses (see Bucher et al 1997). Such an indirect comparison should only be attempted when the study populations, common comparator/reference, and settings are very similar in the two studies (Song et al 2000).

Case-control study – people with the outcome or disease (cases) and an appropriate group of controls without the outcome or disease (controls) are selected and information obtained about their previous exposure/non-exposure to the intervention or factor under study.

Case series – a single group of people exposed to the intervention (factor under study).

Post-test – only outcomes after the intervention (factor under study) are recorded in the series of people, so no comparisons can be made.

Pre-test/post-test – measures on an outcome are taken before and after the intervention is introduced to a series of people and are then compared (also known as a 'before- and-after study').

Cohort study – outcomes for groups of people observed to be exposed to an intervention, or the factor under study, are compared to outcomes for groups of people not exposed.

Prospective cohort study – where groups of people (cohorts) are observed at a point in time to be exposed or not exposed to an intervention (or the factor under study) and then are followed prospectively with further outcomes recorded as they happen.

Retrospective cohort study – where the cohorts (groups of people exposed and not exposed) are defined at a point of time in the past and information collected on subsequent outcomes, eg. the use of medical records to identify a group of women using oral contraceptives five years ago, and a group of women not using oral contraceptives, and then contacting these women or identifying in subsequent medical records the development of deep vein thrombosis.

Cross-sectional study – a group of people are assessed at a particular point (or cross-section) in time and the data collected on outcomes relate to that point in time ie proportion of people with asthma in October 2004. This type of study is useful for hypothesis-generation, to identify whether a risk factor is associated with a certain type of outcome, but more often than not (except when the exposure and outcome are stable eg. genetic mutation and certain clinical symptoms) the causal link cannot be proven unless a time dimension is included.

Diagnostic (test) accuracy – in diagnostic accuracy studies, the outcomes from one or more diagnostic tests under evaluation (the *index test*/s) are compared with outcomes from a *reference standard test*. These outcomes are measured in individuals who are suspected of having the condition of interest. The term *accuracy* refers to the amount of agreement between the index test and the reference standard test in terms of outcome measurement. Diagnostic accuracy can be expressed in many ways, including sensitivity and specificity, likelihood ratios, diagnostic odds ratio, and the area under a receiver operator characteristic (ROC) curve (Bossuyt et al 2003)

Diagnostic case-control study – the index test results for a group of patients already known to have the disease (through the reference standard) are compared to the index test results with a separate group of normal/healthy people known to be free of the disease (through the use of the reference standard). In this situation patients with borderline or mild expressions of the disease, and conditions mimicking the disease are excluded, which can lead to exaggeration of both sensitivity and specificity. This is called spectrum bias because the spectrum of study participants will not be representative of patients seen in practice. *Note: this does not apply to well-designed population based case-control studies.*

Historical control study – outcomes for a prospectively collected group of people exposed to the intervention (factor under study) are compared with either (1) the outcomes of people treated at the same institution prior to the introduction of the intervention (ie. control group/usual care), or (2) the outcomes of a previously published series of people undergoing the alternate or control intervention.

Interrupted time series with a control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and then compared to the outcomes at the same time points for a group of people that do not receive the intervention (factor under study).

Interrupted time series without a parallel control group – trends in an outcome or disease are measured over multiple time points before and after the intervention (factor under study) is introduced to a group of people, and compared (as opposed to being compared to an external control group).

Non-randomised, experimental trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention group or a control group, using a non-random method (such as patient or clinician preference/availability) and the outcomes from each group are compared.

This can include:

(1) **a controlled before-and-after study**, where outcome measurements are taken before and after the intervention is introduced, and compared at the same time point to outcome measures in the control group.

(2) **an adjusted indirect comparison**, where two randomised controlled trials compare different interventions to the same comparator ie. the placebo or control condition. The outcomes from the two interventions are then compared indirectly. *See entry on adjusted indirect comparisons.*

Pseudo-randomised controlled trial - the unit of experimentation (eg. people, a cluster of people) is allocated to either an intervention (the factor under study) group or a control group, using a pseudo-random method (such as alternate allocation, allocation by days of the week or odd-even study numbers) and the outcomes from each group are compared.

Randomised controlled trial – the unit of experimentation (eg. people, or a cluster of people[7]) is allocated to either an intervention (the factor under study) group or a control group, using a random mechanism (such as a coin toss, random number table, computer-generated random numbers) and the outcomes from each group are compared. Cross-over randomised controlled trials – where the people in the trial receive one intervention and then cross-over to receive the alternate intervention at a point in time – are considered to be the same level of evidence as a randomised controlled trial, although appraisal of these trials would need to be tailored to address the risk of bias specific to cross-over trials.

Reference standard - the reference standard is considered to be the best available method for establishing the presence or absence of the target condition of interest. The reference standard can be a single diagnostic method, or a combination of methods. It can include laboratory tests, imaging tests, and pathology, but also dedicated clinical follow-up of individuals (Bossuyt et al 2003).

Screening intervention – a screening intervention is a public health service in which members of a defined population, who do not necessarily perceive that they are at risk of, or are already affected by a disease or its complications (asymptomatic), are asked a question or offered a test, to identify those individuals who are more likely to be helped than harmed by further tests or treatment to reduce the risk of a disease or its complications (UK National Screening Committee, 2007). A screening intervention study compares the implementation of the screening intervention in an asymptomatic population with a control group where the screening intervention is not employed or where a different screening intervention is

[7] Known as a cluster randomised controlled trial

employed. The aim is to see whether the screening intervention of interest results in improvements in patient-relevant outcomes eg survival.

Study of diagnostic yield – these studies provide the yield of diagnosed patients, as determined by the index test, without confirmation of the accuracy of the diagnosis (ie. whether the patient is actually diseased) by a reference standard test.

Systematic review – systematic location, appraisal and synthesis of evidence from scientific studies.

Test - any method of obtaining additional information on a person's health status. It includes information from history and physical examination, laboratory tests, imaging tests, function tests, and histopathology (Bossuyt et al 2003).

Two or more single arm study – the outcomes of a single series of people receiving an intervention (case series) from two or more studies are compared. *Also see entry on unadjusted indirect comparisons.*

Unadjusted indirect comparisons – an unadjusted indirect comparison compares single arms from two or more interventions from two or more separate studies via the use of a common reference ie A versus B and B versus C allows a comparison of A versus C but there is no statistical adjustment for the common reference (B). Such a simple indirect comparison is unlikely to be reliable (see Song et al 2000).

REFERENCES RELATING TO EXPLANATORY NOTES AND GLOSSARY

Bandolier editorial. Diagnostic testing emerging from the gloom? *Bandolier*, 1999;70:70-5. Available at: http://www.jr2.ox.ac.uk/bandolier/band70/b70-5.html

Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Lijmer JG, Moher D, Rennie D, de Vet HCW for the STARD Group. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *AJR*, 2003; 181:51-56

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *J Clin Epidemiol*, 1997;50:683-91.

Elwood M. (1998) *Critical appraisal of epidemiological studies and clinical trials*. Second edition. Oxford: Oxford University Press.

Glasziou P, Irwig L, Bain C, Colditz G. (2001) *Systematic reviews in health care. A practical guide.* Cambridge: Cambridge University Press.

Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JHP, Bossuyt PMM. Empirical evidence of design-related bias in studies of diagnostic tests. *JAMA*, 1999; 282(11):1061-6.

Medical Services Advisory Committee (2005). *Guidelines for the assessment of diagnostic technologies*. [Internet] Available at: www.msac.gov.au

Mulherin S, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. *Ann Intern Med*, 2002;137:598-602.

NHMRC (1999). A guide to the development, implementation and evaluation of clinical practice guidelines. Canberra: National Health and Medical Research Council.

NHMRC (2000a). How to review the evidence: systematic identification and review of the scientific literature. Canberra: National Health and Medical Research Council.

NHMRC (2000b). How to use the evidence: assessment and application of scientific evidence. Canberra: National Health and Medical Research Council.

Phillips B, Ball C, Sackett D, Badenoch D, Straus S, Haynes B, Dawes M (2001). *Oxford Centre for Evidence-Based Medicine levels of evidence (May 2001)*. Oxford: Centre for Evidence-Based Medicine. Available at: http://www.cebm.net/levels_of_evidence.asp

Sackett DL, Haynes RB. The architecture of diagnostic research. *BMJ*, 2002;324:539-41.

Song F, Glenny A-M, Altman DG. Indirect comparison in evaluating relative efficacy illustrated by antimicrobial prophylaxis in colorectal surgery. *Controlled Clinical Trials*, 2000;21(5):488-497.

UK National Screening Committee (2000). The UK National Screening Committee's criteria for appraising the viability, effectiveness and appropriateness of a screening programme. In: Second Report of the UK National Screening Committee. London: United Kingdom Departments of Health. Pp. 26-27. Available at: http://www.nsc.nhs.uk/

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003; 3(1): 25. Available at:
http://www.biomedcentral.com/1471-2288/3/25

Relevance of Paper 1 to the thesis

The original (1999) NHMRC hierarchy of evidence provided the Australian standard for the assessment of design-related bias in studies included in systematic reviews of the literature. The hierarchy was only relevant for use in the critical appraisal of medical tests if the evidence being appraised was derived from clinical trials measuring the health impact of testing (direct evidence of diagnostic effectiveness). Use of the NHMRC hierarchy was mandated by the Australian Government for the assessment of medical services being reviewed by the Medical Services Advisory Committee (MSAC) for a public funding decision (MSAC 2000a).

As clinical trials assessing the effectiveness of medical tests are rare, it became increasingly apparent that if MSAC only relied on direct evidence for test evaluation then it would be unlikely that any tests would have an evidence base sufficient to warrant a positive public funding decision.

Paper 1 (reproduced above) describes the revision of the 1999 NHMRC hierarchy of evidence. The revised evidence hierarchy recognised that the performance of a test and its impact on patient health outcomes can be measured through the use of direct evidence (ie clinical trials reporting on health outcomes) or through a linked evidence approach (see Chapter 4) that incorporates evidence from test accuracy studies. An 'intervention' column was included in the revised hierarchy to address direct evidence of diagnostic effectiveness while a 'diagnostic accuracy' column was included so that linked evidence could also be used and appraised. This latter column ranked the design of test accuracy studies according to the probability that bias will affect the test performance results (eg sensitivity, specificity). The inclusion of this column in the evidence hierarchy contributed to the proper evaluation of the risk of bias of those test accuracy studies included in the health technology assessments (HTAs) supplied to MSAC. The revised evidence hierarchy also became the new Australian standard for appraising the risk of design-related bias in studies included in systematic reviews of the literature.

In terms of addressing the research questions for this thesis, the revised evidence hierarchy was an *enabler* for the use of the linked evidence approach methodology when evaluating medical tests for reimbursement decisions. The evidence base needed to be appraised properly in order to determine whether the test performance results could be relied upon. When writing a systematic review it is helpful to rank the results of the included studies according to the likelihood that bias has affected the results. This is particularly helpful when a meta-analysis cannot be undertaken and the results need to be synthesised into a narrative and communicated to the policy maker for a decision regarding the clinical effectiveness and cost-effectiveness of the test.

In the context of Australian HTA, the revised evidence hierarchy ensured that one tool could be used to assess study design-related bias when both direct and linked evidence was used in the same HTA of a medical test. The hierarchy was always intended to be used in conjunction with a more finely grained critical appraisal instrument (eg QUADAS or QUADAS—2 for test accuracy studies (Whiting P 2003; Whiting et al. 2011)) as there are numerous ways that bias can be introduced into a study or affect the findings of a study – an evidence hierarchy would be unable to capture them all.

This inability to capture all aspects of bias in an evidence hierarchy has been recognised in recent years. The biases associated with the design and execution of research has been considered as a whole, rather than relying on evidence hierarchies alone or on complementary approaches of evidence hierarchies and critical appraisal tools (Sanderson, Tatt & Higgins 2007).

The most commonly used approach to determine the strength of a *body of evidence* in the field of clinical practice guideline development, is known as GRADE (Brozek, Akl, Alonso-Coello, et al. 2009; Guyatt et al. 2008). It has, however, had limited[8] traction in the health technology assessment field where evidence for the safety and effectiveness of new and emerging technologies is often limited and of poor quality. GRADE was developed at around the same time as the NHMRC evidence hierarchy was developed and it is distinguished by assessing bias on a 'per outcome' basis, rather than a 'per study' basis. It also allows clinical expert opinion to inform all judgements of bias, and to downgrade or upgrade quality appraisals on the basis of expert opinion. Although guidance has been produced in recent times on the application of GRADE to test accuracy studies (Brozek, Akl, Jaeschke, et al. 2009; Schunemann, H. J. et al. 2008), there has been minimal recognition to date as to how the system can be used to estimate test effectiveness ie to quantify the clinical utility, or impact of a test on patient health outcomes. Similarly, Cochrane systematic reviews of test

[8] Although its use is increasing in Europe eg the World Health Organization

accuracy do not yet include a method for predicting the clinical utility of a test (Gopalakrishna et al. 2014).

Next, Chapter 4 describes how the clinical utility of a test is estimated in Australian health technology assessments using the linked evidence approach.

CHAPTER 4

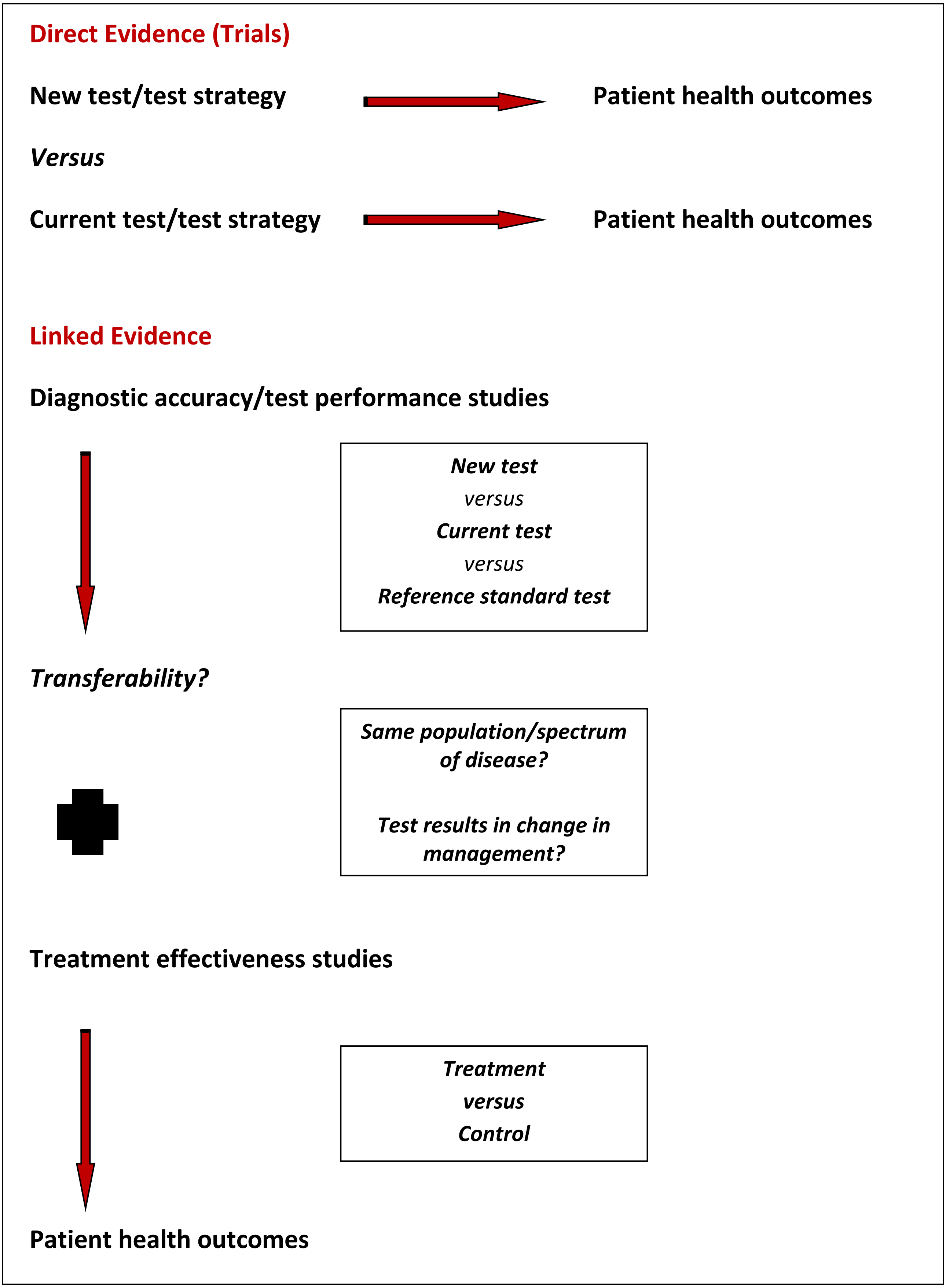
FEASIBILITY OF THE LINKED EVIDENCE APPROACH AND THE IMPACT ON POLICY

In 2005 the Medical Services Advisory Committee (MSAC) in Australia commissioned a Guideline for the assessment and evaluation of diagnostic tests (MSAC 2005a). This work was influenced by Fryback and Thornbury's 6-tiered hierarchical model of efficacy as it relates to diagnostic imaging (Fryback & Thornbury 1991). The hierarchy outlines several levels of efficacy including technical efficacy, diagnostic accuracy, diagnostic thinking (change in diagnosis), therapeutic efficacy (change in management) and patient outcome efficacy (change in health outcomes). The MSAC Guideline proposed that medical tests could be evaluated for clinical utility by linking evidence addressing each of these efficacy levels under certain conditions. This *linking* of evidence would occur in instances where direct trial evidence of test effectiveness (impact of the test on patient health outcomes) was not available (MSAC 2005a).

The MSAC Guideline also suggested that evidence of test accuracy or test performance, alone, may be a sufficient predictor of test effectiveness if there is reasonable justification to assume that the population receiving the test (and within which its accuracy has been tested) is to all intents and purposes the same population that would receive treatment for the condition – and there is good evidence that treatment impacts positively on the health outcomes in this population. This was the **transferability** assumption. If the new test resulted in additional cases being detected, and thus the spectrum of disease in the diagnosed population changed, then evidence of the effectiveness of treatment options in this broader population (via a systematic literature review) would need to be determined. If these data were unavailable, then a *linked* evidence approach could not be undertaken (MSAC 2005a).

The use of *direct* evidence compared to *linked* evidence in the evaluation of medical tests in is illustrated in Box 1.

BOX 1 APPROACHES TO THE EVALUATION OF MEDICAL TESTS IN HEALTH TECHNOLOGY ASSESSMENT



Direct Evidence (Trials)

New test/test strategy → **Patient health outcomes**

Versus

Current test/test strategy → **Patient health outcomes**

Linked Evidence

Diagnostic accuracy/test performance studies

New test
versus
Current test
versus
Reference standard test

Transferability?

Same population/spectrum of disease?

Test results in change in management?

+

Treatment effectiveness studies

Treatment
versus
Control

Patient health outcomes

With the publication of the MSAC Guidelines, a novel approach was available for the assessment of medical tests in a health technology assessment context – including estimating the effect of these tests on patient health outcomes in the face of limited evidence. Although this Guidance was produced to inform the approaches taken by health technology assessment agencies in Australia when evaluating medical tests, there was no subsequent evaluation of the impact of this Guidance. It was unclear whether HTA practices changed or whether these hypothesised changes affected policy decisions concerning test reimbursement.

Relevant research questions

1. *Is the use of the linked evidence approach (LEA) feasible when assessing medical tests for public funding decisions?*
2. *What effect (if any) has LEA methodology had on Australian policy makers' decisions to publicly fund medical tests?*

PROBLEM IDENTIFIED AND ADDRESSED

> *Was there a change in the approach to test evaluation in Australian HTAs after the introduction of the Medical Services Advisory Committee (MSAC) diagnostic guidelines in 2005? And, if so, has the use of LEA methodology affected MSAC decision-making concerning the public funding of medical tests?*

Paper 2 (page 75) is a before-and-after study of all of the HTA reports submitted to MSAC for a public funding decision between 1999 and 2014. The focus of the study was to document the various approaches used to evaluate medical tests. It was hypothesised that MSAC's Guidance on the use of LEA would change the way that tests were evaluated for their clinical utility.

It was also hypothesised that the change in methodological approach might provide more certainty in decision-making, in that there would be an increase in definitive positive or negative public funding recommendations because of the increased quantity and coherence (prediction of clinical utility) of the information provided to decision-makers.

Statement of authorship

Merlin T, Ryan P, Hiller JE. Impact of the 'linked evidence approach' method on policies to publicly fund diagnostic, staging and screening medical tests. *Medical Decision Making* [under review – submitted April 2015]

AUTHORS' CONTRIBUTIONS

Tracy Merlin (Candidate)

Conceived and instigated the research project, wrote the study protocol and analysis plan, undertook the data extraction, statistical analysis and interpreted the results. Drafted the manuscript and incorporated feedback on the manuscript from co-authors.

Philip Ryan

Provided statistical expertise regarding the analysis plan and provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Janet Hiller

Provided suggestions regarding presentation of the results and provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Merlin T, Ryan P, Hiller JE. Impact of the ‘linked evidence approach’ method on policies to publicly fund diagnostic, staging and screening medical tests. *Medical Decision Making* [under review – submitted May 2015]

The pre-publication version of **Paper 2** is reproduced as follows -

Impact of the 'linked evidence approach' method on policies to publicly fund diagnostic, staging and screening medical tests.

Running head: Impact of test evaluation method on funding policy

Tracy L Merlin[1] MPH, Janet E Hiller[2,3] PhD and Philip Ryan[3] FAFPHM

[1]Adelaide Health Technology Assessment (AHTA), School of Public Health, University of Adelaide, Adelaide, South Australia, Australia

[2] Faculty of Health Sciences, Swinburne University, Melbourne, Victoria, Australia

[3] School of Public Health, University of Adelaide, Adelaide, South Australia, Australia

Corresponding author contact details:

A/Prof Tracy Merlin, Director, Adelaide Health Technology Assessment (AHTA)

School of Public Health, Mail Drop DX 650 545

University of Adelaide, South Australia, Australia 5005

Email: tracy.merlin@adelaide.edu.au

Word count: abstract – 300 including headings; text - 3618, excluding abstract, references, figures and tables.

Key words: Diagnostic Test Approval, Evaluation Methodology, Systematic Review, Reimbursement Mechanisms, Decision-making, Policy

ABSTRACT

Background: The Linked Evidence Approach (LEA) is used in health technology assessment (HTA) to evaluate the clinical utility of new medical tests in the absence of direct trial evidence. Our goal was to determine whether use of the method affects the type of public funding decision.

Methods: All HTAs on technologies submitted in 1999-2014 for funding under Australian Medicare were screened for inclusion. Data were extracted from eligible HTAs produced before and after LEA was mandated in 2005. Logistic regression analyses were undertaken to determine the impact of LEA and possible clinical predictors (selected *a priori*) on funding decisions. Regression diagnostics were performed to estimate model fit, model specification and to inform model selection. The unit of analysis was per clinical indication for each new test.

Results: 83 evaluations of medical tests (for 173 clinical indications) were eligible from the 259 HTAs available. When health policy was compared before and after 2005, there was an 11% reduction in overall positive funding decisions, including a 25% decrease in 'interim' funding decisions. The odds of obtaining interim funding reduced by 98% (OR=0.02, 95%CI 0.0005, 0.17, χ^2=26.44, p<0.001) but there was no change in the direction of funding decisions (OR=1.36, 95%CI 0.62, 3.01, χ^2=0.69, p=0.406). Across both time periods, when LEA was used there was a very strong likelihood that the medical test would *not* receive interim funding (X^2=12.63, df=1, p=0.0004). For all positive funding decisions and all *new* positive funding decisions, the strongest predictors were whether or not the test would replace an existing test and whether the available evidence was limited (χ^2=6.63, df=2, p=0.036; X^2=7.22, df=2, p=0.027, respectively).

Conclusions: The use of LEA did not predict the direction of reimbursement decisions. The method did predict that a 'coverage with evidence development' decision was unlikely. This suggests that LEA may reduce decision-maker uncertainty.

ACKNOWLEDGEMENTS

Many thanks to Samuel Lehman for conducting duplicate data extraction for 59 of the 173 clinical indications reviewed in this study.

Financial disclosure: The salary of the primary author (TM) was funded by the University of Adelaide. PR and JH are titleholders of the University but received no funding for their role in this study. No external funding for this study was received. As such, the authors' were independent in designing the study, interpreting the data, writing, and publishing the report. TM holds an existing contract with the Australian Government Department of Health to conduct health technology assessments of medical tests.

INTRODUCTION

In the past, evaluations of medical tests for public funding decisions have been largely restricted to assessments of test accuracy or performance with little consideration given to the impact on patients. This can include receiving a false negative test result – leading to a delay in treatment – or a false positive result – leading to inappropriate treatment (1). These test evaluations have been limited mainly because there has been a lack of primary research assessing the impact of testing on patient health outcomes.

Di Ruffano et al (2011) conducted a capture-recapture analysis using two searches (broad and specific) from the Cochrane CENTRAL hand searched trial database to estimate the number of randomised controlled trials published on diagnostic tests between 2004-2007. Of the 23,888 randomised controlled trials retrieved, 135 were found to be diagnostic randomised controlled trials. The capture-recapture analysis estimated 37 diagnostic trials were published per year for the 4 years (2). This is in contrast to the 5,938 therapeutic trials per year known to have been published.

In general, if a test performs poorly in a diagnostic effectiveness trial, false positive and false negative test results will be reflected in the measured health outcomes of patients. However, as trial evidence of the impact of medical tests on the health outcomes of patients is often scarce, policy makers are faced with making decisions on access to, and reimbursement of, diagnostic, staging and screening tests on the basis of incomplete and uncertain information.

To address this lack of evidence, in 2005 a methodology was published that aimed to provide the maximum amount of information to inform assessments of test effectiveness and cost-effectiveness to Australian policy makers (3, 4). This “linked evidence approach” (LEA) involves the narrative linking of evidence assessing components of a test-treatment pathway in order to predict the likely impact of testing on patient health outcomes. The approach aims to evaluate the clinical utility of medical tests in the absence of direct trial evidence such as randomised controlled trials.

Using LEA, systematic literature reviews and meta-analyses (where possible) are conducted on the available existing research to determine test accuracy relative to appropriate reference standards, the impact of the new test on clinical decision-making, and – where circumstances are appropriate (5) – the impact of likely treatment choices on patient health outcomes. These data are used in decision analytic models to determine the effectiveness and cost-effectiveness of the new tests, relative to existing tests.

The method was informed by criteria developed by Fryback and Thornbury (1991) to assess the efficacy of diagnostic imaging tests (6). The method also built upon the analytic frameworks pioneered by the United States Preventive Services Task Force (USPSTF) and used in clinical practice guideline development (7). These frameworks address both the harms and benefits of medical testing on the patient (8-10).

Although LEA was used sporadically in the assessment of medical tests in Australia from 1999, in 2005 the approach was mandated by the federal government (11) when health technology assessments of medical tests were commissioned to inform public funding policy decisions.

The objective of this study was to determine what effect (if any) the use of LEA methodology and other evidentiary factors had on Australian policy makers' decisions to publicly fund diagnostic, staging and screening tests.

METHODS

The independent committee in Australia that makes decisions regarding the funding of medical tests, through the Medicare Benefits Schedule, is the Medical Services Advisory Committee (MSAC). This committee began making recommendations on the reimbursement of new health technologies in 1999. This study therefore covered the period of HTA production and decision-making from 1999 until December 2014. Guidance on the assessment of diagnostic technologies using LEA was introduced in August 2005 (11) but, as various drafts were produced prior to this release date, the whole of 2005 was considered a change-over period in some of the analyses that have been conducted. Policies and practice with regard to the HTA of medical tests were compared before and after 2005.

HTA reports were included in this study if they met the following criteria:

- Considered by MSAC between 1999[9] and December 2014;
- The whole assessment report was publicly available on the MSAC website (www.msac.gov.au) at the cut-off date of December 30, 2014;
- The evaluation was a 'contracted assessment'[10] commissioned by the Australian Government Department of Health, irrespective of whether the health technology was identified through an internal referral, an external application for public funding, or was an update of an application previously considered by MSAC; and
- The report concerned the assessment of a diagnostic, screening or staging test. Definitions of these types of tests have been reported previously (5).

HTAs were excluded from consideration if:

- The test being assessed was used to monitor response to therapy;
- The test being assessed was pharmacogenetic - as the use of LEA for these tests has been reported elsewhere (12, 13); or
- The HTA was commercial in confidence, withdrawn or not produced.

Independent duplicate selection and data extraction occurred for 59 of the 173 (approximately 1/3) test clinical indications that were eligible. The unit of analysis was test evaluation *per clinical indication*, as tests were often used for multiple purposes and

[9] The first report considered by MSAC was in May 1999.

[10] From 2011 MSAC processes changed, allowing industry applicants to choose whether to provide their own full HTA reports for a reimbursement decision. To reduce the potential for confounded results, these 'submission-based assessments' have been excluded; although the majority are not in the public domain in any event.

thus several evaluations may have been included in one HTA report. In addition, information was extracted from public summary documents on the final MSAC funding decision for each medical test. There were five types of funding decision – funding supported, funding rejected, interim funding (approximately 5 years of funding before the decision is reviewed or new evidence is presented), keep current funding (after a funding decision is reviewed favourably), or no decision required (these generally occurred when MSAC was asked for an evaluation but the funding decision rested at a jurisdictional level).

The HTA methodological approach was coded as:

- 'direct evidence only' – reporting only on direct clinical trials assessing the impact of a test on patient health outcomes;
- 'direct evidence plus full LEA' - reporting on direct clinical trials and supplementing this with a linkage of evidence on the accuracy of the medical test, its impact on clinical decision-making (eg changes in patient management), and the effectiveness of consequent treatment options;
- 'direct evidence plus LEA but full linkage not required' - reporting on direct clinical trials and supplementing this with an abridged LEA. An abridged LEA would search for evidence on the accuracy of the medical test and of its impact on clinical decision-making, but would refrain from evaluating the effectiveness of the consequent treatment options. The latter would be considered unnecessary if the new medical test identified patients with a similar spectrum of disease to patients currently receiving standard treatment after diagnosis with existing tests (5);
- 'components of LEA' – reporting on isolated aspects of the test effectiveness pathway (most commonly, test accuracy alone) with no rationale given for selecting only those components; and
- 'direct evidence plus components of LEA' - reporting on direct clinical trials of the test and supplementing this with reporting on isolated aspects of the test

effectiveness pathway (most commonly, test accuracy alone) with no rationale given for selecting only those components.

Data were analysed using Microsoft Excel 2013 and Stata version 13. Logistic regression analyses, with robust variance estimation to account for the non-independence of clustered data, was performed to determine whether use of LEA, or other factors apparent from the evidence-base in each test evaluation, predicted a decision to reject or support the test for public funding. Clustered variances were likely as the same test was often used for multiple clinical indications and so evaluation methodologies were likely to be similar in each report. Independent variables selected *a priori* as possible predictors included: test purpose (add-on test, replacement test, triage test (14)), methodological approach, year of decision, quality of the evidence base (poor/not poor quality, limited/not limited data, low/high applicability, heterogeneity/homogeneity of findings), reference standard (imperfect/accurate). The dependent variable was a positive funding decision. However, as a MSAC funding decision can mean a *new* positive funding decision, the maintenance of funding (ie an interim funded test that is being reviewed) or the decision to interim fund, this dependent variable was a composite. It was disaggregated for various sensitivity analyses.

Regression diagnostics were conducted to confirm model specification and to determine model fit. The Wald statistical test was used to test the hypothesis that the maximum likelihood estimate of the parameters of interest in each model predicted the proposed true value (15). Model selection was primarily informed by Akaike Information Criterion measures to estimate minimisation of information loss (16).

There was no external funding source for this study. The authors performed the research independently and as part of their role as academics at The University of Adelaide.

RESULTS

Of the 259 HTAs available on the MSAC website, 83 were found to meet the eligibility criteria and reported on the use of a test for diagnosis (61%), staging (23%) or screening (12%) purposes for 173 clinical indications. Nearly one half of these were ‘add on’ tests (42%), while approximately one quarter (26%) were ‘replacement’ tests.

39 evaluations of diagnostic, staging or screening tests conducted before LEA was introduced (May 1999 to August 2005), and comprising 63 clinical indications, were compared to 44 evaluations of tests (110 clinical indications) conducted after LEA was introduced (August 2005 – December 2014).

HTA Methodology

A comparison of evaluation methodologies, before and after the use of LEA was recommended by government, indicates that use of the “components of LEA” approach reduced significantly (Figure 1). “Components of LEA” predominantly only considers diagnostic accuracy data and not the downstream effects of a test. Between 2005 and 2010 the use of “components of LEA” ceased completely, only to re-emerge – albeit to a lesser extent – between 2011 and 2014.

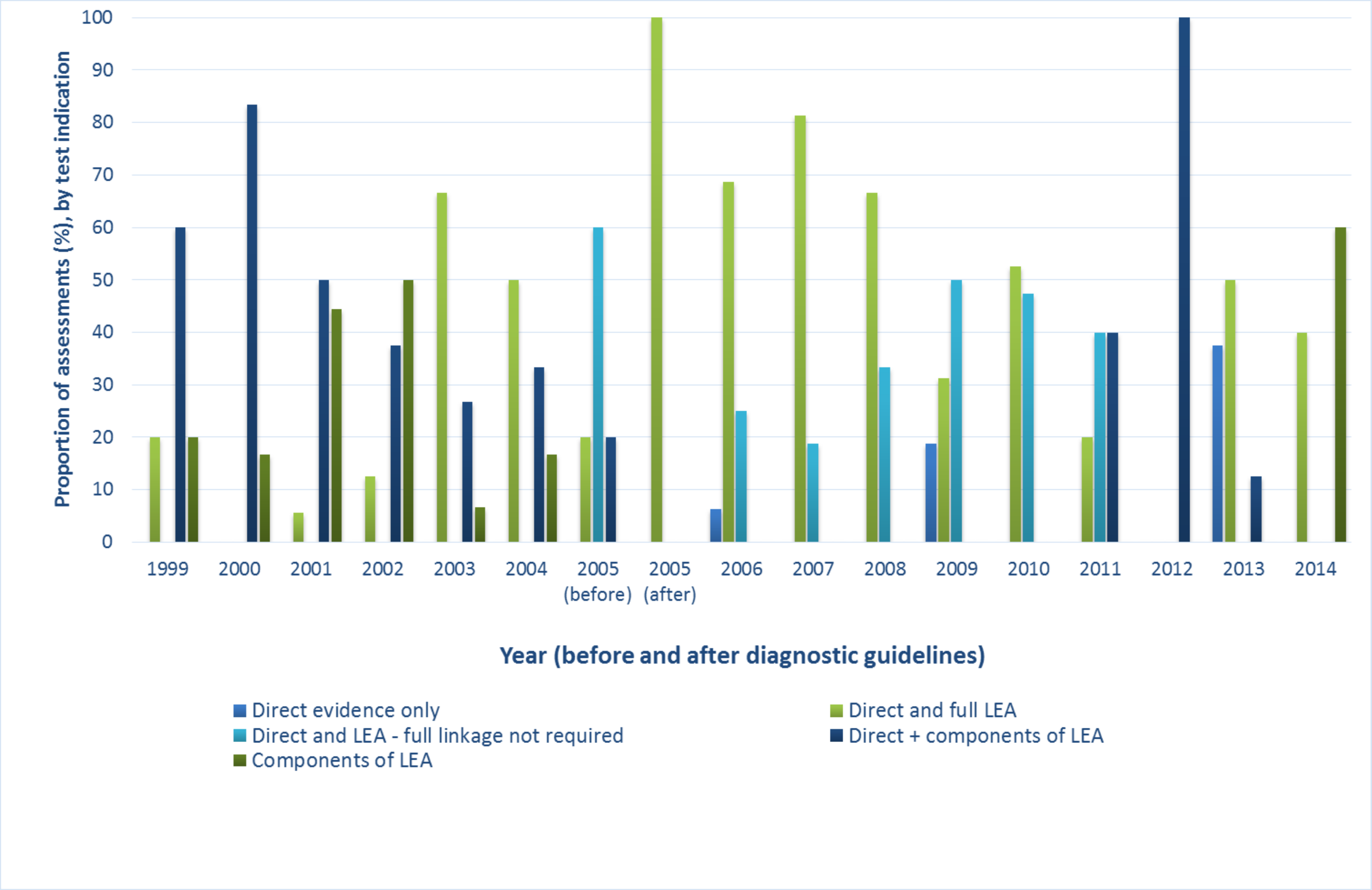


FIGURE 1 **CHANGE IN EVALUATION METHODOLOGY OVER TIME**

MSAC Funding Decisions

Before the introduction of the MSAC guidelines for evaluating diagnostic tests (May 1999 – July 2005), 63 clinical indications for eligible diagnostic, staging or screening tests were assessed to determine if there was sufficient evidence of test safety, effectiveness and cost-effectiveness to warrant public funding through the Medicare Benefits Scheme. 63.5% of the funding decisions were positive. This included 27% where the decision was conditional upon a review in 5 years (interim funding) and 4.8% where the original interim funding decision was confirmed after review. 34.9% of the funding decisions were negative.

After the LEA methodology was formally introduced (August 2005 – December 2014), 110 specific uses of diagnostic, staging or screening tests were evaluated for a public funding decision. 59.1% of these funding decisions proved to be positive. This included 0.9% where the decision was conditional upon a review in 5 years (interim funding) and 10.9% where the original interim funding decision was confirmed after review. 38.2% of the funding decisions were negative.

The most common methodological approach in the reports supporting these decisions was a search for direct evidence, supplemented by a linked evidence approach. In most of these instances there was limited direct evidence available, or it concerned a population or intervention that was not perfectly applicable, and so a full LEA was undertaken to redress shortfalls in the evidence-base.

A comparison of funding decisions before and after the use of LEA was recommended indicates that the proportion of funding decisions informed by the method increased substantially (Figure 2). The odds of an ensuing negative funding decision was five times higher than for a positive funding decision with HTAs that only used “components of LEA” during the period 2005-2014, although there was significant uncertainty about the estimate (unadjusted OR 5.37, 95%CI 0.50, 269.41).

Predicting test funding

Various logistic regression models were tested to determine whether LEA and/or the other pre-specified independent variables predicted the composite dependent variable where either a new positive funding decision or an interim funding decision or a decision to maintain existing funding defined “success”. The models with the best fit are depicted in Table 1 but the prediction capability of three of these four depicted models were still consistent with chance.

The model (Model 1) that independently predicted a public funding decision consisted of two factors - the absence/presence of limited data and the use of the new test as a replacement for an existing test (χ^2=6.63, df=2, p=0.036). In a comparison between Models 1 and 2, Model 1 had a slightly better fit to the data but the difference between the models was not marked.

Other combinations of the pre-specified independent variables, including methodological approach (specifically, the use of LEA), were not significant predictors (data not shown). There was no apparent association between decision year (1999-2014) and public funding recommendations, nor was there an apparent difference in decisions between time periods (before or after introduction of the LEA Guidelines). Results were similar irrespective of whether the year of introducing the LEA Guidelines (2005) was included when time periods were compared (data not shown).

With regard to test purpose, both add-on tests and replacement tests predicted public funding (add on tests – unadjusted OR 2.8, 95%CI 0.97, 8.10, p=0.058; replacement tests – unadjusted OR 4.66, 95%CI 1.40, 15.57, p=0.012), although replacement tests were the stronger predictor. This was not surprising as replacement tests were more likely to be cost-effective or cost-saving than add-on tests. Tests that were undertaken to triage patients were not significant predictors of public funding (unadjusted OR 2.0, 95%CI 0.59, 6.84, p=0.269) but this result was based on only 32 triage tests out of the 173 tests considered for public funding decisions.

FIGURE 2 FUNDING DECISIONS BY METHODOLOGICAL APPROACH

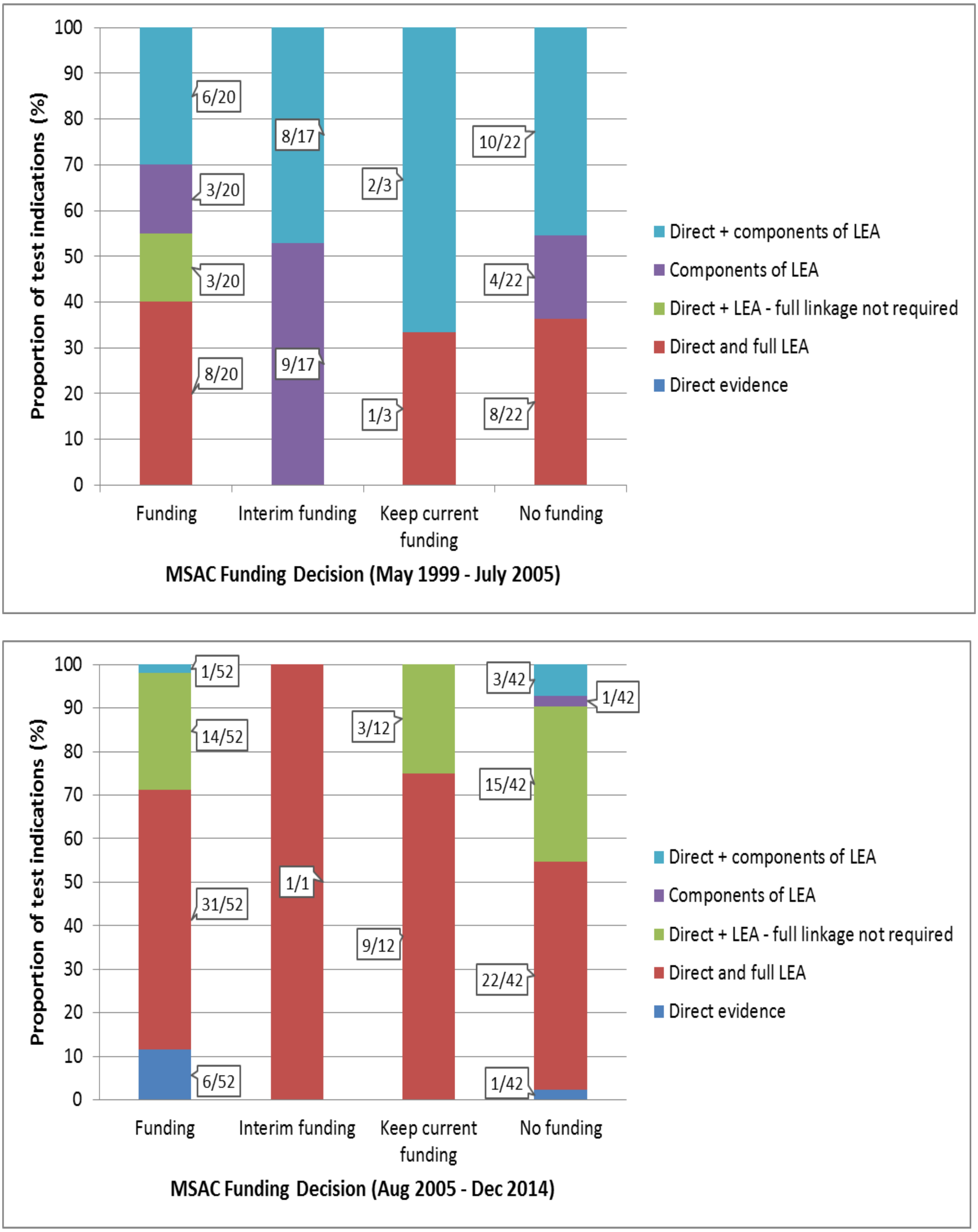


TABLE 1. PREDICTING FUNDING OF MEDICAL TESTS IN AUSTRALIA

Variable	Model 1		Model 2		Model 3		Model 4	
	β [SE]	Robust OR$_{adj}$ [95%CI]	β [SE]	Robust OR$_{adj}$ [95%CI]	β [SE]	Robust OR$_{adj}$ [95%CI]	β [SE]	Robust OR$_{adj}$ [95%CI]
Predicting public funding								
Constant	0.882 [0.409]	2.42 [1.08, 5.39]	1.015 [0.477]	2.76 [1.08, 7.02]	0.812 [0.526]	2.25 [0.80, 6.31]	1.038 [0.483]	2.82 [1.10, 7.28]
Test purpose *Replacement vs not replacement for existing test*	0.794 [0.434]	2.21 [0.94, 5.18]	0.777 [0.442]	2.17 [0.91, 5.17]	0.870 [0.454]	2.39 [0.98, 5.81]	0.833 [0.445]	2.30 [0.96, 5.50]
Reference standard *Accurate vs imperfect*			-0.293 [0.414]	0.75 [0.33, 1.68]				
Evidence quality *Poor vs good quality*							-0.214 [0.413]	0.81 [0.36, 1.82]
Limited vs not limited data available	-0.775 [0.422]	0.46 [0.20, 1.05]	-0.755 [0.428]	0.47 [0.20, 1.09]	-0.916 [0.455]	**0.40** **[0.16, 0.98]**	-0.820 [0.435]	0.44 [0.19, 1.03]
LEA vs no LEA methodology					0.192 [0.518]	1.21 [0.44, 3.35]		
No. of observations	169, adjusted for 81 clusters		169, adjusted for 81 clusters		162, adjusted for 77 clusters		169, adjusted for 81 clusters	

Variable	Model 1	Model 2	Model 3	Model 4
Log pseudo-likelihood	-107.79	-107.39	-103.46	-107.60
Wald Test	**X^2=6.63, df=2, p=0.036**	X^2=7.61, df=3, p=0.055	X^2=6.97, df=3, p=0.073	X^2=6.53, df=3, p=0.089
Pseudo R^2	3.86%	4.22%	4.43%	4.03%
AIC	1.311	1.318	1.327	1.321
AIC*n	221.588	222.781	214.926	223.199

Bold indicates statistically significant predictor. OR_{adj} = odds ratio adjusted for other predictors in the model; CI = confidence interval; β = beta coefficient; SE = standard error; LEA = linked evidence approach; AIC = Akaike information criterion.

Predicting new test funding

Following the introduction of LEA, new positive funding decisions reduced overall by 11% driven by a 25% reduction in time-limited 'interim' funding decisions. 'Definitive' positive funding decisions increased by 15% but the ratio to negative funding decisions was not significantly different between both time periods (unadjusted OR=1.36, 95%CI 0.62, 3.01, χ^2=0.69, p=0.406). The impact of various independent variables, including the use of LEA, on decisions to fund new tests ie new positive funding decisions and interim funding decisions, was also tested.

Four of the better models at predicting new public funding decisions are given in Table 2. Two of these models (Models 1 and 2) demonstrated a greater than chance ability to predict funding of new tests. Both models were driven by the association between decision-making and the presence/absence of limited data. Funding was also affected by whether the new test was a replacement for an existing test (Model 1: X^2=7.22, df=2, p=0.027). Model 2 also incorporated an appropriate reference standard as a prediction variable (X^2=8.88, df=3, p=0.031). In a comparison between Model 1 and 2, Model 2 had a slightly better fit to the data but the difference was weak. Other possible combinations of the independent variables, including methodological approach, did not predict new funding decisions greater than chance (data not shown). The use of LEA did not appear to predict a new positive funding decision.

TABLE 2. PREDICTING *NEW* FUNDING OF MEDICAL TESTS IN AUSTRALIA

Variable	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]
Predicting new funding								
Constant	0.843 [0.414]	2.32 [1.03, 5.23]	1.015 [0.477]	2.81 [1.07, 7.35]	0.945 [0.562]	2.57 [0.86, 7.75]	0.887 [0.483]	2.43 [0.94, 6.26]
Test purpose								
Replacement vs not replacement for existing test	0.724 [0.448]	2.06 [0.86, 4.97]	0.777 [0.442]	2.08 [0.85, 5.07]	0.819 [0.477]	2.27 [0.89, 5.77]	0.732 [0.452]	2.08 [0.86, 5.05]
Reference standard								
Accurate vs imperfect			-0.293 [0.414]	0.66 [0.29 1.53]	-0.465 [0.454]	0.63 [0.26, 1.53]		
Evidence quality								
Poor vs good quality							-0.061 [0.423]	0.94 [0.41, 2.15]
Limited vs not limited data available	-0.919 [0.426]	**0.40** **[0.17, 0.92]**	-0.755 [0.428]	**0.40** **[0.17, 0.94]**	-1.106 [0.467]	**0.33** **[0.13, 0.83]**	-0.929 [0.432]	**0.40** **[0.17, 0.92]**
LEA vs no LEA methodology					0.311 [0.569]	1.36 [0.45, 4.16]		
No. of observations	154, adjusted for 75 clusters		154, adjusted for 75 clusters		147, adjusted for 71 clusters		154, adjusted for 75 clusters	

Variable	Model 1	Model 2	Model 3	Model 4
Log pseudo-likelihood	-99.90	-99.15	-94.66	-99.88
Wald Test	**X^2=7.22, df=2, p=0.027**	**X^2=8.88, df=3, p=0.031**	X^2=9.16, df=4, p=0.057	X^2=7.16, df=3, p=0.067
Pseudo R^2	4.44%	5.15%	5.70%	4.45%
AIC	1.336	1.318	1.356	1.349
AIC*n	205.799	222.781	199.321	207.769

Bold indicates statistically significant predictor. OR_{adj} = odds ratio adjusted for other predictors in the model; CI = confidence interval; β = beta coefficient; SE = standard error; LEA = linked evidence approach; AIC = Akaike information criterion.

Interim funding decisions

Relative to the period before LEA was mandated, the odds of interim funding reduced by 98% (unadjusted OR=0.02, 95%CI 0.0005, 0.17, χ^2=26.44, p<0.001).

Four of the better models at predicting *interim* public funding decisions are given in Table 3. The simplest model (Model 1), with only LEA methodological approach as a predictor, was strongly explanatory of interim funding decisions (X^2=12.63, df=1, p=0.0004). The other three models, with additional independent variables included, were statistically significantly predictive of interim funding, but while the fit of Models 1, 2 and 3 were similar, Model 1 is preferred on the grounds of parsimony.

Models that included LEA were, on the whole, statistically significant because of the strong association between LEA and interim funding decisions – when LEA methodology was used, medical tests did not receive interim funding. There was only one model tested (Model 4) that demonstrated the ability to predict interim funding in the absence of the LEA variable. In this model if there was poor quality evidence and limited data in the test evaluation, as well as an imperfect reference standard, then it was likely that interim funding would not be received (X^2=8.53, df=3, p=0.036).

TABLE 3. PREDICTING *INTERIM* (5 YEAR TIME-LIMITED) FUNDING OF MEDICAL TESTS IN AUSTRALIA

Variable	Model 1		Model 2		Model 3		Model 4	
	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]	β [SE]	Robust OR_{adj} [95%CI]
Predicting interim funding								
Constant	-0.663 [0.523]	0.52 [0.18, 1.44]	-0.316 [0.613]	0.73 [0.22, 2.42]	-0.064 [0.788]	0.94 [0.20, 4.39]	-0.918 [0.813]	0.40 [0.08, 1.96]
Reference standard								
Accurate vs imperfect							-1.217 [0.660]	0.30 [0.08, 1.08]
Evidence quality								
Poor vs good quality					-0.322 [0.581]	0.72 [0.23, 2.26]	-0.019 [0.496]	0.98 [0.37, 2.59]
Limited vs not limited data available			-0.677 [0.591]	0.51 [0.16, 1.62]	-0.732 [0.612]	0.48 [0.14, 1.60]	-1.304 [0.564]	**0.27** **[0.09, 0.82]**
LEA vs no LEA methodology	-4.082 [1.148]	**0.02** **[0.002, 0.16]**	-3.913 [1.131]	**0.02** **[0.002, 0.18]**	-3.943 [1.155]	**0.02** **[0.002, 0.19]**		
No. of observations	166, adjusted for 79 clusters		166, adjusted for 79 clusters		166, adjusted for 79 clusters		173, adjusted for 83 clusters	
Log pseudo-likelihood	-37.80		-37.12		-36.97		-51.91	
Wald Test	**X^2=12.63, df=1,**		**X^2=12.74, df=2,**		**X^2=13.33, df=3,**		**X^2=8.53, df=3,**	

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**
	p=0.0004	**p=0.002**	**p=0.004**	**p=0.036**
Pseudo R^2	33.65%	34.84%	35.07%	10.14%
AIC	0.480	0.483	0.494	0.646
AIC*n	79.602	80.249	81.933	111.814

Bold indicates statistically significant predictor. OR_{adj} = odds ratio adjusted for other predictors in the model; CI = confidence interval; β = beta coefficient; SE = standard error; LEA = linked evidence approach; AIC = Akaike information criterion.

DISCUSSION

We hypothesised that the mandated use of LEA in Australia would change the way that medical tests were evaluated for their clinical utility. The results of our study have confirmed this. LEA methodology was the most common method for presenting data to policy makers between 2005 and 2014. Commissioned HTA assessors have followed the 2005 *MSAC Guidelines for the assessment of diagnostic technologies* (3). Decision-making was based on the linkage of systematically reviewed evidence on test performance, relative to an accepted reference standard, to evidence on the impact of the test on treatment decisions and through careful consideration of the likely impact of the test on patient health outcomes – including the impact of false positive and false negative test results. At the least, it is likely that this could have led to more informed decision-making.

After 5 years the presentation of "unlinked" component evidence (primarily the presentation of technical accuracy alone) has re-emerged. There are several reasons why this may have occurred. HTA processes in Australia were reviewed in 2009 (17) and reforms of the process led to the introduction of an option for applicants for public funding to submit their own assessments of medical tests, which were then critiqued by independent contracted assessors, to facilitate a potentially faster review by the decision-making body (18). The option to have the assessment conducted by independent assessors, essentially "free of charge" was still available but timeliness could not be guaranteed.

To mitigate the effects of this change in process, our study has only included contracted assessments, not submission-based assessments. But unintended consequences of the change process may also have affected contracted assessments. New guidance needed to be developed to assist both applicants and contracted assessors in their assessment of medical tests. The 2005 guidance was archived on the MSAC website but the new guidance (which incorporates LEA methodology) and templates for presenting HTAs on investigative technologies have not yet been released. It is possible that assessors who have been commissioned since 2010 would not have access to "best practice" guidance on the use of LEA in the analysis of tests. Alternatively, it is possible that the proportion of direct evidence available for these assessments was sufficient so that an explicit linkage of evidence was not needed and test accuracy data were provided only for the sake of completeness. Irrespective of the reason it is apparent that contracted assessments that did not use LEA in recent years tended to report on tests that were subsequently rejected for public funding, although the numbers were too small to establish this as occurring above chance.

The results of the logistic regression models indicated that the choice of methodological approach is unlikely to affect the direction (positive or negative) of funding decisions but use of LEA is strongly associated with a negative likelihood of a medical test obtaining interim funding.

It is possible that this change in decision-making is not attributable to the introduction of LEA guidance but rather to other factors. The reduction in interim funding after 2005 could have been the consequence of a policy change at the government level or due to turn-over in the composition of the decision-making committee. However, if this were the case then the association between LEA and the reduction in interim funding would likely only hold for the 87% (n=110 indications) of HTAs that used the method in the period after 2005. It is clear, though, that this also held for the period prior to 2005 where approximately one-third of HTAs (32%, n=63 indications) also used LEA (Figure 2).

The additional information provided in a LEA could plausibly reduce the uncertainty associated with decision-making and therefore reduce the need to make interim funding decisions ie those decisions that are time-limited and are reviewed subsequent to the provision of additional information. The observed concomitant increase in more definitive positive or negative public funding recommendations might have been the result of an increase in the quantity and coherence (prediction of clinical utility) of the information provided to decision-makers. However, the data used in this study were uncontrolled and so other explanations for the change in policy behaviour cannot be ruled out.

If LEA is one of the causes for this change in policy behaviour, then use of the approach might obviate the need for "coverage with evidence development" arrangements for some services that involve medical tests. Coverage with Evidence Development "is characterized by restricted coverage for a new technology in parallel with targeted research when the stated goal of the research or data collection is to provide definitive evidence for the clinical or cost-effectiveness impact of the new technology" (19). In the case of medical tests it is probable that uncertainty will be the norm, as direct evidence of the impact of testing on health outcomes is rare (2) and so decision-maker uncertainty is likely to be high. However, if sufficient linked evidence is already available then there is no need to generate new information to reduce that decision-maker uncertainty. The available evidence simply needs to be identified and selected appropriately and used systematically in decision modelling to predict likely health outcomes.

This does not mean that the use of LEA will always result in certainty. LEA is also affected by the availability of information. If a positive result using the new test results in additional cases being detected, and thus the spectrum of disease in the diagnosed population changes, then evidence of how existing treatments perform in this broader population would be needed. If these data are unavailable, then a linked evidence approach will not be informative (5, 11) and there may be a case for coverage with evidence development, particularly in areas of high unmet clinical need (19).

CONCLUSION

The use of LEA did not appear to affect the direction of reimbursement decisions to any great extent. However, fewer interim funding decisions after introduction of the methodology tends to suggest greater decision-maker certainty regarding the clinical utility of medical tests; although other explanations for this finding cannot be ruled out. Whether the use of LEA in HTA has resulted in *better* decision-making with regards to the funding of medical tests is an issue for future research.

REFERENCES

1. Staub L, Dyer S, Lord S, Simes RJ. Linking the Evidence: Intermediate Outcomes in Medical Test Assessments. International Journal of Technology Assessment in Health Care. 2012;28(1):52-8.
2. di Ruffano L, Davenport C, Eising A, Hyde C, Deeks J. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of Clinical Epidemiology. 2012;65(3):282-7.
3. Medical Services Advisory Committee (MSAC). Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia, 2005 August 2005.
4. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making. 2009;29(5):E1-E12.
5. Merlin T, Lehman S, Hiller JE, Ryan P. The "linked evidence approach" to assess medical tests: a critical analysis. Int J Technol Assess Health Care. 2013;29(3):343-50.
6. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making. 1991;11:88-94.
7. Harris R, Helfand M, Woolf S, Lohr K, Mulrow C, Teutsch S, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35.
8. Nelson H, Huffman L, Fu R, Harris E. Genetic Risk Assessment and BRCA Mutation Testing for Breast and Ovarian Cancer Susceptibility: Systematic Evidence Review for the U.S. Preventive Services Task Force. Annals of Internal Medicine. 2005;143:362-79.
9. Nelson H, Pappas M, Zakher B, Mitchell J, Okinaka-Hu L, Fu R. Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer in Women: A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. Annals of Internal Medicine. 2014;160:255-66.
10. Whitlock E, Garlitz B, Harris E, Bell T, Smith P. Screening for Hereditary Hemochromatosis: A Systematic Review for the U.S. Preventive Services Task Force. Annals of Internal Medicine. 2006;145:209-23.
11. MSAC. Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia, 2005 August 2005.

12. Merlin T, Farah C, Schubert C, Mitchell A, Hiller J, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing codependent technologies. Medical Decision Making 2012.

13. Merlin T. The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines. Personalized Medicine. 2014;11(4):435-48.

14. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. British Medical Journal. 2006;332:1089-92.

15. Wald A. Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. Transactions of the American Mathematical Society. 1943;54:426-82.

16. Akaike H. A new look at the statistical model identification IEEE Transactions on Automatic Control. 1974;19(6):716-23.

17. Australian Government Department of Health and Ageing. Review of Health Technology Assessment in Australia - A discussion paper. Canberra, ACT: Commonwealth of Australia; 2009.

18. Australian Government Department of Health and Ageing. An Overview of the New Arrangements for Listing on the Medicare Benefits Schedule 2012. Available from: http://www.msac.gov.au/internet/msac/publishing.nsf/Content/B8E1F7C44BE7E25BCA257A7D002477C3/$File/Overview-new-arrangements-MSAC.pdf.

19. Trueman P, Grainger D, Downs KE. Coverage with Evidence Development: Applications and issues. International Journal of Technology Assessment in Health Care. 2010;26(1):79-85.

Relevance of Paper 2 to the thesis

Paper 2 has addressed the first two research questions posed in this thesis.

That is, it is clear from the results of the analysis of Australian HTAs in Paper 2 that the linked evidence approach (LEA) methodology is feasible for HTA assessors to undertake when assessing medical tests for public funding decisions. It is also apparent that LEA has become the standard method for evaluating medical tests since the introduction of the 2005 *MSAC Guidelines for the assessment of diagnostic technologies* (MSAC 2005a).

In terms of the second research question for the thesis, the impact of LEA methodology on MSAC's decisions to publicly fund medical tests appeared to be through reduced uncertainty in decision-making. The stark reduction in interim funding decisions whenever the methodology was used in the evaluation of medical tests lends support for this conclusion; however, it is possible that other factors may have had contributory effects such as changes in the composition of MSAC membership or changes in government policy.

The research presented in Paper 2 compared the different test evaluation approaches, and their impact on policy and reimbursement decisions, both before and after the introduction of the 2005 MSAC Guidelines. Apart from the impact of LEA on interim funding decisions, the two key factors that predicted the *direction* of funding decisions was whether the evidence was adequate to inform a decision (limited evidence) and whether the test being assessed would replace an existing test.

Paper 2 does not attempt to identify whether those implementing LEA in the evaluation of medical tests had any difficulty in applying the method. To determine whether there are specific contexts or test types where the application of LEA is more problematic than others, it was essential that the Australian HTAs that used LEA were analysed in considerable depth. Chapter 5 investigates the strengths and weaknesses of LEA and provides evidence-based guidance on how LEA should be practically applied in order to ameliorate any limitations with the method.

CHAPTER 5

DEVELOPING A DECISION FRAMEWORK FOR THE LINKED EVIDENCE APPROACH

HTA agencies internationally have recently adopted or recommended methodological approaches to the evaluation of medical tests that are similar to those elaborated in the Australian MSAC Diagnostic Guidelines (MSAC 2005a).

The European Network for Health Technology Assessment (EUnetHTA) developed the *HTA Core Model for Diagnostic Technologies* in 2008. This guidance manual for evaluating diagnostic tests bases its recommended methodology on the approach suggested by MSAC, and states -

> *"When direct trial evidence is not available, other study types that provide evidence about test safety, accuracy, impact on management and the effectiveness of the treatment, are relevant to the assessment of effectiveness. Evidence from these studies can be linked to yield an estimate of effectiveness of the diagnostic technology **(linked evidence).** " (EUnetHTA 2008)*

Two years later the Agency for Healthcare Research and Quality (AHRQ), in the U.S. Department of Health and Human Services, produced the *Methods Guide for Medical Test Reviews* (AHRQ 2010). AHRQ noted that systematic reviews of medical tests are more challenging than reviews of therapeutic interventions because of the indirect impact of medical tests on important health outcomes. AHRQ suggests the use of analytic frameworks (clinical pathways) when determining what research questions should be evaluated by systematic reviews when evaluating medical tests.

> *"Because of the often-convoluted linkage to clinical outcomes, research studies mostly focus on intermediate outcomes such as diagnostic accuracy. The analytic framework can help users to understand how these intermediate outcomes fit in the pathway to influencing clinical outcomes, and to consider whether these downstream issues may be relevant to the review."(p5, Paper 2) (AHRQ 2010).*

Like the MSAC Guidelines, the AHRQ Methods Guide indicates that in some circumstances diagnostic accuracy studies alone may be adequate for evaluating a medical test but that seven questions should be asked before making this determination:

1. Are extra cases detected by the new, more sensitive test similarly responsive to treatment?
2. Are trials available that selected patients with the new test?
3. Do trials assess whether the new test results predict response?
4. If available trials selected only patients assessed with the old test, do extra cases represent the same spectrum or disease subtypes as trial participants?
5. Are tests' cases subsequently confirmed by the same reference standard?
6. Does the new test change the definition or spectrum of disease (eg earlier stage)?
7. Is there heterogeneity of test accuracy and treatment effect (ie do accuracy and treatment effects vary sufficiently according to levels of a patient characteristic to change the comparison of the old and new test)?

These questions all appear to be aimed at determining something similar to the **transferability** condition as proposed in the MSAC Diagnostic Guidelines (MSAC 2005a).

The National Institute of Health and Care Excellence (NICE) Centre for Health Technology Evaluation has also developed a formal methods guide for its Diagnostics Assessment Programme. This guide provides less detail than that provided by AHRQ or MSAC in their guidance documents but does suggest that -

> *"If, as is likely, there are no end-to-end studies available for a diagnostic technology, then different types of evidence are collected and a linked evidence approach taken." (p71, NICE 2011). (NICE 2011)*

Although these HTA guidance documents for evaluating medical tests provide a suggested methodological approach for the assessment of the body of evidence relating to the effectiveness of a medical test, there has been a stark lack of detail on how the linked evidence framework should be applied.

There have been recent advances in one component of the linkage of evidence to determine test effectiveness, namely the quantitative synthesis (meta-analysis) of test accuracy studies. Another AHRQ methodological overview concerning the meta-analysis and reporting of test accuracy notes -

> *"The many existing frameworks for assessing the value of testing propose a stepwise appraisal process, moving from analytic validity (technical test performance), to clinical*

> *validity (diagnostic and predictive accuracy), clinical utility (effect on clinical outcomes) and overall cost-effectiveness assessment.*[2] *Primary studies that directly address all components of the assessment framework are very uncommon. Therefore, systematic reviewers are typically faced with the task of putting together the pieces of the puzzle by synthesizing studies that address each component of the framework." (AHRQ 2012)*

Staub et al (2012) observe that current guidelines on conducting and reporting medical test HTAs do not provide explicit criteria about when to include intermediate outcomes, such as markers of changes in patient management (Staub et al. 2012). It is unclear what assumptions are necessary when linking evidence of test accuracy with intermediate outcomes and health outcomes, and how to assess the quality of primary studies that examine intermediate outcomes. Fifty percent of the 149 international HTAs collated by Staub and colleagues reported evidence about the consequences of testing beyond test performance, with 41 percent also considering intermediate outcomes such as change in patient management as a consequence of the test. However, overall only 60 percent of the HTAs drew clear conclusions about the clinical effectiveness of the test based on the totality of the evidence available.

Although progress had been made in the theoretical use of the linked evidence approach, there was a lack of guidance on how to practically apply the method. Guidance was needed on how a narrative synthesis of test accuracy studies should be undertaken, when and how the findings of this synthesis should be linked to accumulated evidence on the impact of testing on patient management (intermediate outcomes), as well as when a systematic review of treatment effectiveness studies was needed.

Relevant research question

3. *Are there any specific situations where the use of the linked evidence approach (LEA) is inadequate? If so, are there ways that the approach can be improved?*

PROBLEM IDENTIFIED AND ADDRESSED IN THE PEER-REVIEWED PUBLICATION

What problems have been identified when applying LEA methodology to test evaluations? And, if there have been problems, how can these be overcome? Can the application of LEA be done consistently so that it will facilitate consistent decisions by policy makers across different types of tests?

Paper 3 (page 109) systematically reviews the test evaluations submitted to MSAC since the introduction of LEA methodology but concentrates on authors' reports of the challenges associated with applying the methodology. The aim was to identify particular circumstances when the application of LEA is problematic.

With this information a decision framework was able to be developed that could be used when applying LEA to medical test evaluations. The decision framework provides guidance on the amount and type of evidence that is necessary to inform a reimbursement decision that is intended to be based on the predicted clinical utility of a test. In addition, the framework identifies circumstances when LEA is inadequate to the task and when direct evidence is crucial to inform decision-making.

Statement of authorship

Merlin T, Lehman S, Hiller JE, Ryan P. The 'linked evidence approach' to assess medical tests: critical appraisal. *International Journal of Technology Assessment in Health Care*, July 2013; 29(3):343-350, doi: 10.1017/S0266462313000287. Available at: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8955598

AUTHORS' CONTRIBUTIONS

Tracy Merlin (Candidate)

Conceived and designed the project, wrote the protocol for the systematic review and analysis, undertook the data extraction and analysis, interpreted the results, and developed the decision framework. Drafted the manuscript and incorporated feedback on the manuscript from co-authors and peer-reviewers.

Samuel Lehman

Undertook 50% duplicate data extraction and provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Janet Hiller

Contributed to the interpretation of results and provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Philip Ryan

Contributed to the interpretation of results and provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Merlin T, Lehman S, Hiller JE, Ryan P. The 'linked evidence approach' to assess medical tests: a critical analysis. *International Journal of Technology Assessment in Health Care*, July 2013; 29(3):343-350, doi: 10.1017/S0266462313000287. Available at: http://journals.cambridge.org/action/displayAbstract?fromPage=online&aid=8955598

As such, I have the right to re-use my journal contribution or an adapted version of it in any volume of which I am an editor or author, whether or not for commercial gain, and permission is automatically given to the publisher of such a volume subject to normal acknowledgement.

Paper 3 is reproduced as follows -

The ‘linked evidence approach’ to assess medical tests: a critical analysis

Tracy Merlin[1], Samuel Lehman[1], Janet E Hiller[2] and Philip Ryan[3]

Short title: Analysis of the ‘linked evidence approach’

[1] Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, School of Population Health, University of Adelaide, Adelaide, South Australia, Australia.

[2] Faculty of Health Sciences, Australian Catholic University, Fitzroy, Victoria, Australia

School of Population Health, University of Adelaide, Adelaide, South Australia, Australia

[3] Data Management & Analysis Centre (DMAC), Discipline of Public Health, School of Population Health, University of Adelaide, Adelaide, South Australia, Australia

Corresponding author contact details:

A/Prof Tracy Merlin, BA(Hons), MPH

Managing Director, Adelaide Health Technology Assessment (AHTA)

Discipline of Public Health, Mail Drop DX 650 545,

School of Population Health, University of Adelaide,

Adelaide, South Australia, 5005, Australia.

Tel: 61-8-8313 3575, Fax: 61-8-8313 6899, Email: tracy.merlin@adelaide.edu.au

Total word count of text and abstract: 3963

ABSTRACT

Objectives: A linked evidence approach (LEA) is the synthesis of systematically acquired evidence on the accuracy of a medical test, its impact on clinical decision-making and the effectiveness of consequent treatment options. We aimed to assess the practical utility of this methodology and to develop a decision framework to guide its use.

Methods: As Australia has lengthy experience with LEA, we reviewed HTA reports informing reimbursement decisions by the Medical Services Advisory Committee (August 2005 - March 2012). Eligibility was determined according to pre-determined criteria, and data were extracted on test characteristics, evaluation methodologies and reported difficulties. 50% of the evidence-base was independently analysed by a second reviewer.

Results: Evaluations of medical tests for diagnostic (62%), staging (27%), and screening (6%) purposes were available for 89 different clinical indications. 96% of the evaluations used either the full LEA methodology or an abridged version (where evidence is linked through to management changes but not patient outcomes). 61% had the full evidence linkage. 25% of test evaluations were considered problematic; all involving LEA (n=22). Problems included: determining test accuracy with an imperfect reference standard (41%); assessing likely treatment effectiveness in test positive patients when the new test is more accurate than the comparator (18%); and determining probable health benefits in those symptomatic patients ruled out using the test (13%). A decision framework was formulated to address these problems.

Conclusions: LEA is useful for evaluating medical tests but a stepped approach should be followed to determine what evidence is required for the synthesis.

Word Count: 250

Key words: Diagnostic Test Approval, Evaluation Methodology, Systematic Review, Reimbursement Mechanisms

Acknowledgements: Thanks to Dr Sarah Lord for permission to reproduce Supplementary Figure 1; and to Ms Skye Newton for her comments on the draft prior to submission.

Source of funding: Nil.

INTRODUCTION

Guidance on the assessment of medical tests has been produced only recently (2008-2011) in the United States (1-3) and Europe (4), including an interim methods guide from England (5). Australia developed its own guidance for the assessment of medical tests for reimbursement purposes in 2005 (6-7), proposing a “linked evidence approach”, which has subsequently been recommended in each of these international guidance documents.

A recent review (8) of 149 English-language health technology assessments (HTAs) of medical tests, conducted by 18 agencies in 8 countries indicated that the majority of HTAs using LEA follow the Australian evaluation framework. As policies regarding public funding are dependent on the quality and quantity of information provided to the decision-maker, it is timely to reflect on the lessons learned from the application of LEA.

What is the Linked Evidence Approach (LEA) and when is it used?

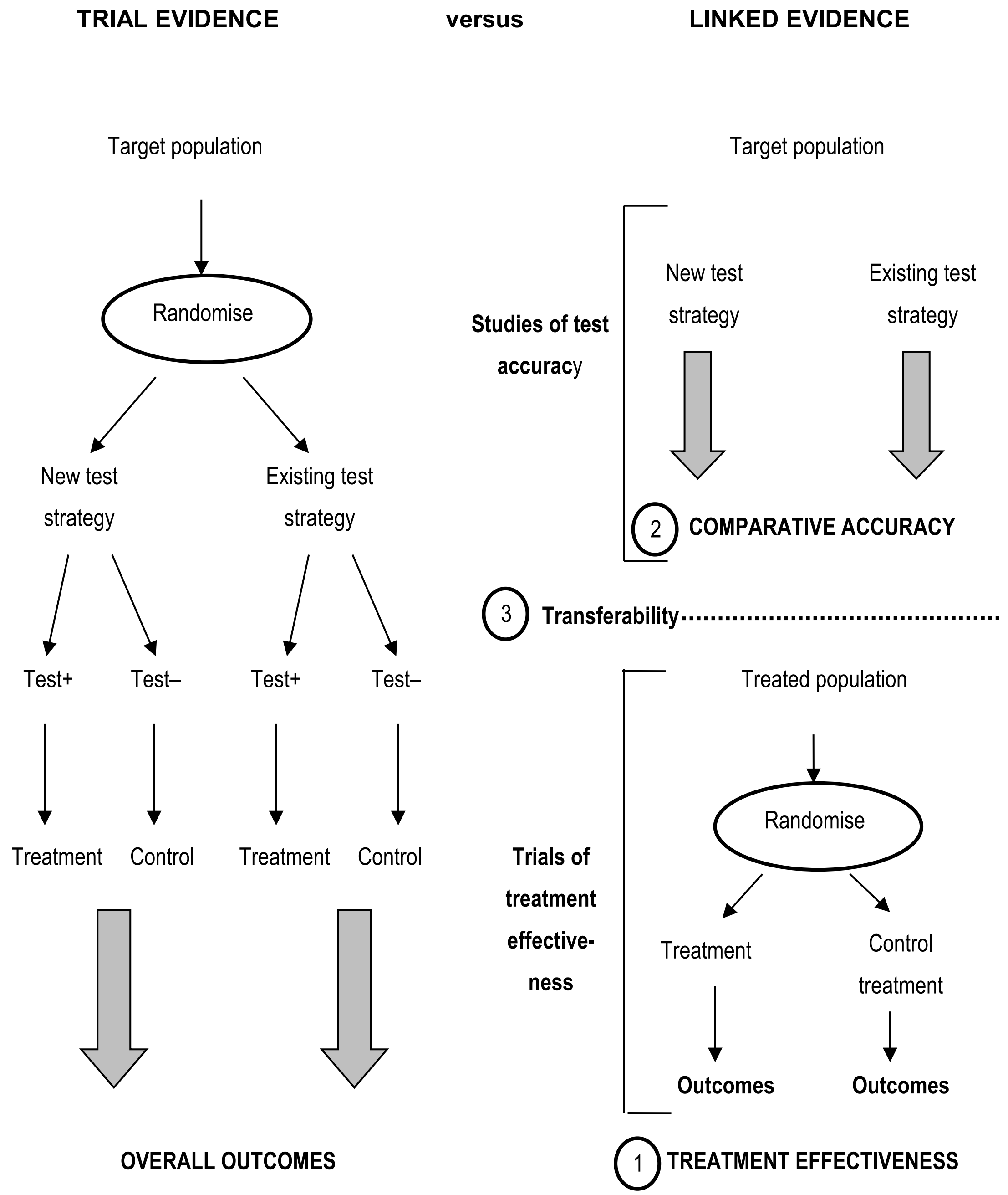
LEA methodology in Australia (6-7) was based on analytic frameworks used by the United States Preventive Services Task Force (USPSTF) in the development of clinical practice guidelines (9), as well as criteria developed by Fryback and Thornbury (1991) to assess the efficacy of diagnostic imaging tests (10). Fryback and Thornbury’s efficacy criteria includes technical efficacy, diagnostic accuracy, diagnostic thinking (change in diagnosis), therapeutic efficacy (change in management) and patient outcome efficacy (change in health outcomes). “Outcome efficacy” or clinical effectiveness is the factor that is of the greatest relevance to policy makers for public funding decisions, and to clinicians determining the best use of testing in managing their patients.

The paramount method of determining the clinical effectiveness of a test is through the direct impact of the test on patient health outcomes. This is, ideally, a randomised controlled trial whereby patients are randomised to assessment with or without use of the medical test and, subsequent to treatment, their health outcomes are measured. However, this type of direct evidence is often lacking (11).

Di Ruffano et al noted this lack, stating “*policy and decision makers frequently need to resort to lower grade evidence, such as decision models to provide guidance on test selection and use*” (11). The Australian, and more recent US and European, test evaluation guidance outlines methods to deal with this type of evidence.

The Australian Medical Services Advisory Committee (MSAC) guidelines recommend the systematic review and narrative linking of key aspects of Fryback and Thornbury's efficacy criteria, under certain conditions. This linking of evidence would occur in instances where direct trial evidence of the clinical effectiveness of a test is not available, or is inadequate for decision making purposes (6). In some cases, evidence of test accuracy would be considered a sufficient proxy for diagnostic effectiveness if there is reasonable justification to assume that the population receiving the test (and within which its accuracy has been tested) is to all intents and purposes the same population that would receive treatment for the condition – and there is good evidence that treatment impacts positively on the health outcomes in this population. This is the *transferability* assumption (see Figure 1).

FIGURE 1 THE USE OF DIRECT EVIDENCE COMPARED TO LINKED EVIDENCE IN THE EVALUATION OF DIAGNOSTIC TESTS



Source: Lord S, Ghersi D, Simes J, Irwig L. (2005) *Medical Services Advisory Committee: Guidelines for the assessment of diagnostic technologies*. Canberra: Commonwealth of Australia. Reproduced with permission.

Transferability cannot be assumed if a positive result using the new test leads to earlier, new or alternative treatments that have not been evaluated in clinical trials. If the new test results in additional cases being detected, and thus the spectrum of disease in the diagnosed population changes, then evidence of treatment effectiveness in this broader population (via a systematic review of treatment effectiveness) would be needed. If these data are unavailable, then a linked evidence approach is not informative (6).

Objective

The aim of this study was to determine whether LEA is feasible and to identify situations where its use may be problematic for informing reimbursement decisions. The objective was to use these data to inform the development of a decision framework to be used when applying LEA.

METHODS

HTA reports commissioned by MSAC, and conducted by predominantly independent academic evaluation groups, were included in the analysis if they met the following criteria:

- Considered by MSAC between August 2005 and March 2012;
- Publicly available on the MSAC website (www.msac.gov.au) between February and March 2012;
- A test requested for reimbursement through government referral, industry application, or an update of a previous assessment; and
- A test used for diagnostic, screening or staging purposes.

Diagnostic tests were considered to identify new pathological conditions in symptomatic patients; screening tests were considered to identify new pathological conditions in asymptomatic or apparently healthy persons; and staging tests were considered to characterise the stage of disease in a patient previously diagnosed. Diagnostic tests that may have a therapeutic component were included eg a biopsy that happened to capture all of the diseased tissue and so effectively treated the condition.

HTAs were excluded in the following circumstances:

- The test being assessed was used for monitoring a specific treatment eg titrating a drug according to a biomarker concentration;
- The test being assessed was pharmacogenetic ie part of a co-dependent technology pairing (12); or
- The HTA was commercial in confidence, withdrawn or not produced.

Monitoring and pharmacogenetic tests were excluded because the relationship with a single (usually drug) treatment is closer, thus the likelihood of direct evidence being available is higher than with diagnostic, staging or screening tests.

All agencies commissioned to undertake evaluations of medical tests for MSAC were required to follow the MSAC diagnostic guidelines following their implementation in 2005 (6).

Independent duplicate selection and data extraction occurred for 50% of all identified HTAs. The unit of analysis was test evaluation *per clinical indication*, as tests were often used for multiple purposes and thus several evaluations may have been included in one HTA report. Data were extracted and coded for the following variables: report details, author, test type (high sensitivity and specificity, rule in, rule out, not enough information, other) and purpose (triage, replacement, add-on), the target population for the test (clinical indications), year of MSAC consideration, the comparator test, identified reference standard to determine test accuracy, quality of reference standard (as discussed in the report), methodological approach, and methodological issues encountered (problems with LEA as discussed in the report). Methodological approach was coded as:

- 'direct evidence only' – reporting only on direct clinical trials, from test to measurement of patient health outcomes;
- 'direct evidence plus full LEA' - reporting on direct clinical trials and supplementing this with a linkage of evidence on the accuracy of the medical test, its impact on clinical decision-making (eg changes in patient management), and the effectiveness of consequent treatment options;
- 'direct evidence plus LEA but full linkage not required' - reporting on direct clinical trials and supplementing this with an abridged LEA. An abridged LEA would search

for evidence on the accuracy of the medical test and of its impact on clinical decision-making, but would not then assess the effectiveness of consequent treatment options due to the treatment being well established and the patient spectrum of disease being similar to those patients currently receiving treatment;

- ‘components of LEA’ – reporting on isolated aspects of the test effectiveness pathway (most commonly, test accuracy alone) with no rationale given for selecting only those components
- ‘direct evidence plus components of LEA’ - reporting on direct clinical trials and supplementing this with reporting on isolated aspects of the test effectiveness pathway (most commonly, test accuracy alone) with no rationale given for selecting only those components

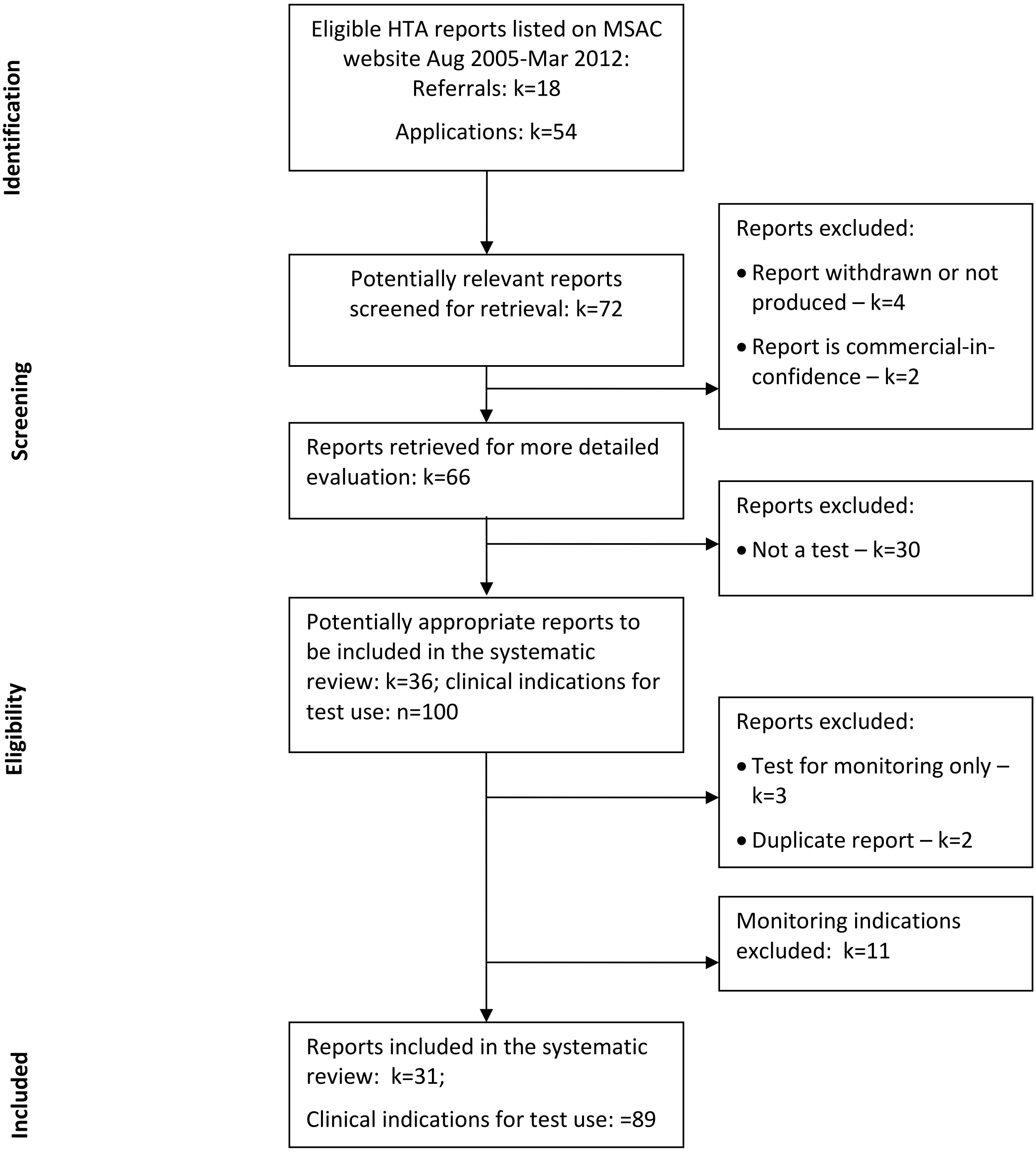
Tests were characterised as having high sensitivity and specificity if, relative to an appropriate reference standard, both parameters were 85% or higher. ‘Rule in’ tests were defined as having high positive predictive value (as reported by the authors), or in the absence of prevalence data, high specificity. ‘Rule out’ tests were defined as having high negative predictive value (as reported by the authors), or in the absence of prevalence data, high sensitivity.

Descriptive statistics were calculated and results were analysed qualitatively.

RESULTS

Figure 2 outlines the process used to select eligible HTAs for the review.

FIGURE 2 PRISMA FLOWCHART. ADAPTED FROM LIBERATI ET AL. (2009)


Identification
Screening
Eligibility
Included
Eligible HTA reports listed on MSAC website Aug 2005-Mar 2012:
Referrals: k=18
Applications: k=54
Potentially relevant reports screened for retrieval: k=72
Reports excluded:
• Report withdrawn or not produced – k=4
• Report is commercial-in-confidence – k=2
Reports retrieved for more detailed evaluation: k=66
Reports excluded:
• Not a test – k=30
Potentially appropriate reports to be included in the systematic review: k=36; clinical indications for test use: n=100
Reports excluded:
• Test for monitoring only – k=3
• Duplicate report – k=2
Monitoring indications excluded: k=11
Reports included in the systematic review: k=31;
Clinical indications for test use: =89

We identified test evaluations for 89 clinical indications in 31 eligible HTA reports. Testing was reported as being undertaken for diagnostic purposes (62%), staging (27%) and for screening (6%). 4% of tests were classified as both diagnostic and staging, while 1% were jointly diagnostic and therapeutic.

Of the 89 test evaluations, 96% used either an abridged (where evidence is linked through to management changes but not patient outcomes) or full LEA methodology, with 61% undertaking the full linkage. Overall, 35% of test evaluations were reported as not requiring a full linkage of evidence. This was usually because the test did not identify patients with a different spectrum of disease (ie different marker or stage of disease) and, as treatment effectiveness was already well known in that patient population, evidence of the impact of treatment did not need to be re-evaluated. The proportion of abridged LEA evaluations increased from 19% in 2007 to 47-50% two to three years later.

In 25% (N=89) of the test evaluations, the HTA authors reported difficulties with methodology. These difficulties all involved the use of an abridged or full LEA. None of these evaluations involved the 4% of HTAs that used an approach that synthesised direct evidence alone.

In the 'problematic' HTAs using LEA (N=22), five main challenges were identified.

1. *Imperfect reference standard*

In 34% of cases where there was not enough information to determine test accuracy, problems in applying LEA were reported. Test accuracy could not be determined because there were insufficient or only low quality studies available or the reference standard was imperfect. Where evidence was lacking, most HTA authors did not report a fault with the LEA approach, they simply reported that the evidence-base was limited. However, when problems with LEA were reported (N=22), 41% of the problems identified involved an imperfect reference standard against which test accuracy (the first component of the linkage) was benchmarked. These included HTAs on optical coherence tomography (13) and molecular testing for myeloproliferative disease (14).

2. *Spectrum of disease differences*

When the new test was more accurate than the designated comparator, inability to assess likely treatment effectiveness in test positive patients was a frequently reported difficulty (18%, N=22). Current treatment options would have only been trialled in populations with a spectrum of disease identified by the less accurate comparator test. Overall, 33% (N=15) of

HTAs of highly sensitive and specific tests reported difficulties using LEA. These included positron emission tomography for staging cervical cancer (15), and magnetic resonance imaging for breast cancer screening in high risk women (16).

3. 'Rule out' tests

Determining probable health benefits in symptomatic patients that are ruled out from the target condition can also be difficult using LEA. Evidence cannot practically be obtained on the myriad of treatment options that may be offered a patient testing negative. Perhaps they receive an early and accurate differential diagnosis to explain their symptoms or, if triage tested, avoid further unnecessary, and potentially invasive, testing. Approximately half (43%) of the handful of HTAs of 'rule out' tests (N=7) reported difficulties applying LEA.

Example HTAs where this problem was reported include brain natriuretic peptide testing to rule out heart failure (17) and positron emission tomography to rule out glioma (18). In the remaining HTAs of this test type there was insufficient information to fully complete the evidence linkage ie there was no apparent change in patient management as a consequence of the test or the data were insufficient to come to any conclusions regarding a change in management. Therefore, problems that would normally be faced when addressing the third linkage (impact on patient health outcomes) did not eventuate.

4. Established tests

Medical tests that are already in established practice but have not previously received public funding were considered difficult to assess. In this situation, nominating the appropriate comparator test strategy was reported as the main difficulty. This issue was reported in HTAs of urinary metabolic profiling for the detection of metabolic disorders (19).

5. Surrogate outcomes

Evaluating the clinical impact of tests when the evidence was limited to surrogate outcomes was reported as an issue. Additional information would be required in the linkage to address the validity of the surrogate outcome. For example, hepatitis B virus (HBV) DNA testing and the use of serum HBV DNA levels as a surrogate for clinical outcomes (20) would require – in the absence of direct evidence – information on the prognostic value of serum HBV DNA levels.

No problems were identified using LEA for 'rule in' tests.

Development of a decision framework to apply LEA

A decision framework was developed to help guide the implementation of LEA (Figure 3). This framework was developed on the basis of information obtained on LEA during the systematic review, most notably the increasing use of abridged LEA, indicating that evaluators are applying their own 'rules' when using a linked evidence approach.

The framework incorporates three scenarios:

A. Optimisation

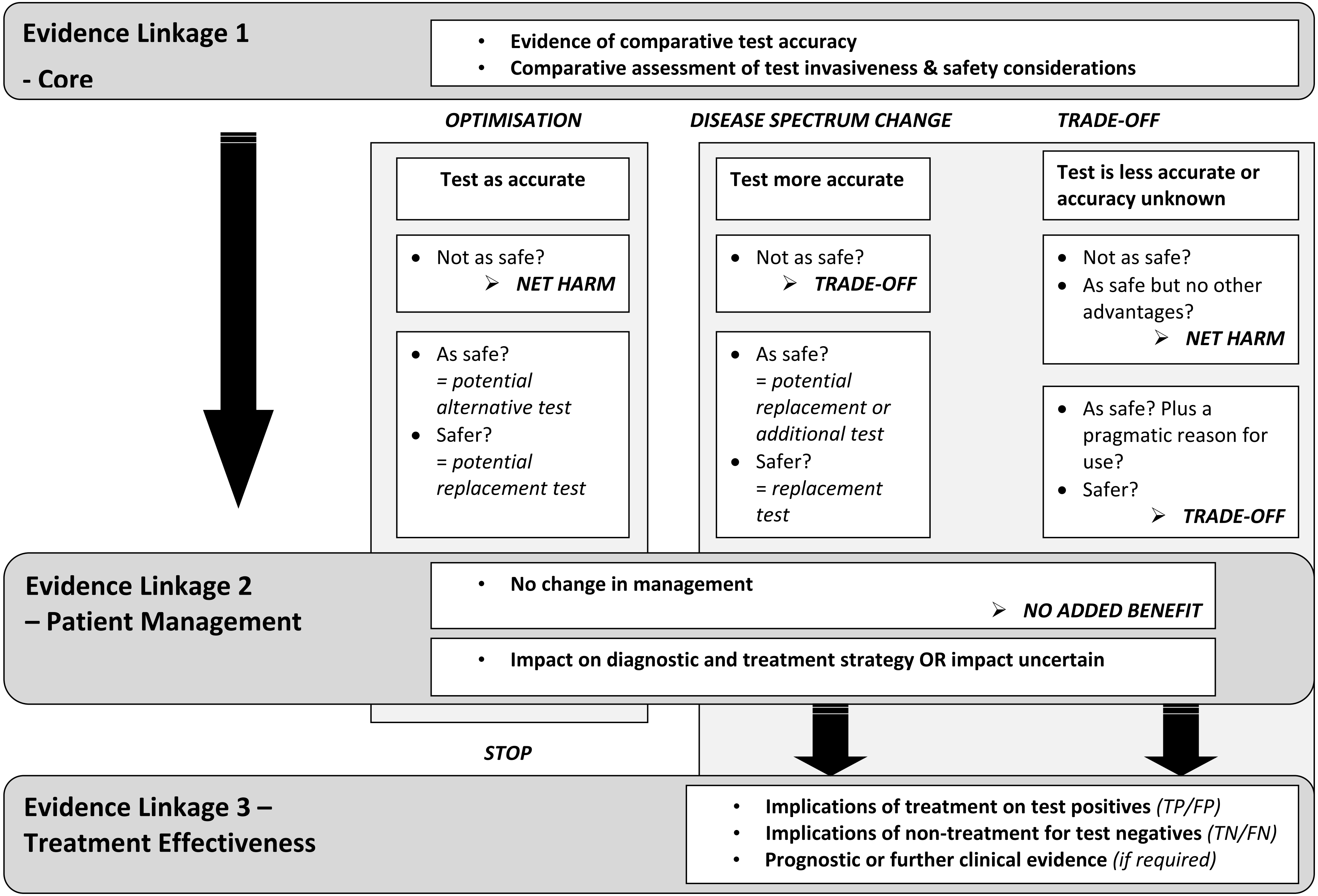
In this scenario if the test is found to be as accurate as the comparator test but not as safe, the result is a net harm; any additional evidence to inform the policy-maker (including cost information) is likely to be superfluous. Conversely, an assessment of the impact of the new test on patient management is recommended when safety is *not* a concern as decision makers will be interested in whether the test has any advantages over its comparator in terms of utilisation (and thus cost implications). As the spectrum of disease in patients receiving these tests is unlikely to differ from that in the existing treated population (given test accuracy is similar), a review of treatment effectiveness would not be required as the treatment options are unlikely to change. At best, if there are safety or accessibility benefits with the new test, the management and treatment of tested patients will be optimised.

B. Trade-off

When the test being assessed is less accurate than the comparator test, then an assessment of test invasiveness or safety is needed to determine whether there is a net harm or a trade-off in safety and test performance. The trade-off analysis will need to determine the consequences of treating or not treating, respectively, the likely increase in false positive (FP) or false negative (FN) diagnoses. Treatment options for patients with a true positive (TP) or true negative (TN) diagnosis are unlikely to change as a consequence of the test and so do not need assessment.

When it is impossible to determine test accuracy (eg imperfect reference standard) a conservative approach is needed to determine all the possible consequences of testing. The implications of false negatives and positives need to be explored, as well as, conversely, the potential to uncover a spectrum of disease for which the natural history (and therefore impact of treatment) is largely unknown (see Scenario C below). Sequential linkages of evidence are required to build a picture of the overall clinical effectiveness of the test. With each linkage in the synthesis, the uncertainty regarding the transferability between linkages is increased.

FIGURE 3 DECISION FRAMEWORK TO IMPLEMENT THE LINKED EVIDENCE APPROACH WHEN EVALUATING MEDICAL TESTS



Evidence Linkage 1

- Core

- **Evidence of comparative test accuracy**
- **Comparative assessment of test invasiveness & safety considerations**

OPTIMISATION

Test as accurate

- Not as safe?
  - ➢ ***NET HARM***

- As safe?
  = potential alternative test
- Safer?
  = potential replacement test

DISEASE SPECTRUM CHANGE

Test more accurate

- Not as safe?
  - ➢ ***TRADE-OFF***

- As safe?
  = potential replacement or additional test
- Safer?
  = replacement test

TRADE-OFF

Test is less accurate or accuracy unknown

- Not as safe?
- As safe but no other advantages?
  - ➢ ***NET HARM***

- As safe? Plus a pragmatic reason for use?
- Safer?
  - ➢ ***TRADE-OFF***

Evidence Linkage 2

– Patient Management

- **No change in management**
  - ➢ ***NO ADDED BENEFIT***

- **Impact on diagnostic and treatment strategy OR impact uncertain**

STOP

Evidence Linkage 3 –

Treatment Effectiveness

- **Implications of treatment on test positives** *(TP/FP)*
- **Implications of non-treatment for test negatives** *(TN/FN)*
- **Prognostic or further clinical evidence** *(if required)*

C. Disease spectrum change

Of all the scenarios, the one where a randomised controlled trial is most needed is when the new test proves to be more accurate than the comparator test. In the absence of direct evidence, the consequences of treatment, or avoidance of treatment, in all patients receiving a more accurate test are difficult to determine because the absolute benefit of the treatment in the new cases detected is not likely to be known. This benefit is likely to depend on the patient prognosis without the treatment, as well as the comparative effectiveness and risks of the treatment in these particular patients (7).

If the test is more accurate but less safe than the comparator test, there is a trade-off situation and a cost-effectiveness analysis is likely to be warranted. If the test has similar safety it may be used as an additional test for patients testing negative on the comparator. If the test has better performance and safety, then a cost-effectiveness analysis may be performed to determine whether it is a suitable replacement for the comparator.

If the test is more sensitive, prognostic or clinical evidence is needed to determine treatment effectiveness in patients diagnosed with the new test. Evidence is also needed on the impact of early versus delayed treatment to determine if there are benefits associated with the reduction in false negatives. If the test is more specific, prognostic or clinical evidence is needed to determine if there are better health outcomes in true negatives. Evidence is also needed on the consequences of inappropriate treatment of false positives to determine if there are benefits associated with the reduction in false positives. With each linkage in the synthesis, the uncertainty regarding the transferability between linkages increases.

DISCUSSION

Feasibility of LEA

In most cases where *direct* evidence of a medical test's impact on patient health outcomes is limited or lacking, LEA can provide a transparent evidence synthesis to inform public funding decisions regarding the clinical effectiveness of the test. Further, because the data has been systematically acquired, it can then be used as inputs in the decision analytic modelling underpinning an economic analysis, leading to arguably less biased representation of inputs and transition probabilities in economic models (21).

However, there are some situations where the LEA synthesis may mislead policy-makers as to the clinical effectiveness of the test, either because insufficient information is presented to address areas of uncertainty or because these uncertainties have not been explicated. Some of these situations were anticipated by the MSAC diagnostic guidelines (6); namely, that LEA may be inadequate to act as a proxy for *direct* evidence in instances where there are spectrum of disease differences between the tested population and the treated population (ie the test identifies new cases that cannot be identified with existing tests); and where there is an imperfect reference standard against which to determine test accuracy.

We have identified two circumstances where evidence additional to the standard LEA synthesis is considered necessary – (i) 'rule out' tests, and (ii) when evidence only reports on surrogate outcomes.

Currently the traditional linked evidence approach is based on the assumption that the test predicts the disease and that this will impact on the health of patients with that disease. The framework does not take into account the benefits or harms from being 'ruled out' from the disease and/or investigated for a different condition, as would occur with *direct* evidence. Health outcomes in test-treatment trials are captured for all patients who test positive *and* negative for the condition in both the new test and existing test trial arms (Figure 1). This is of particular relevance to triage testing as the benefits of a triage test often reside in those patients 'ruled out' from the diagnosis, through not having unnecessary, usually invasive, 'gold standard' testing and/or earlier differential diagnosis and management of the cause of their symptoms (22). Inability to measure the health benefits from being ruled out can be particularly critical when assessing the cost-effectiveness of a triage test. It is important that some attempt is made to identify if there are any health benefits from 'ruling out' symptomatic patients from a condition through use of the test.

This is not a concern in a 'well' or screening population that is receiving the triage test. Those that are 'ruled out' (assuming the test has a low false negative rate) are simply confirmed as healthy. They do not need to be investigated for alternative diagnoses and so treatment effectiveness in the 'ruled out' arm is not an issue. In a screening population, the main issue is false positive and true positive diagnoses and these factors would be considered under LEA.

In instances where an HTA reports spectrum of disease differences between tested and treated populations, or when outcomes reported in the evidence base are surrogates for clinical endpoints, it has been suggested that additional information is provided to address likely patient prognosis following treatment in the tested population. This could take the form of, respectively, a short-term randomised controlled trial comparing treatment outcomes in those receiving the new test versus the comparator test (7), or observational evidence demonstrating an association between the surrogate outcome and patient-relevant clinical outcomes (23).

When undertaking an HTA of an established test, LEA was reported as challenging because of the difficulty in identifying the relevant comparator test. This problem arises simply because the evidence base (whether 'direct' or LEA) assumes that the established test is the benchmark and thus it is either incorporated in the comparator or the only available comparators are new/unassessed tests. In these cases, historical comparators may be used (eg by assuming a scenario where the test was never established)(14) or surveillance of clinical outcomes in patients receiving the established test could be used to supplement the linkage.

Decision framework to apply LEA

The draft methods guide, released by the Agency for Healthcare Research and Quality (AHRQ) (1), suggests that analytic frameworks (9) and/or decision trees and flow charts should be created as a matter of principle when reviewing medical tests. Complementary to this approach, Lord et al suggest using the principles of randomised controlled trial design as a hypothetical framework to identify what types of comparative evidence are required to evaluate medical tests (7).

These frameworks for evaluating medical tests rightly suggest that all relevant areas of evidence-based enquiry should be mapped out prior to collating and selecting evidence. However, little attention is given as to whether it is still relevant to pursue the planned synthesis once there are findings that negate the need to continue with the linkage.

Our review of MSAC HTAs, although potentially limited by duplicate data extraction of only half of the assessments, found that over time there was a reduction in the proportion of evidence syntheses that undertook a full linkage of evidence. In later years only approximately half reported that a full linkage was either possible or warranted. No formal decision framework was presented to justify this abridged linkage, although the logic for truncating the synthesis was invariably provided. These abridged linkages may have increased over time as a consequence of growing familiarity with LEA by the HTA evaluation groups or the evidence may just not have been available to proceed with a full linkage and so the LEA was truncated by necessity.

On the basis of these observations, we have proposed a formal decision framework for applying LEA. The framework is Bayesian in that prior information affects subsequent evidence synthesis decisions. Although the work is limited to Australian HTAs, the identified benefits and limitations with LEA are likely to be broadly applicable to any HTA of medical tests; although, this would need to be tested.

Policy implications of this research

Medical tests are complex interventions, simply because of the downstream consequences associated with testing. General methods for dealing with complex interventions have been proposed (24), as well as methods specific to medical tests (1, 7). These include conceptualising *a priori* the overall theoretical basis for linking evidence (7), as well as the optimal study designs needed to address or measure assumptions inherent in the synthesis plan (8).

Where this paper differs from previous research is by proposing that any *a priori* conceptualisation of questions relevant to an evidence synthesis for a medical test should subsequently be tailored according to the evidence that is found. We have formulated a framework that recognises the necessary pre-conditions for determining the clinical effectiveness of a test. When these conditions are not met, it is wasteful of resources and potentially confusing to policy-makers to proceed with the collation of evidence as outlined in the synthesis plan.

These pre-conditions appear to have been informally implemented, to a greater or lesser extent, with growing frequency in recent Australian HTAs. The decision framework we have proposed incorporates the lessons learned with LEA, and aims to facilitate transparency and standardised use of the methodology.

REFERENCES

1. AHRQ. Methods Guide for Medical Test Reviews. Rockville, MD: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; 2010.

2. CDC. ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing [internet]: Centers for Disease Control and Prevention (CDC), Office of Public Health Genomics; 2010. Available from: http://www.cdc.gov/genomics /gtesting/ACCE/acce_proj.htm.

3. FDA. In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff [internet]. Rockville, Maryland: Food and Drug Administration (FDA), U.S. Department of Health and Human Services; 2011.

4. EUnetHTA. HTA Core Model for Diagnostic Technologies. Work Package 4: European Network for Health Technology Assessment2008.

5. NICE. Interim methods statement (pilot). Version 8 [internet]. London: National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme; 2010.

6. MSAC. Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia2005 August 2005.

7. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making. 2009 Sep-Oct;29(5):E1-E12.

8. Staub L, Dyer S, Lord S, Simes RJ. Linking the Evidence: Intermediate Outcomes in Medical Test Assessments. International Journal of Technology Assessment in Health Care. 2012;28(1):52-8.

9. Harris R, Helfand M, Woolf S, Lohr K, Mulrow C, Teutsch S, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35.

10. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making. 1991;11:88-94.

11. di Ruffano L, Davenport C, Eising A, Hyde C, Deeks J. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of Clinical Epidemiology. 2012;65(3):282-7.

12. Merlin T, Farah C, Schubert C, Mitchell A, Hiller J, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing codependent technologies. Medical Decision Making 2012 August 22, 2012.

13. Marinovich L. Optical Coherence Tomography [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au.

14. Buckley L, Wang S, Merlin T. Molecular testing for myeloproliferative disease. Part A – Polycythaemia vera, essential thrombocythaemia and primary myelofibrosis. Part B - Systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au.

15. Schoeppe S, Lewis S, Marinovich L, Wortley S. Positron emission tomography for cervical cancer [internet]. Canberra: Commonwealth of Australia; 2010. Available from: www.msac.gov.au.

16. Lord S, Lei W, Griffiths A, Walleser S, Parker S, Thongyoo S, et al. Breast magnetic resonance imaging [internet]. Canberra: Commonwealth of Australia; 2007. Available from: www.msac.gov.au.

17. Merlin T, Moss J, Brooks A, Newton S, Hedayati H, Hiller J. B-type natriuretic peptide assays in the diagnosis of heart failure [internet]. Canberra: Commonwealth of Australia; 2008. Available from: www.msac.gov.au.

18. Marinovich L, Wortley S. Positron emission tomography for glioma. Canberra: Commonwealth of Australia; 2010.

19. Gillespie J, Guarnieri C, Phillips H, Bhatti T. Urinary metabolic profiling for detection of metabolic disorders [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au

20. Gillespie J, Smala A, Walters N, Birinyi-Strachan L. Hepatitis B virus DNA testing [internet]. Canberra: Commonwealth of Australia; 2007. Available from: www.msac.gov.au.

21. Craig D, McDaid C, Fonseca T, Stock C, Duffy S, Woolacott N. Are adverse effects incorporated in economic models? A survey of current practice. International Journal of Technology Assessment in Health Care. 2010;26(03):323-9.

22. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. British Medical Journal. 2006;332:1089-92.

23. Micheel C, Ball J, editors. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Washington DC: National Academy of Sciences; 2010.

24. Anderson L, Petticrew M, Rehfuess E, Armstrong R, Ueffing E, Baker P, et al. Using logic models to capture complexity in systematic reviews. Research Synthesis Methods. 2011;2:33-42.

Relevance of Paper 3 to the thesis

Paper 3 has addressed the third research question posed in this thesis, that is:

Are there any specific situations where the use of the linked evidence approach is inadequate? If so, are there ways that the approach can be improved?

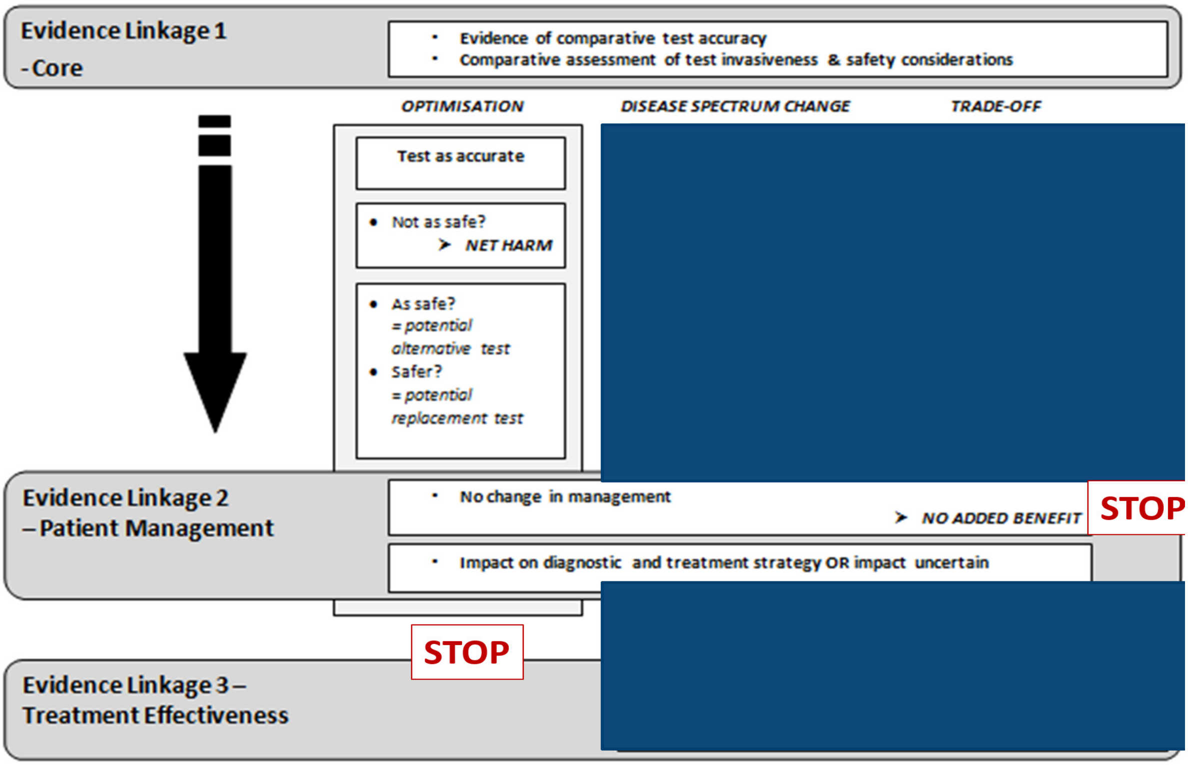
The research found that in most circumstances the use of LEA in medical test evaluation is straightforward. In the 25 percent of test evaluations where there were difficulties in applying LEA, this occurred when:

- the test identified patients with the condition at an earlier stage of the disease than the comparator test and so the safety and effectiveness of the currently available treatments for these patients would be unknown;
- the reference standard test for determining test performance was imperfect;
- the test was used for triaging symptomatic patients;
- only surrogates were available to measure the health impact of treatment;
- the test being evaluated was already established in clinical practice.

The identification of these inadequacies with LEA led to the formulation of a decision framework that proposed specific strategies to circumvent test evaluations that may be problematic. Three common scenarios were described to provide guidance to HTA practitioners as to the amount and type of evidence required to address a policy question *within the context* of findings explicated earlier in the test-treatment pathway (ie from previous evidence in the linkage).

Figure 3 (page 123) of Paper 3 was originally four separate figures. During peer review of this manuscript it was asked that these figures be amalgamated into one decision algorithm. This was done for the published paper but I still think it is easier for the reader to understand how to apply the decision framework, if each scenario is considered separately. I have presented each of the three scenarios separately below (but based on the published decision algorithm) to provide added clarity.

FIGURE 4 OPTIMISATION SCENARIO – TEST AS ACCURATE


Evidence Linkage 1 - Core
Evidence of comparative test accuracy
Comparative assessment of test invasiveness & safety considerations
OPTIMISATION
DISEASE SPECTRUM CHANGE
TRADE-OFF
Test as accurate
Not as safe?
NET HARM
As safe? = potential alternative test
Safer? = potential replacement test
Evidence Linkage 2 – Patient Management
No change in management
NO ADDED BENEFIT
STOP
Impact on diagnostic and treatment strategy OR impact uncertain
STOP
Evidence Linkage 3 – Treatment Effectiveness

- In the Optimisation Scenario, evidence linkage 1 is initially compiled – namely, evidence of the accuracy of the test and of the comparative invasiveness and safety of the test.

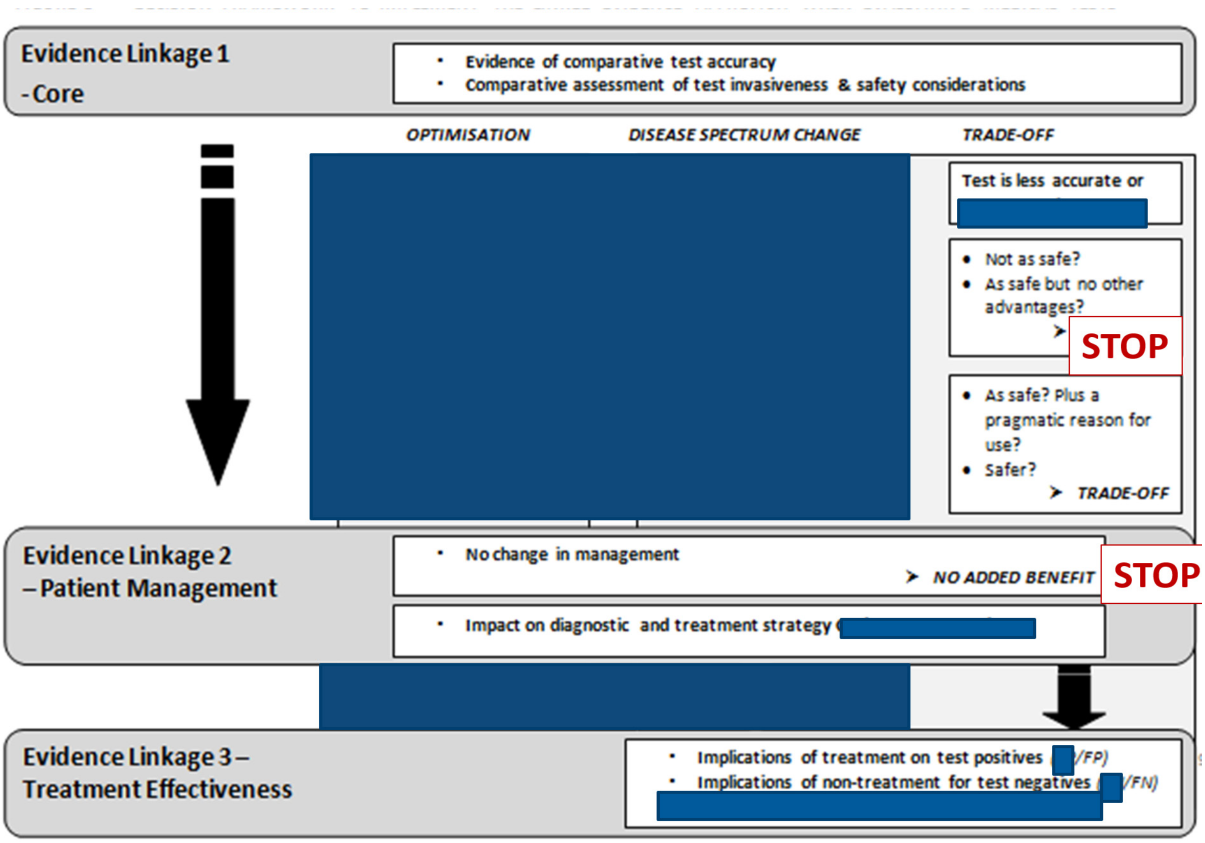
- Then, if the test is as accurate as the comparator test but not as safe, the balance of benefits and harms would be considered a net harm; any additional evidence to inform the policy-maker (including cost information) is likely to be superfluous.

- If the test is as accurate as the comparator test and as safe, an assessment of the impact of the new test on patient management is recommended (evidence linkage 2). If the test does not impact on the management of a patient then the evidence collation should cease. A test is not warranted if it does not affect medical decision-making. However, if the test does impact on decision-making it is particularly important to determine whether the test has any advantages over its comparator in terms of utilisation and cost implications.

- As the spectrum of disease in patients receiving these tests is unlikely to differ from that in the existing treated population (given test accuracy is similar), a review of treatment effectiveness (evidence linkage 3) is not required. The overall treatment benefits and harms are unlikely to differ.

- If there happens to be safety and/or accessibility benefits with the new test, the management and treatment of tested patients is likely to be optimised.
- If there are concerns that these safety and/or accessibility benefits might encourage some patients to receive the new test that would not have received the current test, and therefore change the type of population receiving the treatment (ie the spectrum of disease is different) then the Disease Spectrum Change scenario (Figure 7, page 136) should be addressed.

FIGURE 5 TRADE-OFF SCENARIO – POORER ACCURACY

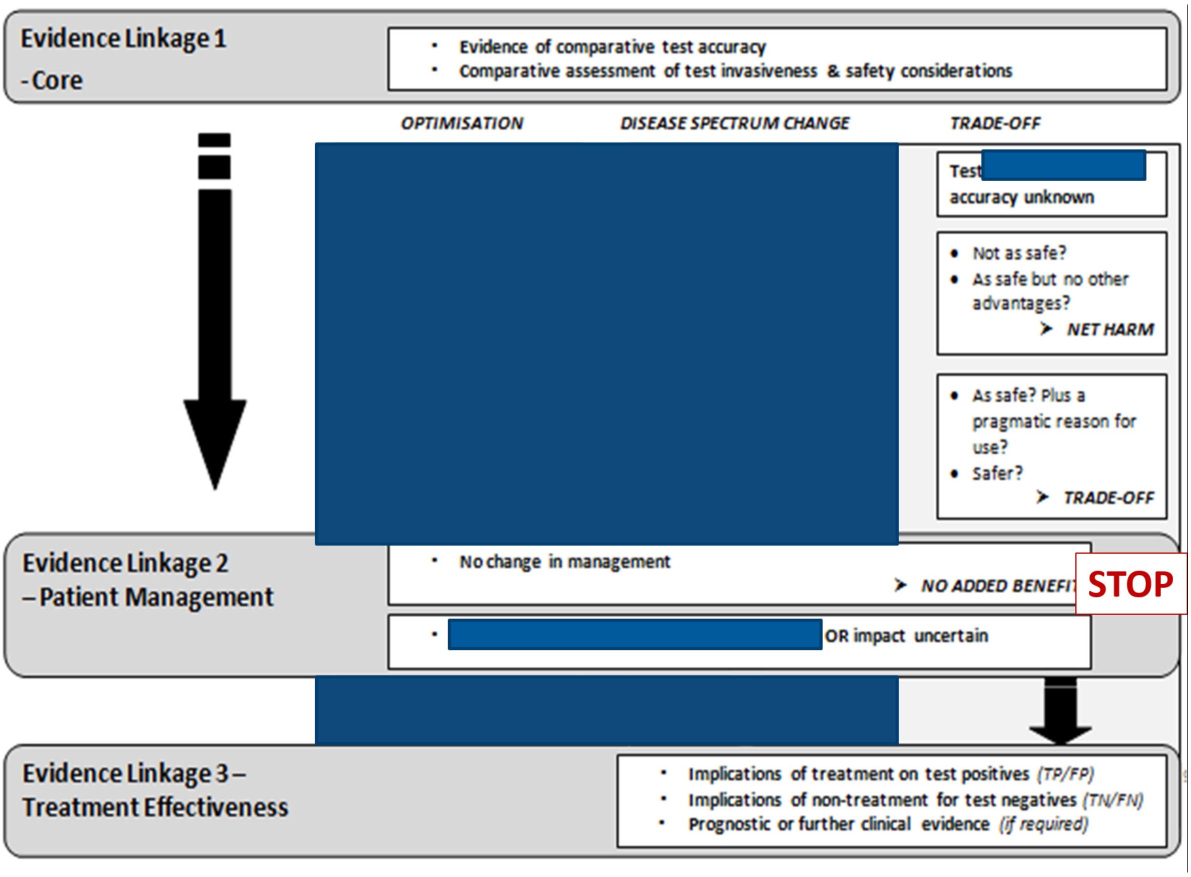

Evidence Linkage 1
- Core
Evidence of comparative test accuracy
Comparative assessment of test invasiveness & safety considerations
OPTIMISATION
DISEASE SPECTRUM CHANGE
TRADE-OFF
Test is less accurate or
Not as safe?
As safe but no other advantages?
STOP
As safe? Plus a pragmatic reason for use?
Safer?
TRADE-OFF
Evidence Linkage 2
– Patient Management
No change in management
NO ADDED BENEFIT
STOP
Impact on diagnostic and treatment strategy
Evidence Linkage 3 –
Treatment Effectiveness
Implications of treatment on test positives
/FP)
Implications of non-treatment for test negatives
/FN)

- In the Trade-Off Scenario, evidence linkage 1 is collated to determine the accuracy of the test and of the comparative invasiveness and safety of the test.
- Then, if the test is less accurate than the comparator test, an assessment of test invasiveness and safety is undertaken to determine whether there is a net harm or a trade-off between safety and test performance.
- Irrespective of this trade-off, if the test is unlikely to change patient management (evidence linkage 2) then the evidence collation should cease. A test is not warranted if it does not affect medical decision-making. However, if the test does impact on

decision-making it is particularly important to determine whether the test has any advantages over its comparator in terms of utilisation and cost.

- The trade-off analysis needs to determine the consequences of treating or not treating, respectively, the likely increase (due to poorer test accuracy) in false positive (FP) or false negative (FN) diagnoses. This is a partial analysis of evidence linkage 3.
- Treatment options for patients with a true positive (TP) or true negative (TN) diagnosis are unlikely to change as a consequence of the test and so do not need formal assessment.

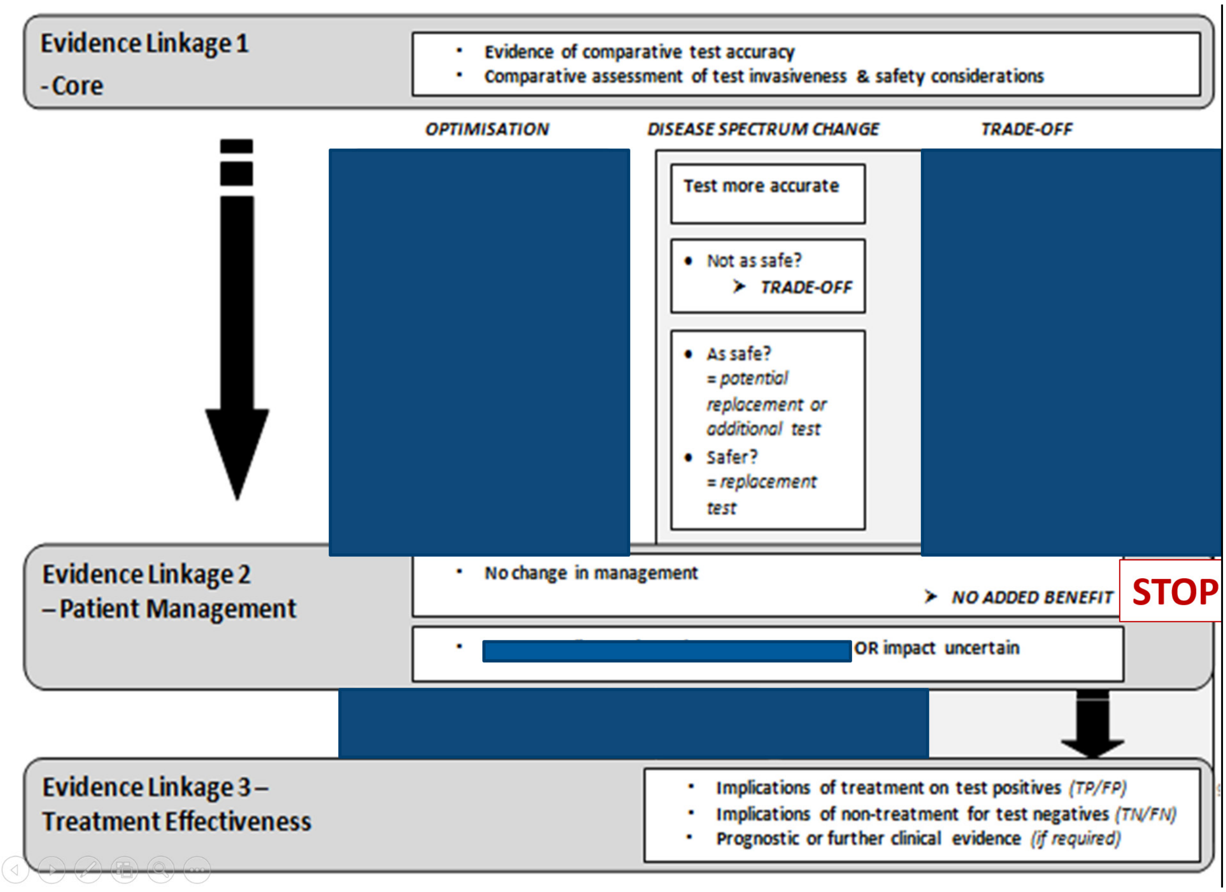
FIGURE 6 IMPERFECT REFERENCE STANDARD – UNCERTAIN ACCURACY


Evidence Linkage 1
- Core
• Evidence of comparative test accuracy
• Comparative assessment of test invasiveness & safety considerations
OPTIMISATION
DISEASE SPECTRUM CHANGE
TRADE-OFF
[redacted]
Test [redacted] accuracy unknown
• Not as safe?
• As safe but no other advantages?
➤ NET HARM
• As safe? Plus a pragmatic reason for use?
• Safer?
➤ TRADE-OFF
Evidence Linkage 2
– Patient Management
• No change in management
➤ NO ADDED BENEFI
STOP
• [redacted] OR impact uncertain
Evidence Linkage 3 –
Treatment Effectiveness
• Implications of treatment on test positives (TP/FP)
• Implications of non-treatment for test negatives (TN/FN)
• Prognostic or further clinical evidence (if required)

- In this scenario, evidence linkage 1 is addressed ie evidence of test performance is collated and data on the comparative invasiveness and safety of the test are obtained.

- When it is impossible to determine test accuracy (eg due to an imperfect reference standard) a conservative approach is needed to determine all the possible consequences of testing.
- Firstly, it needs to be considered whether the test is likely to change patient management (evidence linkage 2). Even with the uncertainty regarding the performance of the new test, if it is apparent that decision-making regarding patient care will not change as a consequence of the information provided by the test, then the evidence collation should cease. However, if the test is likely to impact on decision-making it is important to determine whether the test has any advantages over its comparator in terms of utilisation and cost implications.
- Secondly, the implications of false negatives and positives need to be explored, as well as, conversely, the potential to uncover a spectrum of disease (ie different types of true positives and true negatives) for which the natural history and the impact of treatment is largely unknown. This is evidence linkage 3.
- Sequential linkages of evidence are required to build a picture of the overall clinical effectiveness of the test. With each linkage in the synthesis, the uncertainty regarding the transferability between these multiple linkages is increased.

FIGURE 7 DISEASE SPECTRUM CHANGE SCENARIO – MORE ACCURATE


Evidence Linkage 1 - Core
Evidence of comparative test accuracy
Comparative assessment of test invasiveness & safety considerations
OPTIMISATION
DISEASE SPECTRUM CHANGE
TRADE-OFF
Test more accurate
Not as safe?
TRADE-OFF
As safe? = potential replacement or additional test
Safer? = replacement test
Evidence Linkage 2 – Patient Management
No change in management
NO ADDED BENEFIT
STOP
[redacted] OR impact uncertain
Evidence Linkage 3 – Treatment Effectiveness
Implications of treatment on test positives (TP/FP)
Implications of non-treatment for test negatives (TN/FN)
Prognostic or further clinical evidence (if required)

- In the Disease Spectrum Change Scenario, evidence linkage 1 is addressed by compiling evidence on the accuracy of the test and on its comparative invasiveness and safety.

- Then, if the test is more accurate than the comparator (relative to the reference standard) an assessment of test invasiveness and safety is undertaken to determine whether there is a trade-off between safety and test performance – in which case the Trade-Off Scenario (Figure 5) needs to be addressed – or there is a net benefit.

- If there is a net clinical benefit and the new test is a potential alternative, additional or replacement test to the comparator, then an assessment of whether the new test will impact on clinical decision-making needs to be undertaken (evidence linkage 2). If it is apparent that decision-making regarding patient care will not change as a consequence of the information provided by the test, then the evidence collation should cease. However, if the test is likely to impact on decision-making it is important to determine whether the test has any advantages over its comparator in terms of cost and utilisation.

- Of all the scenarios, the one where a randomised controlled trial is most needed is when the new test proves to be more accurate than the comparator test. In the absence of this information, evidence linkage 3 needs to be evaluated comprehensively.
    - If the test is more sensitive than the comparator, prognostic or clinical evidence is needed to determine treatment effectiveness in patients diagnosed with the new test. Evidence is needed on the impact of early versus delayed treatment to determine if there are benefits associated with the reduction in false negatives.
    - If the test is more specific, prognostic or clinical evidence is needed to determine if there are better health outcomes in true negatives. Evidence is also needed on the consequences of inappropriate treatment of false positives to determine if there are benefits associated with the reduction in false positives.
- Sequential linkages of evidence are required to build a picture of the overall clinical effectiveness of the test. With each linkage in the synthesis, the uncertainty regarding the transferability between these multiple linkages is increased.

The use of this decision framework may provide more consistent application of LEA in the evaluation of tests being considered for public funding, ensure that all the relevant questions associated with a test's utility are addressed and, perhaps, result in more consistent decisions.

The decision framework is not restricted to a single new test versus a single comparator test, but – as is the case with 'add on' tests - can be used to compare a new test strategy with an existing test strategy. In this instance, the same scenarios would apply. For example, in the case of an 'add on' test to the existing test strategy, if the new test strategy proved to be more accurate and as safe as the existing test strategy, then the disease spectrum change scenario would apply. There would be a consequent need to determine the impact of treating the newly identified test positives, as well as the impact of not treating the reduced number of test negatives.

The scenarios described above help to identify the strengths and weaknesses of LEA and to address specific methodological limitations. It was important to explore the strengths and limitations, so that LEA could be adapted to a new type of testing situation – that is, the situation where the link between test and treatment is highly co-dependent. The lessons learned in applying LEA to different types of tests and testing situations (ie the decision framework) then informed the extension of the method to the evaluation of companion genetic tests that target a tailored pharmaceutical treatment. This adaptation of LEA to 'personalised medicines' is discussed in Chapter 6.

CHAPTER 6

NOVEL APPLICATION OF THE LINKED EVIDENCE APPROACH

In the last few years there have been increasing requests for the public funding of co-dependent technologies, in particular personalised medicines. Personalised medicine has been defined as

> '*…the use of genetic or other biomarker information to improve the safety, effectiveness, and health outcomes of patients via more efficiently targeted risk stratification, prevention, and tailored medication and treatment-management approaches*' (Faulkner et al. 2012).

These medicines target a group of patients that can only be identified with the help of a companion diagnostic test. This test isolates a particular genetic or other biomarker that predicts whether the medicine is likely to be beneficial to the patient.

Up until 2010, no-one had developed a formalised methodological framework for evaluating these pharmacogenetic technologies (drug and test) for reimbursement or public funding purposes. This may be due to:

- many countries having limited experience with evaluating the safety, effectiveness and cost-effectiveness of the medical test component of the technology pairing,
- the HTA processes used in some countries are not integrated for the evaluation of both test and drug, or
- the available research on these co-dependent technologies is often unsuited or inadequate for the systems developed to evaluate new technologies for reimbursement decisions (Faulkner et al. 2012).

There are other types of co-dependent technologies – for example a device and a pharmaceutical, like drug-eluting stents – and all test-treatment combinations have some level of co-dependency because most tests are undertaken in order to inform a treatment decision. However, the high level of co-dependency in pharmacogenetic technologies was particularly interesting from a methodological perspective. How could this technology pairing be evaluated for a reimbursement decision?

In Australia we have a robust process for evaluating medical tests, as well as an established approach to the evaluation of pharmaceuticals but until 2010 the processes were not integrated. The evaluation methods used within each HTA ‘silo’ were quite different. In 2010 I worked with the Australian Government to develop a process for integrating the decision-making of these ‘silos’ so that both PBAC and MSAC decision-making concerning the drug and companion test were consistent and congruent (Australian Government Department of Health and Ageing 2010). Part of this involved developing an outline of what type of information would be required to inform a decision from both of the Australian decision-making committees should there be an application for the public funding of a pharmacogenetic technology.[11] These requests for information were based on research conducted for this PhD to determine whether the linked evidence approach had broader relevance for all types of medical tests and testing situations, including the companion tests used to target personalised medicines.

Relevant research question

4. *Can the linked evidence approach (LEA) be feasibly adapted to the evaluation of personalised medicines ie the use of a genetic test to target a pharmaceutical treatment?*

PROBLEM IDENTIFIED AND ADDRESSED IN THE PEER-REVIEWED PUBLICATION

Can LEA be applied to companion diagnostic tests? If so, does the linked evidence approach need further modification to inform public funding decisions when two (or more) technologies work together in an integrated fashion to impact on the health of a patient?

Paper 4 (page 151) describes the development of an analytic framework incorporating LEA methodology that was formulated to assess the safety, effectiveness and cost-effectiveness

[11] http://www.health.gov.au/internet/hta/publishing.nsf/Content/14B1C87A7C197EE6CA2577A00012C52D/$File/codependents.pdf

of personalised medicines to inform government subsidy decisions. The research was done in order to solve an immediate problem for the Australian federal government. It was receiving requests from industry applicants to have specific personalised medicines funded under the Medicare Benefits Schedule (for the biomarker test) and the Pharmaceutical Benefits Schedule (for the drug) but had no method or process by which this could occur. As I had already proposed a research protocol for investigating and evaluating 'co-dependent' pharmacogenetic technologies some six months' earlier, I was asked by the government whether I could work with them to address this problem. The research protocol that had been developed was implemented and a method for evaluating these technologies was developed. Paper 4 below describes the research that was undertaken.

Statement of authorship

Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing codependent technologies. *Medical Decision Making*, April 2013; 33(3):333-342. doi: 10.1177/0272989X12452341. Available at: http://mdm.sagepub.com/content/33/3/333

AUTHORS' CONTRIBUTIONS

Tracy Merlin (Candidate)

Conceived the research project, wrote the research concept plan and revised the proposal to apply/adapt it to a policy process, drafted the data extraction template, undertook data extraction and interpreted the results, drafted the evaluation framework and incorporated feedback from co-authors, policy-makers and public submissions. Drafted the manuscript and incorporated feedback on the manuscript from co-authors and peer-reviewers.

Claude Farah

Provided input to the data extraction template, undertook data extraction, provided feedback on the draft evaluation framework, and approved the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Camille Schubert

Provided input to the data extraction template, undertook data extraction, provided feedback on the draft evaluation framework, and approved the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Andrew Mitchell

Provided the opportunity to apply and adapt the research project to a policy process and to receive input from policy makers and the public; provided input on the draft evaluation framework and contributed to the interpretation of results; provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Janet Hiller

Provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Philip Ryan

Provided feedback on the manuscript. I give consent for Tracy Merlin to present this paper for examination towards the Doctor of Philosophy.

Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing co-dependent technologies. *Medical Decision Making*, April 2013; 33(3):333-342,

doi: 10.1177/0272989X12452341.

Available at: http://mdm.sagepub.com/content/33/3/333

Paper 4 is reproduced as follows –

Assessing personalised medicines in Australia: A national framework for reviewing co-dependent technologies

Running head: Assessing personalised medicine for reimbursement[12]

Tracy Merlin MPH[1], Claude Farah MMedSci[1], Camille Schubert BEc[1], Andrew Mitchell MMedSci,[2] Janet E Hiller PhD[3] and Philip Ryan MBBS[4]

[1] Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, School of Population Health and Clinical Practice, University of Adelaide, Adelaide, South Australia, Australia

[2] Australian Government Department of Health and Ageing, Canberra, Australian Capital Territory, Australia

[3] Faculty of Health Sciences, Australian Catholic University, Fitzroy, Victoria, Australia; and School of Population Health and Clinical Practice, University of Adelaide, Adelaide, South Australia, Australia

[4] Data Management & Analysis Centre (DMAC), Discipline of Public Health, School of Population Health and Clinical Practice, University of Adelaide, Adelaide, South Australia, Australia

Text Word Count: 4178 words

[12] Financial support for this project was provided in part by a contract with the Australian Government Department of Health and Ageing. The funding agreement ensured the authors' independence in designing the framework, interpreting the data, writing and publishing the report. The following author is employed by the sponsor: Andrew Mitchell.

Corresponding author contact details / Address for reprint requests:

Tracy Merlin, BA(Hons), MPH

Managing Director, Adelaide Health Technology Assessment (AHTA)

Senior Lecturer, Discipline of Public Health, Mail Drop DX 650 545,

School of Population Health and Clinical Practice, University of Adelaide,

Adelaide, South Australia, 5005, Australia.

Tel: 61-8-8313 3575, Fax: 61-8-8313 6899, Email: tracy.merlin@adelaide.edu.au

ABSTRACT

Background: Since the mapping of the human genome in 2003, the development of biomarker targeted therapy and clinical adoption of 'personalised medicine' has accelerated. Models for insurance subsidy of biomarker/test/drug packages ('co-dependent technologies' or technologies that work better together) are not well-developed. Our aim was to create a framework to assess the safety, effectiveness and cost-effectiveness of these technologies for a national coverage or reimbursement decision.

Methods: We extracted information from assessments of recent Australian reimbursement applications that concerned genetic tests and treatments to identify items and evidence gaps considered important to the decision-making process. Relevant international regulatory and reimbursement guidance documents were also reviewed. Items addressing causality theory were included to help explain the relationship between biomarker and treatment. The framework was reviewed by policy-makers and technical experts, prior to a public consultation process.

Results: The framework consists of five components – context, clinical benefit, evidence translation, cost-effectiveness, and financial impact – and a checklist of 79 items. To determine whether the biomarker test, the drug, both or neither should be subsidised, we considered it crucial to identify whether the biomarker is a treatment effect modifier or a prognostic factor. To aid in this determination, the framework explicitly allows the linkage of different types of evidence to examine whether targeting the biomarker varies the likely clinical benefit of the drug, and if so, to what extent.

Conclusions: The first national framework to assess personalised medicine for coverage or reimbursement decisions has been developed and introduced, and may be a suitable model for other health systems.

Abstract word count: 254

INTRODUCTION

Until recently health professionals have had limited information about the likely response of a patient to therapy. Treatment strategies were generally based on aggregated information and subsequently modified according to individual response. With increased understanding of genetics, it is now possible to personalise medicine so that the risk profile of a patient can be determined prospectively to guide treatment so that it is more effective from initiation, is only used in those who will respond, and/or with fewer side effects [1-4].

Several drugs, particularly for cancer, have been developed and marketed with a 'companion diagnostic' - a test to determine whether a patient has a biomarker that will predict response to a drug [5-6]. Examples include trastuzumab and HER2 testing for breast cancer; cetuximab and K-RAS mutation testing for metastatic colorectal cancer; and gefitinib and EGFR testing for lung cancer [7-10]. Such treatment is potentially more clinically and cost-effective as it only targets patients likely to respond [11-13].

The US Federal Drug Administration (FDA) has made preliminary efforts to provide guidance on prospective, scientifically robust co-development of a drug and companion diagnostic [5] and there is growing international discussion investigating ways of dealing with these co-dependent technologies from an assessment and reimbursement perspective [14-16]. However, there is also growing frustration from industry and health professionals that personalised medicine is not living up to its promise [17], partly because the current models of assessment internationally are inadequate to inform coverage or reimbursement decisions regarding these distinctive technologies [18].

Both the treatment and companion diagnostic in a personalised medicine need to be assessed for performance in order to make a coverage or reimbursement decision. This is not a straightforward process [16, 19]. In order to determine what factors influence these decisions, Meckley and Neumann (2010) selected six personalised medicine case studies and extracted data on the quality of evidence supporting each case study, type of regulatory oversight each received, whether clinical guidelines supported the technology, and whether the technology had been found to be cost-effective [18]. They noted there was poor evidentiary support - in the form of randomised controlled trials assessing the *direct impact of testing* on health outcomes - for most of these technologies and that the key factor influencing a positive reimbursement decision appeared to be the strength of the evidence base.

The recent *Review of Health Technology Assessment in Australia* similarly recognised that co-dependent technologies (or technologies that work better together), such as personalised medicines, are problematic to assess for reimbursement decisions [20]. As a consequence, research was undertaken to develop an assessment framework to assist policy-makers to make evidence-based decisions about subsidised access to these emerging technologies.

Three objectives were formulated to ensure that the assessment framework was feasible:

1. To identify the different decision-making scenarios that would apply specifically to a personalised medicine ie targeting drug therapy on the basis of a biomarker;
2. To identify the criteria needed to inform an assessment of these technologies; and
3. To formulate an approach that recognises the scarcity of direct evidence, ie randomised trials assessing the impact on health outcomes of testing versus no testing for the biomarker to guide treatment with the new drug.

METHODS

1st Stage

Five co-dependent technologies that had previously been assessed for coverage or reimbursement decisions were reviewed (Table 1):

1. EGFR/gefitinib for non small cell lung cancer;
2. K-RAS/cetuximab for metastatic colorectal cancer;
3. K-RAS/panitumumab for metastatic colorectal cancer;
4. PDGFR rearrangements/imatinib for primary or secondary clonal eosinophilia (systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia) [21]; and
5. KIT D816V/imatinib for aggressive systemic mast cell disease without eosinophilia [21].

These case studies were selected as they were the most recent co-dependent technologies to be assessed for a reimbursement decision by our national committees (either for the test or drug). In all cases the drug was considered for reimbursement prior to consideration of the biomarker test. Three of the five source documents were only available as commercial-in-confidence.

The information provided in each independent assessment report on these five technology applications was categorised and tabulated. 67 information items were identified as being present in at least one of the five applications. A gap analysis was conducted for each personalised medicine across the 67 items to determine what key information was considered absent on the basis of (a) matters raised within the assessment report[13] and (b) matters raised during the appraisal and decision-making process[14]. Each of the five personalised medicines was independently rated by three experienced evaluators of reimbursement applications, in terms of whether the 67 information items were provided in the application (yes, no, partially) and whether or not the information was needed (yes, no, not applicable). A free text column was used to comment on whether difficulties were likely to arise when reviewing an item.

It was noted that reimbursement was more likely when there were fewer evidence gaps present in the application and when the evidence was of better quality (Table 1). This latter finding is consistent with Meckley and Neuman (2010). However, given that in both Meckley and Neuman's case studies and in our case studies there was a lack of direct randomised controlled trial evidence of the biomarker test impact on patient health outcomes, it was thought that a framework that allowed the linkage of different types of evidence to support a claim for reimbursement might provide policy-makers with fewer evidence gaps, and thus reduce decision-making uncertainty.

[13] mentioned in the independent assessment report (Commentary) of an applicant's submission undertaken on behalf of the Pharmaceutical Benefits Advisory Committee (PBAC) or discussed in the independent assessment report undertaken on behalf of the Medical Services Advisory Committee (MSAC)

[14] relevant MSAC or PBAC meeting minutes or formal advice from the Economics Subcommittee of PBAC – both MSAC and PBAC make decisions regarding whether reimbursement is warranted for, respectively, new medical services (including diagnostics, devices and procedures) and pharmaceutical medicines in Australia

TABLE 1 CASE STUDIES OF PHARMACOGENETIC CO-DEPENDENT TECHNOLOGIES

Case study **(biomarker/ therapy)**	**Decision-making body [therapeutic purpose]**	**Evidence quality**	**Evidence gaps** (N=67 information items)[a]	**Test reimbursed?**[b]	**Drug reimbursed?**[b]
EGFR/gefitinib for non small cell lung cancer (2nd line)	PBAC [targeted treatment]	No direct evidence Linked evidence - moderate quality	8/67 (12%)	Not considered	Yes
K-RAS/ cetuximab for metastatic colorectal cancer (1st line)	PBAC [targeted treatment]	No direct evidence Linked evidence - poor quality	32/67 (48%)	Not considered	No
K-RAS/ panitumumab for metastatic colorectal cancer (2nd line)	PBAC [targeted treatment]	No direct evidence Linked evidence - poor quality	21/67 (31%)	Not considered	No
PDGFR re-arrangements/ imatinib for primary or secondary clonal eosinophilia[c]	MSAC [targeted treatment]	Direct evidence – poor quality *Plus* Linked evidence – moderate quality	3/67 (4%)	Yes	Yes
KIT D816V/ imatinib for aggressive systemic mast cell disease without eosinophilia (2nd line)	MSAC [rule out imatinib treatment]	Direct evidence – poor quality *Plus* Linked evidence – moderate quality	4/67 (6%)	No[d]	Yes

[a] 67 information items (denominator) were collated from submissions at the completion of Stage 1. Evidence gaps (numerator) were defined as a complete absence of information in the submission; however, please note that frequently the information items were only partially/inadequately addressed and in some instances items were not applicable; [b] decision at the time the framework was being developed; [c] systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia; [d] PDGFR rearrangements and the KIT D816V mutation are mutually exclusive so, as the PDGFR test was funded, there was no need to fund the KIT D816V test.

PBAC = Pharmaceutical Benefits Advisory Committee; MSAC = Medical Services Advisory Committee

2nd Stage

In order to ensure that the linkage of evidence was done rigorously, the relationship between biomarker status (as identified by a test) and drug treatment outcomes needed to be adequately explained. The collated list of items was cross-checked against Bradford Hill causality theory [22] to ensure that there were multiple opportunities to explain the association between biomarker and drug treatment outcomes, even in the absence of generally accepted experimental evidence. The Bradford Hill criteria, namely strength, specificity and temporality of the association between the biomarker and drug treatment on health outcomes; consistency and coherence of effect; biological plausability and gradient (eg dose-response); producing the effect upon experimentation or by analogy, were addressed and five additional items were included in the checklist. The list of items was then structured in a format consistent with that used for assessing pharmaceuticals for reimbursement decisions in Australia [23].

Currently available international guidance documents, for the appraisal of technologies or appraisal of applications for test/drug reimbursement, were reviewed to determine whether any further items would be relevant to the framework's development. To identify this literature, Embase and Medline were canvassed, along with internet searches of regulatory and reimbursement agency websites and the health technology assessment database (http://www.crd.york.ac.uk). No new items were identified in the international literature, although some of the documents provided detail that was considered useful as explanatory material in the framework.

3rd Stage

Feedback was sought on the framework from a Steering Committee comprising Chairs and members of the two committees responsible for funding decisions for new technologies[15] in Australia, as well as representatives of the funder (government).

The structure of the framework was considered by the Steering Committee to be consistent with the information needed to make reimbursement decisions. Committee members chose not to prioritise any of the 72 items as all were considered important. The following

[15] Medical Services Advisory Committee (MSAC) and Pharmaceutical Benefits Advisory Committee (PBAC)

amendments were suggested by Committee members and incorporated into the framework: a more precise definition of the biomarker; explanatory detail regarding the proposed test, including a new checklist item on the proposed Medicare descriptor for the test; a new item on the need for testing for new somatic mutations following treatment; a new item on the method and timing of specimen retrieval; a new item concerning the analytic validity of the test; re-ordering items and in some cases collapsing items that contained similar concepts or splitting items that contained multiple concepts; and modifying wording of items or explanations to make it clearer regarding the scope or purpose of some of the items.

4th Stage

On the basis of feedback from the Steering Committee, a finalised assessment framework containing a checklist of 79 items and explanatory material was developed and released for public consultation between 16th September and 17th December 2010[16].

Twelve submissions were received through public consultation. Suggestions were made to extend the scope of the framework in the future and to clarify the government administrative processes for co-dependent technology applications. The key concern specific to the framework was whether the requested evidence could be feasibly provided, particularly for co-dependent technologies targeting rare diseases. In response to this feedback, examples were included in the framework (see Additional File 2) to make it more explicit that the linkage of different types of relevant evidence was encouraged when there were deficiencies in an experimental evidence base. Similarly, guidelines are being produced that will further explain each of the concepts in the framework; and a government process for case managing co-dependent technology applications has been developed.[17]

[16] http://www.health.gov.au/internet/hta/publishing.nsf/Content/whats-new

[17] Summarised at http://www.health.gov.au/internet/hta/publishing.nsf/Content/co-1

RESULTS

Decision-making scenarios

In Australia, both clinical and cost-effectiveness (ie value for money[18]) are considered as part of reimbursement decision-making. Regardless of whether decision-making occurs on the basis of clinical effectiveness or cost-effectiveness, four distinct scenarios can arise when assessing a co-dependent technology:

1. the drug is (cost-)effective in an untested population, but (cost-)ineffective when conditioned on biomarker status as identified by the test. This might occur if: the biomarker does not explain the variation in treatment effect; other prognostic factors are more important in terms of the drug's effect than the identified biomarker; or if the biomarker is highly prevalent, testing may be considered an unnecessary expense. In this scenario the drug is reimbursed but not the test;

2. the drug is (cost-)effective in an untested population, but more (cost-)effective when conditioned on the biomarker identified by the test. In this scenario the drug is reimbursed but the decision to reimburse the test will depend on the level of uncertainty surrounding the relationship between biomarker and the treatment effect of the drug;

3. the drug is not (cost-)effective in an untested population but is (cost-)effective when conditioned on biomarker status as identified by the test. In this scenario, reimbursement of both test and drug will depend on the level of uncertainty surrounding the relationship between the biomarker and the treatment effect of the drug; and

4. the drug is not (cost-)effective in either an untested or tested population. In this scenario neither the drug nor the test is subsidised.

[18] the incremental cost of the new test/treatment strategy over the current test/treatment strategy relative to the incremental health outcomes gained. This incremental cost-effectiveness ratio (ICER) is assessed for each health intervention submitted for reimbursement in Australia but decisions regarding the value for money of the ICER are determined on a case-by-case basis. No specific willingness-to-pay threshold is used in Australia.

Possible applications of the assessment framework

The framework was developed to assess a new personalised medicine in the first instance (prototype situation). However, it was recognised that reimbursement of a drug and its companion test may not occur contiguously, nor would the test and drug be necessarily submitted for funding by the same applicant, in which case the framework needed to be sufficiently flexible to address different reimbursement situations (Table 2). These "situations" are described in more detail below.

TABLE 2 REIMBURSEMENT SITUATIONS REQUIRING DIFFERENT APPLICATIONS OF THE ASSESSMENT FRAMEWORK

Reimbursement situation	Biomarker*	Test	Drug
Prototype situation (See Additional file 1)	Probable new marker	New reimbursement application	New reimbursement application
Situation I	Valid marker	Currently reimbursed	New reimbursement application
Situation II	Valid marker	New reimbursement application	Currently reimbursed
Situation III	Valid marker	New reimbursement application	New reimbursement application
Situation IV	Group of markers	Currently reimbursed ± new reimbursement application	Currently reimbursed ± new reimbursement application

* FDA categorises biomarkers according to "probable" and "valid" [24-25].

Prototype situation

The framework for assessing personalised medicines consists of five domains and a checklist of 79 items (see Additional File 1). Section A provides the rationale for the co-dependent relationship between biomarker test and drug; Section B provides the supporting evidence of clinical benefit (in a manner that allows the linkage of different types of evidence when direct evidence is not available, see also Additional File 2); Section C outlines how the evidence of clinical benefit can be translated to the local setting; Section D provides the economic model incorporating clinical and cost data for the biomarker test and drug, and for the drug without use of the test; and Section E describes the financial or budgetary impact of funding both test and drug.

Extensions of the framework

In addition to the situation where a new test and drug are being submitted for coverage or reimbursement in the context of an as yet unproven biomarker, four other situations were identified.

Situation I. When a new drug is submitted for reimbursement for targeting a previously established (valid) [24-25] biomarker using a test that is currently reimbursed, the aim is to discriminate the superior (or non-inferior) treatment effect of the drug alone. It would be inefficient to address all of the 79 items for this new drug so only items that address specific areas of uncertainty would require assessment. This would mean that some basic information regarding the previous co-dependent technology assessment would need to be in the public domain.

Situation II. When a drug and companion test for an established biomarker have been accepted as cost-effective, evaluation of a new test, for the same biomarker, would only require an assessment of the comparative accuracy of the new and old test. If the spectrum of disease identified by the new test in the patient population does not change, supporting evidence of treatment effectiveness would not be required.

Situation III. When the biomarker has been previously assessed but both the proposed test and the proposed drug are new, it is likely that the majority of the checklist items would need to be assessed, although the biomarker's prognostic or predictive impact may not need review.

Situation IV. When a new biomarker (or group of biomarkers) is identified as part of a new application, the aim is to gauge whether this new biomarker(s), when targeted by the drug, results in further improved patient health outcomes. This scenario could encompass the possibility of a new or currently listed drug, as well as a new or currently listed test (a complex scenario). Thus all of the checklist items would need to be assessed.

DISCUSSION

Key considerations when developing the framework

Examination of the clinical effectiveness of a co-dependent technology requires an innovative approach to assessment. Data are needed to support the claim of a relationship between biomarker status and the treatment effect of the drug, primarily because this directly informs the decision to reimburse the test, the drug, both or neither. The biological plausibility of the relationship is essential. Specifically, the causal pathway could suggest that the test is identifying a biomarker that is an independent prognostic factor (prognostic test), treatment effect modifier (predictive test), or both, (see example in Additional File 3) or the relationship is unknown. [26]

A prognostic factor is a risk factor that affects the likely progress of the patient regardless of the particular treatment they are given. [27] If, for example, a biomarker in a tumour sample acts as an independent prognostic factor for an early death from metastatic colorectal cancer, then regardless of the treatment given (ie the new Drug A or the old Drug B), these patients will have a worse prognosis than those without the biomarker. In a reimbursement/policy framework this indicates that the two health technologies (prognostic and therapeutic) have a low level of co-dependency. By identifying those patients with a better (or worse) prognosis, irrespective of treatment, the test may be used to provide more cost-effective targeting of the new drug, but all possible comparator treatments would need to be considered in making a reimbursement decision as they are also likely to be more cost-effective in the identified subgroup. These health technologies may include established treatments that, following a reimbursement decision, are retrospectively targeted to certain patient groups where there will be an optimal effect in terms of toxicity, uptake, effectiveness and cost-effectiveness. Prognostic impact can be distinguished using the study designs described in Figure 1, Figure 3, and to a lesser extent Figure 4, in Additional File 1.

When treatment effect varies according to biomarker status, the drug and test are considered highly co-dependent. Drug A may have been developed specifically to target a biomarker in order to produce a clinical benefit to the patient (eg survival, quality of life, reduced complications).[19] If this successfully predicts a favourable treatment effect then

[19] In some cases, the development of the test and the treatment is a joint enterprise.

among patients with the biomarker those receiving Drug A in addition to Drug B would have better health outcomes than patients receiving Drug B alone, whereas patients without the biomarker will receive the same clinical benefit regardless of whether Drug A is used in addition to Drug B. If Drug A replaces Drug B, then patients without the biomarker would be effectively untreated.

An adequately powered randomised trial that prospectively assesses comparative treatment effect on patient-relevant health outcomes according to subgroups delineated by the biomarker would be ideal to determine whether effect modification is occurring. As prediction of treatment effect variation suggests that there is a unique relationship between the biomarker and drug, reimbursement of both technologies would be considered, particularly if this predicts a qualitative difference (ie better rather than worse or non-inferior) rather than a quantitative difference (ie the extent of effect is improved). Treatment effect modification can be distinguished using the study designs described in Figure 1, Figure 3, and to a lesser extent Figure 4, in Additional File 1.

A comparison with a 'no testing' arm similarly allows the incremental benefit of the biomarker test to be determined ie receiving Drug A when biomarker positive and Drug B when biomarker negative versus Drug A being administered to everyone (see Figure 1 and Figure 2, Additional File 1). This may assist when there is uncertainty as to whether the biomarker explains the differential treatment effect between Drug A and Drug B or whether some other unmeasured variable is responsible**.**

It is helpful to envision the "ideal" randomised controlled trial evidence that would be needed to answer the decision-makers' question as to whether the biomarker test and/or drug should be subsidised. [28] In this case, a double-randomised controlled trial (Figure 1, Additional File 1) may be considered ideal evidence as it addresses each of the biomarker-drug relationship issues described previously. However, the practicalities are such that trials of this design are rarely, if ever, going to be conducted. As the aim of biomarker-targeting is to maximise the therapeutic effect of a drug, and it is more efficient (both financially and logistically ie in terms of sample size requirements) to measure this effect in a biomarker enriched population (particularly when the biomarker is uncommon), it is unlikely that double-randomised controlled trials would be conducted. Similarly, if the new drug therapy is meant to replace an existing therapy in patients that are biomarker positive, rather than to be used in combination with an existing therapy, then it could be considered unethical to conduct a trial where there is a chance that biomarker negative patients in the untested treatment arm would be receiving a new drug that has no effect on them, apart from perhaps an increased risk of adverse events, and be forgoing a known effective treatment.

Given these practical limitations with Level 1 evidence, it was considered reasonable to take a pragmatic approach and allow an applicant for a co-dependent technology to build a chain of argument through the linkage of different types of evidence (“linked evidence approach”) (see Additional File 2; and Option 2, Section B in Additional File 1). The key is to present this linkage so that decision-makers can see that obvious uncertainties have been addressed, that data are defensibly transferrable across different parts of the linkage, and that available evidence for the linkage has been gathered systematically and transparently and has been executed in an internally valid manner, ie the data are not selectively used or affected by bias and confounding.

Other factors need to be considered when assessing a personalised medicine, including: whether the test or drug is additional to the current tests or treatment being received, or replaces them; if, when using a “linked evidence approach”, there is a reference standard for the biomarker test or whether the test itself is proposed as the reference standard [21, 29]; whether testing can be conducted on biopsied tumour samples taken at diagnosis or after first-line treatment, or; whether the method of sample preservation, storage, or previous treatment or instability of the biomarker state over time will affect the accuracy of the test results. Some biomarkers are only identified through the use of multiple tests, and the positive and negative predictive value of these tests will vary according to the prevalence of a biomarker state in the population being tested. Each of these factors has been identified as requiring an answer in the assessment framework.

International context

The implications of poor primary research, and/or poor assessment frameworks to address personalised medicines include: (1) fragmented or poor decision-making as a consequence of considering the drug and biomarker test independently, rather than as an integrated package (see Table 1); (2) poor guidance to trialists and industry regarding appropriate trial design and thus wasted resources in producing and presenting suboptimal evidence to funding agencies; and (3) poor health outcomes for patients as a consequence of receiving ineffective or potentially harmful treatment if a personalised medicine has not been assessed rigorously and yet is reimbursed.

Most countries that ascribe to the evidence-based assessment of technologies in order to make resource and policy decisions have guidance available on evaluating single interventions, such as drugs [23]. Guidance on the assessment of tests has been produced only recently (2008-2011) in the United States[30] [31-32], England [33] and Europe [34]. Australia developed its own guidance for the assessment of diagnostic tests for reimbursement purposes in 2005 [28, 35], proposing a "linked evidence approach" when assessing tests, which has subsequently been recommended in each of the international guidance documents mentioned above. Although there have been recent developments in regulatory policy in the United States to allow *joint approval* of a co-dependent test and drug [32], to our knowledge, no authority to date has developed a system to evaluate a package of co-dependent test and drug technologies for reimbursement purposes. As can be seen from the framework that has been developed there are many domains where both the biomarker test and drug need to be considered together when evaluating their clinical benefit and cost-effectiveness.

Strengths and Limitations of the framework

The assessment framework that has been developed is novel as it tackles the concept of personalised medicine within a coverage or reimbursement context. A formal assessment framework provides clarity for industry, with regard to policy-makers' expectations, and can drive research aimed at addressing these expectations. It also facilitates consistency in decision-making and helps to identify areas of uncertainty for a reimbursement decision. The framework recognises that often the 'ideal' clinical evidence to address decision-makers' questions is not available. This is both a strength and limitation of the framework. It is a strength in that this pragmatic approach allows potentially beneficial medicines to be subsidised, despite deficiencies in the supporting evidence. [36] However, linking evidence from different studies conducted in different populations can never provide evidence about the impact of a new biomarker test and new drug on patient outcomes with the same strength and quality as a double-randomised controlled trial. A trial would capture the entire causal pathway, including the unexpected and unknown effects. [28] The linkage of individual pieces of evidence to estimate the effect of a trial must therefore be applied and interpreted with caution. Identifiable uncertainties or assumptions concerning linkages in the pathway can be explored using decision analytic modelling, but modelling itself may be prone to oversimplification and potential bias.

An area of economic uncertainty for decision-makers is how to allocate value to the components in a co-dependent technology package [37]. Australia assesses value across a number of technologies, including "diagnostics", but in current practice the Australian system is quite passive. Suppliers of the technologies are allowed to set a price for each component reflective of the supplier's notion of value and then decision-makers judge whether the value of the package as a whole is acceptable in terms of incremental cost-effectiveness. This might be problematic when the supplier differs for each of the technologies in a package; or for health systems which actively allocate value but are inflexible in revising this value when two technologies become linked.

Another potential limitation of the assessment framework is that it has yet to be evaluated over the long-term or empirically assessed as to its utility. Applications for personalised medicines have been accepted in Australia, using the framework, since late 2010 via a newly created Health Technology Assessment Access Point (HTAAP). This process case manages a personalised medicine to ensure that each co-dependent technology is appraised by the relevant decision-making committee and that coordinated advice is provided for a reimbursement decision. Applicants to the HTAAP are encouraged to use the framework[20] as it is the backbone upon which more detailed guidelines are being produced on co-dependent technology evaluation. It is also conceptually consistent with the current Australian guidelines for evaluating diagnostic tests and drugs [23, 35].

Australia is a small market in global terms so, currently, four applications have been evaluated since the co-dependent technology framework was drafted, with another four commencing the submission process. The rate of applicants seeking reimbursement of these technologies has increased rapidly, although it is unclear whether this is because there is now a recognised method outlining the type of evidence that policy-makers expect to see or because there have been more personalised medicines getting regulatory approval. Either way, reports suggest that the framework has assisted in providing valuable guidance to the

[20] Described as the *Draft Information Requests for Assessing Co-Dependent Technologies* on the HTAAP site - http://www.health.gov.au/internet/hta/publishing.nsf/Content/co-1, and in the HTAAP information pack for applicants

decision-maker, facilitating efficient processing of a reimbursement decision for both biomarker test and drug.

ACKNOWLEDGEMENTS

Our thanks go to Brian Richards, Graeme Suthers, Brian K Harrison, Robyn Ward, and Lloyd Sansom for providing input on the items described in the framework.

REFERENCES

[1] Finkelstein Y, Bournissen FG, Hutson JR, Shannon M. Polymorphism of the ADRB2 gene and response to inhaled beta- agonists in children with asthma: a meta-analysis. J Asthma. 2009 Nov;46(9):900-5.

[2] Roden D.M., Altman R.B., Benowitz M.D., et al. Pharmacogenomics: challenges and opportunities. Annals of Internal Medicine. 2006;145:749-57.

[3] Wang B, Wang J, Huang SQ, Su HH, Zhou SF. Genetic Polymorphism of the Human Cytochrome P450 2C9 Gene and Its Clinical Significance. Curr Drug Metab. 2009 Sep;10(7):781-834.

[4] Xie H-G, Frueh FW. Pharmacogenomics steps toward personalized medicine. Future Medicine. 2005;2(4):325-37.

[5] Department of Health and Human Services. Drug-diagnostic co-development: concept paper. Food and Drug Administration (FDA) 2005.

[6] Scartozzi M, Bearzi I, Mandolesi A, Pierantoni C, Loupakis F, Zaniboni A, et al. Epidermal Growth Factor Receptor (EGFR) gene copy number (GCN) correlates with clinical activity of irinotecan-cetuximab in K-RAS wild-type colorectal cancer: a fluorescence in situ (FISH) and chromogenic in situ hybridization (CISH) analysis. BMC Cancer. 2009;9:303.

[7] Karapetis CS, Khambata-Ford S, Jonker DJ, O'Callaghan CJ, Tu D, Tebbutt NC, et al. K-ras mutations and benefit from cetuximab in advanced colorectal cancer. N Engl J Med. 2008 Oct 23;359(17):1757-65.

[8] Lurje G, Lenz HJ. EGFR Signaling and Drug Discovery. Oncology. Feb 2;77(6):400-10.

[9] Shak S. Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin Multinational Investigator Study Group. Semin Oncol. 1999 Aug;26(4 Suppl 12):71-7.

[10] Tamura K, Okamoto I, Kashii T, Negoro S, Hirashima T, Kudoh S, et al. Multicentre prospective phase II trial of gefitinib for advanced non-small cell lung cancer with epidermal growth factor receptor mutations: results of the West Japan Thoracic Oncology Group trial (WJTOG0403). Br J Cancer. 2008 Mar 11;98(5):907-14.

[11] Essers BA, Seferina SC, Tjan-Heijnen VC, Severens JL, Novak A, Pompen M, et al. Transferability of Model-Based Economic Evaluations: The Case of Trastuzumab for the Adjuvant Treatment of HER2-Positive Early Breast Cancer in the Netherlands. Value health. 2010 Jan 15;13(4):375-80.

[12] Meckley LM, Gudgeon JM, Anderson JL, Williams MS, Veenstra DL. A policy model to evaluate the benefits, risks and costs of warfarin pharmacogenomic testing. Pharmacoeconomics.28(1):61-74.

[13] Wong W, Carlson J, Thariani R, Veenstra D. Cost Effectiveness of Pharmacogenomics: A Critical and Systematic Review. PharmacoEconomics. 2010;28(11):1001-13.

[14] Conti R, Veenstra DL, Armstrong K, Lesko LJ, Grosse SD. Personalized Medicine and Genomics: Challenges and Opportunities in Assessing Effectiveness, Cost-Effectiveness, and Future Research Priorities. Med Decis Making. 2010 Jan 4;30(3):328-40.

[15] Parliamentary Office of Science and Technology. Postnote: Personalised medicine. London, UK 2009.

[16] Terasawa T, Dahabreh I, Castaldi P, Trikalinos T. Systematic Reviews on Selected Pharmacogenetic Tests for Cancer Treatment: CYP2D6 for Tamoxifen in Breast Cancer, KRAS for anti-EGFR antibodies in Colorectal Cancer, and BCR-ABL1 for Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia. Draft. Rockville, MD: Agency for Healthcare Research and Quality; 2009.

[17] Laksman Z, Detsky AS. Personalized medicine: understanding probabilities and managing expectations. J Gen Intern Med. 2011;26(2):204-6.

[18] Meckley LM, Neumann PJ. Personalized medicine: factors influencing reimbursement. Health Policy. 2010 Feb;94(2):91-100.

[19] Schmitt F. HER2+ breast cancer: how to evaluate? Adv Ther. 2009 Jul;26 Suppl 1:S1-8.

[20] Australian Government Department of Health and Ageing. Review of Health Technology Assessment in Australia - A discussion paper. Canberra, ACT: Commonwealth of Australia 2009.

[21] Buckley E, Merlin T. Molecular testing for the diagnosis of systematic mast cell disease, hypereosinophilic syndromes and chronic eosinophilic leukaemia. *MSAC application 1125b*. Canberra, ACT: Australian Government 2010.

[22] Bradford Hill A. The environment and disease: association or causation? Proceedings of the Royal Society of Medicine. 1965;58:295-300.

[23] Australian Government Department of Health and Ageing. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee. Canberra, ACT: Commonwealth of Australia 2008.

[24] U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry: Pharmacogenomic Data Submissions. Procedural. Rockville, Maryland: FDA 2005.

[25] Food and Drug Administration. Expert Working Group (Efficacy) of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). Guidance for Industry: E15 Definitions for Genomic Biomarkers, Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories. : U.S. Department of Health and Human Services. .

[26] Lee CK, Lord SJ, Coates AS, Simes RJ. Molecular biomarkers to individualise treatment: assessing the evidence. Med J Aust. 2009 Jun 1;190(11):631-6.

[27] Clark GM, Zborowski, D.M., Culbertson, J.L., Whitehead, M., Savoie, M., Seymour, L., Shepherd, F.A.,. Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib. J Thorac Oncol. 2006;1:837-46.

[28] Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making. 2009 Sep-Oct;29(5):E1-E12.

[29] Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. BMC Med Res Methodol. 2009;9:34.

[30] AHRQ. Methods Guide for Medical Test Reviews. Rockville, MD: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; 2010.

[31] Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics. ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing 2010 [Available from: http://www.cdc.gov/genomics/gtesting /ACCE/acce_proj.htm]

[32] Food and Drug Administration (FDA). In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff. In: U.S. Department of Health and Human Services, ed. Rockville, Maryland: U.S. Department of Health and Human Services 2011.

[33] NICE. Interim methods statement (pilot). Version 8. London: National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme; 2010.

[34] EUnetHTA. HTA Core Model for Diagnostic Technologies. Work Package 4: European Network for Health Technology Assessment; 2008.

[35] Medical Services Advisory Committee (MSAC). Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia; 2005.

[36] Steinberg EP, Tunis S, Shapiro D. Insurance coverage for experimental technologies. Health Affairs. 1995;14:143-58.

[37] Garrison LP, Austin MJF. The Economics of Personalized Medicine: A Model of Incentives for Value Creation and Capture. Drug Information Journal. 2007;41:501-9.

Framework for evaluating co-dependent technologies for a reimbursement decision

This framework is slightly modified from the version published for consultation[21]. It concentrates on the pairing of a proposed test with a proposed drug with reference to a specific genetic biomarker. The context is the Australian health system[22] however the questions are likely to be adaptable to any context where reimbursement of a personalised medicine is being considered.

The left hand column of the table below presents a sequence of information requests which are intended to meet the evidentiary requirements of policy makers when assessing a co-dependent technology (drug and biomarker test) for a reimbursement decision. The right hand column comments on and extends each information request in various ways to help interpret what is meant by the information request.

An initial indication is given as to whether each request mainly relates to the test (T), the drug (D) and/or the overlap (O) between them.

The following schematic outlines the *general* flow of information requests:

Section A [1-19]→ Section B [20]

→ Option 1 [21-24]→ Section C [40-42]→ Section D [43-71]→ Section E [72-79]

→ Option 2 [25-39]→ Section C [40-42]→ Section D [43-71]→ Section E [72-79]

Often, however, listing circumstances will vary between co-dependent technologies. The information required to reduce decision-maker uncertainty will vary as a consequence.

It is planned that this initial set of information requests will be further developed into more comprehensive guidelines as experience with the new process grows.

[21] http://www.health.gov.au/internet/hta/publishing.nsf/Content/whats-new

[22] Hence, reference to the Pharmaceutical Benefits Advisory Committee (PBAC) and Medical Services Advisory Committee (MSAC) – both responsible for reimbursement decisions in Australia, as well as the Therapeutic Goods Administration which is the regulatory agency in Australia

Section A

Context for the submission

1) (T) **Who is the test sponsor?**	*Identify the source of the test (e.g. commercial sponsor, research laboratory, widespread pathology practice). This includes clinical sponsors of tests, given that tests not only guide the initiation of therapy but also the cessation of therapy.*
2) (D) **Who is the drug sponsor?**	*This enables a different sponsor to be identified if necessary for each component of a pair of co-dependent technologies.*
3) (D) **What is the proposed drug?** • What is the requested Pharmaceutical Benefits Schedule (PBS) restriction?	*Provide a description of the drug, its background, mechanism of action, etc. as per 2008 Pharmaceutical Benefits Advisory Guidelines (PBAC) Guidelines† subsection A.1. Specify the drug's Therapeutic Goods Administration (TGA) registration status.*
4) (O) **What is the biomarker?**	*This initial scenario considers genetic DNA biomarkers only, i.e. the assessment of one genetic locus at a time. Note that the Food and Drug Administration (FDA) in the United States of America has provided specific definitions of genomic biomarkers (i.e. assessment across the genome, testing hundreds or thousands of loci simultaneously). Genomic testing is beyond this initial scope and would be applicable to more complex scenarios.*
5) (T) **What is the proposed test?** • What is the requested Medicare Benefits Schedule (MBS) item descriptor?	*This relates to a description of a single test or assay. However, often tests are done in series when assessing genetic biomarkers or there may be an algorithm-based computation of the results of a number of tests. Describe the test method in sufficient detail that a laboratory technician would be able to perform it.* *Specify the range of techniques available to measure the biomarker (e.g. polymerase chain reaction (PCR), high resolution melting (HRM)), and indicate which method, if any, is regarded as the reference or 'gold' standard.*
6) (T & D) **Is the test (or drug) currently reimbursed through the MBS (or PBS)?**	*Describe current reimbursement arrangements for the test and the drug. This determines the extent of information needed for the current technology.*
7) (T & D) **What is the medical condition or problem being managed, i.e. the patient indication?**	*Describe the patient indication being addressed. If different test result thresholds are likely, or if eligibility for the drug is determined subjectively, consider providing alternative indications.*

Rationale for the submission

8) (O) **Is there a clear definition of the biomarker(s) (e.g. specific genetic DNA mutation(s))?**	*Describe the nature of the genetic DNA biomarker (e.g. single nucleotide polymorphisms (SNPs), mutation, or copy number variation (CNV)). Where relevant, include the following elements describing the context for a biomarker: (i) the general clinical area, (ii) the specific use of the biomarker, and (iii) the critical parameters which define when and how the biomarker should be used. Describe exactly what the test is identifying in cases where there is no "specific mutation", e.g. an expression micro-array of tumour tissue which identifies cancer with activation of a particular pathway, and susceptibility to a certain drug, but does not identify a specific mutation as such. Categorise the mutation as either a germline or somatic mutation. If the mutation is classified as a germline mutation, then consider issues related to heritability, e.g. testing of relatives and genetic counselling would need to be considered and assess the ethical and medico-legal implications of testing.*
9) (O) **What is the biological rationale for targeting that biomarker(s) with the drug?**	*Present the initial evidence that was relied on to select the biomarker. Describe and explain the overall approach to the selection of the biomarker including methods and relevant aspects of study design and statistical analysis. Describe the rationale for the selection of the population sample studied in the biomarker qualification. Present the criteria used for selection of candidate genes (e.g. candidate by position, by function, based on expression profiling data). Justify, using molecular biological or pharmacological principles, the plausibility of treatment effect modification (or interaction) between the biomarker itself and the drug, or alternatively between the drug and another factor for which the biomarker is a proxy. Advise whether this rationale precedes the specification of the data collection which forms the primary source of evidence.*
10) (O) **Do any other biomarker(s) predict variation in the comparative treatment effect (between using the drug and not using the drug)?** In the case of another biomarker that is a genetic mutation: • Have details on the specific mutation and the nature of the mutation been provided? • Is the effect of treatment on this other mutation consistent with the effect under consideration?	*(Note that this may be relevant even if the other biomarker(s) are claimed, but are not proven and/or are not reimbursed.)* *If testing for other biomarkers is reimbursed, this would move to a more complex scenario.*

11) (O) **What is the prevalence of a true positive biomarker in the population likely to receive the test?**	*The source population would be those who are eligible according to the requested MBS item descriptor and PBS restriction and follow the corresponding clinical pathway to the point of being offered the test – or the drug in the absence of the test. An estimate of the prevalence of a true positive biomarker is relevant to calculating the performance of a test in terms of its negative and positive predictive value. Indicate where there is no 'gold' standard to determine this true positive status of the biomarker and use an alternative appropriate methodology to estimate it.*

Proposed impact on current clinical practice

12) (T & D) **What are the relevant clinical pathways? That is, is there a description and comparison of the proposed clinical management of a typical patient up to the point of being offered the proposed test and subsequent therapy with the proposed drug, as compared to the currently existing clinical pathway(s) where the proposed test is not offered and the proposed drug is not available?**	*In these clinical pathways, outline all alternative tests/test strategies (whether in series or occurring concurrently) and all alternative treatments (including non-drug treatments) for the patient indication both with and without knowledge of the patient's biomarker status. If it is important for patients with a rapidly progressive disease to ensure that a timely test result is available to determine drug eligibility, indicate whether the test is therefore likely to be performed earlier in disease progression in a broader population than might otherwise be considered as potentially eligible for the drug. Identify tests and treatments that are commonly used and likely to be supplemented or replaced by the pair of co-dependent technologies (see Information Requests 13 and 14).*
13) (T) **Can the proposed test be used with other treatments and/or for other purposes?** (Refer to the clinical pathways provided in response to Information Request 12.)	*If other treatments or purposes are relevant, this would move to a more complex scenario.*
14) (T) **Is the test an additional test to other(s) currently defining the condition? Or a replacement test? Or both (i.e. depending on the test result, replace some tests or be additional to other tests)?** (Refer to the clinical pathways provided in response to Information Request 12.)	*Most commonly, the test would be an additional test; although occasionally if the biomarker is a strong predictor, then it could replace another test in the workup.*
15) (T) **How is it suggested that the test will be offered in Australia?**	*Specify the TGA registration status of the test. Assess access and quality assurance issues. Identify how many laboratories offering the test have NATA accreditation for that test. (Note that a way of determining this is not yet available.) Indicate whether the test accessibility is likely to be widespread or only available in a few selected laboratories across the country. Explain how the test would be undertaken in practice and what*

	impact it would have on patient and health professionals. Discuss the practicalities of a non-MBS accessible test.
16) (T) **Have the following been identified:** **i) the biospecimen required to perform the test?** **ii) whether this specimen needs to be collected specifically for the purposes of performing the test or has already been collected for another purpose?**	i) *For example: blood, tumour material (formalin-fixed paraffin embedded (FFPE) or fresh), bone marrow, cytology specimen, mouth swab.* ii) *For example: tumour already removed can be tested if archival FFPE is available and the test can identify the biomarker from this tissue.* *If a new specimen needs to be collected, specify the costs, risks and feasibility of collecting the sample. In some instances, such as a blood sample, the costs and risks would be trivial. In other instances, such as when a new biopsy is required, there may be significant costs as well as safety risks for the patient.*
17) *(If relevant)* (T) **What is the potential need for subsequent testing to identify new somatic mutations which may guide dosage or cessation of therapy with the co-dependent drug?**	*This will impact on the clinical need for the proposed test as well as its potential use to guide drug dosage titration and treatment continuation. If subsequent testing is needed, this would move to a more complex scenario.*
18) (T) **Are the test results expected to be consistent over time, including over the course of the disease?**	*Where test results may change over time, provide sufficient detail to clarify the relationship and timeframes between test results and the appropriateness of treatment. For example; Kirsten rat sarcoma viral oncogene homolog (K-RAS) testing of the primary colorectal cancer tumour is usually representative of the findings in metastases. However epidermal growth factor receptor (EGFR) results change with exposure to radiotherapy etc and so the results of testing the primary tumour may not be representative of what is happening in non-small cell lung cancer metastases.*
19) (O) **Can the proposed drug be used with other specific tests for that biomarker, other than the test proposed? What methodologies are available to test for the marker?**	*If other tests are publicly funded, this would move to a more complex scenario.*

SECTION B

Clinical benefit of the pair of co-dependent technologies in terms of patient health outcomes

20) (O) **Is there direct evidence of prognostic impact associated with different biomarker status?**	*This is used to discriminate prognostic impact as an alternative (or in addition) to treatment effect modification. It requires a comparison of outcomes in patients receiving usual care conditioned on the presence or absence of biomarker positive status.*

When presenting the body of evidence to address clinical benefit, two different options (Option 1 and Option 2) are provided so that available information can be used to maximum effect to inform a reimbursement decision.

OPTION 1. Is there 'direct evidence'* of the proposed test's impact on patient health outcomes? For example, patients randomised to the proposed test or to no test and followed through to allocation of the proposed drug or usual care and the subsequent impact of that treatment on their health outcomes. • **Level 1**: Is a trial available that randomised to use of the test or not, and then randomised to use of the drug or its main drug comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 1 at end of table – double-randomised controlled trial. • **Level 2**: If not, is a trial available that randomised to the use of the test or not, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 2 at end of table – single-randomised controlled trial of test. • **Level 3**: If not, is a trial available that prospectively tested eligible patients, and then randomised test positive or negative patients to use of the drug or its main comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes)? See Figure 3 at end of table –biomarker-stratified design / randomised trial of drug only (with the eligibility of all subjects determined by test	*Direct evidence is used to determine whether the pair of co-dependent technologies are (cost-) effective and safe. If randomised to use of the test, then biomarker status would be known and, on that basis, subsequent targeted therapy or usual care could be decided for the patient. If randomised to not using the test, then the patient would receive treatment that is not targeted by the biomarker result. 'Direct evidence' does not exclude the need for an assessment of translational issues (see Information Requests 40-42). Translation steps (applicability, transformation and extrapolation):* • *address external validity concerns of trials usually conducted in a different setting or with a different population (i.e. spectrum of disease)* • *address concerns that usually relate to the length of follow-up of the direct evidence, to the use of surrogate outcomes and most importantly to capture the point estimate and confidence limits of the treatment effect taking into account the impacts of incorporating the test results.* *Given that Level 2 direct evidence does not provide information on the test(biomarker)-drug relationship ie evidence that the biomarker is a treatment effect modifier or prognostic factor, therefore consider supplementing with Level 3 or 4 direct evidence (also see Information Requests 34 and 35).* *Given that Level 3 and 4 direct evidence effectively involve uncontrolled study designs (i.e. there is no trial arm provided to assess the impact of not testing biomarker status), consider providing a supplementary 'linked evidence' approach (see Option 2 below) so that at least a comparison of the proposed test/test*

result). • **Level 4**: If not, is a trial available that randomised eligible patients to use of the drug or its main comparator, and then followed participants to measure clinical outcomes (whether surrogate outcomes or directly patient relevant outcomes), and then analysed results across subgroups of patients defined by whether they are positive for the test (or biomarker) or whether they are negative to the test (or biomarker)? See Figure 4 at end of table – biomarker-stratified design / randomised trial of drug only (with the test result determined through subgroup analysis). • **Level 5**: If not, then move to corresponding guidance on 'linked analyses' (see **Option 2**, below).	*strategy and existing test/test strategy can be made with respect to their relative diagnostic accuracy.* *Level 4 direct evidence may use archival tissue/sampling to determine biomarker status. Exercise caution when interpreting results from Level 4 studies where biomarker status might change over time, including where there is evidence that intervening treatment may modify the biomarker.*
21) (O) **Is the direct evidence presented and selected in a comprehensive and unbiased manner?**	For example, present a systematic review of direct evidence concerning this pair of proposed test and proposed drug for this biomarker with pre-specified inclusion/exclusion criteria and a PRISMA^ flowchart indicating how trials were selected and the reasons why any potentially relevant trials were excluded.
22) (O) **Is the direct evidence of good quality?**	*Assess bias, confounding, the impact of chance on results and whether the analyses were pre-specified and/or exploratory. Use an intervention study design critical appraisal checklist to cover all issues likely to affect the internal validity of the presented trial results.*
23) (O) **Does the direct evidence provided show a clinically important and statistically significant impact on patient-relevant health outcomes?**	*Assess both effectiveness and safety. Describe outcomes in the studies (primary and secondary outcomes) and statistical methods used. Provide an extended assessment of comparative harms. Assess the balance of benefits and harms and interpret findings from the body of evidence.*
24) (O) **Is the direct evidence provided applicable to the requested MBS and PBS populations?**	*Describe patient characteristics in the trials and indicate whether they are relevant to the Australian situation. Indicate whether the requested technologies were provided in a setting similar to the Australian setting of use. Also see Section C.*
OPTION 2. Is there 'linked evidence'[#] available of the test's impact on patient health outcomes? In other words, can different types of evidence from different sources be linked in a chain of argument to	*For example, this might involve linking evidence of test accuracy with evidence that the test result changes patient management, and with evidence that the alternative treatments have different effectiveness and safety profiles.* *Further background is provided in the 2005 Medical*

estimate this impact?	*Services Advisory Committee (MSAC) Guidelines‡ for the assessment of diagnostic technologies. Note that a full linked evidence approach is only meaningful when the evidence for the proposed test and the evidence for the proposed drug have been generated in similar patient populations and so it is clinically sensible to link the two data sets. If the test identifies patients earlier or with a different spectrum of disease than the patients in whom the drug has been trialled, then it is not clinically sensible to link this evidence. In such circumstances direct evidence is needed.*

What is the test effectiveness and safety?

25) (T) **What is the analytical test performance?**	*Analytical test performance assesses how accurately and how consistently the test identifies biomarker status, e.g. the coefficient of variation and other appropriate statistics. Present any differences across laboratories in how they characterise test results (e.g. a kappa statistic or other concordance statistic). Identify whether there is an external quality assurance program by which laboratories can benchmark their assays.*
26) (T) **Is there a clinical reference standard or a 'gold' standard against which test performance can be measured?**	*Indicate whether this clinical reference standard is also the relevant diagnostic comparator, i.e. the current test/test strategy being used in the absence of the proposed test.*
Option A (if no reference standard): test performance is determined using predictive accuracy. **OR**	*If a reference standard is not available or unacceptable for the requested use and/or requested population: consider whether one can be constructed. If so, calculate estimated sensitivity and specificity under the constructed standard. In this situation: specify the designated reference standard that was constructed; create the new reference standard independently from the analysis of results of the proposed test (ideally, in advance of collecting any specimens); and consult with statisticians and health professionals prior to constructing the reference standard. (FDA, 2007)††* *If a reference standard is not available and cannot be constructed: calculate and report measures of agreement (the terms sensitivity and specificity are not appropriate to describe these comparative results). Instead, the same numerical calculations are made, but the estimates are called positive percent agreement and negative percent agreement, rather than sensitivity and specificity. (FDA, 2007)†† This reflects that the estimates are not of accuracy but of agreement of the proposed test with the non-reference standard. In addition, quantities such as positive predictive value, negative predictive value, and positive and negative likelihood ratios cannot be computed since the subjects' condition status (as*

	determined by a reference standard) is unknown. In this situation: report the 2x2 table of results comparing the candidate test with the comparative method; describe the comparative method and how it was performed; and report the agreement measures along with their confidence intervals (FDA, 2007)†† or kappa statistics. Alternatively odds ratios could be reported indicating the likelihood of an outcome, given that particular test result.
Option B (if a reference standard is available): test performance is determined using diagnostic accuracy measures.	*Test performance measures include: sensitivity, specificity, likelihood ratios, positive and negative predictive values, area under curve (AUC). Designate a reference standard and compare the proposed test to the designated reference standard by cross classifying the test results of patients who are representative of the intended use population. Include confidence intervals and significance levels to quantify the statistical uncertainty in these estimates due to the subject/sample selection process. This type of uncertainty decreases as the number of participants in the study increases.* *Assess whether there is a test performance level below which the test should not be used (for example, either false positives are too great or false negatives are too great) so that other better performing tests are needed.* *If a reference standard is available, but impractical: use it to the extent possible. Calculate estimates of sensitivity and specificity adjusted to correct for any (verification) bias that may have been introduced by not using the reference standard to its fullest extent. (FDA, 2007)†† If it is determined that using a reference standard on all subjects is impractical or not feasible, obtain estimates of sensitivity and specificity using the proposed test and a comparative method (other than a reference standard) on all subjects, and use the reference standard on just a subset of subjects (sometimes called partial verification studies or two-stage studies). In this instance, the usual formulas for calculating sensitivity and specificity would give biased estimates of sensitivity and specificity, i.e. verification or workup bias. However, if the designated reference standard is applied to a random subset of all subjects, or to all subjects where the proposed test and the comparative method disagree and to a random sample of subjects where they agree, then it is possible to compute adjusted estimates (and variances) of sensitivity and specificity. In this case,*

	retest a sufficient number of subjects to estimate sensitivity and specificity with reasonable precision. (FDA, 2007)††
27) (T) **Is the evidence of diagnostic or predictive accuracy presented and selected in a comprehensive and unbiased manner?**	*For example, present a systematic review of diagnostic accuracy studies for this test with inclusion/exclusion criteria delineated and a PRISMA^ flowchart indicating how trials were selected and reasons why any potentially relevant trials were excluded.*
28) (T) **Is the evidence of diagnostic or predictive accuracy of good quality?**	*See QUADAS~ checklist items: 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14.*
29) (T) **Are there any safety considerations that will impact on the entire process of testing?**	
30) (T) **Is the evidence of test accuracy and safety applicable to the requested MBS and PBS populations?**	*Assess whether test accuracy was determined in the correct population. See QUADAS~ checklist items: 1, 12.*
31) *(If relevant for a comparison of tests)* (T) **Which test has the best test performance (in terms of accuracy and/or clinical benefit)?**	*Assess trade-offs in false positives, false negatives, and in positive predictive value and negative predictive value.* *If other tests are publicly funded, this would move to a more complex scenario.*
32) *(If relevant for a comparison of tests)* (T) **Which test is most accessible/ available/ used?**	*Assess access and quality assurance issues.* *If other tests are publicly funded, this would move to a more complex scenario.*
33) (O) **Will knowledge of the test result cause a change in the management of the patient by the treating clinician? Are there instances where management would not change, despite the test indicating the biomarker is present?**	*There may be 'leakage' issues identified through an assessment of the 'change in management' part of a linkage. Often a test is done to rule out a drug (e.g. to avoid potential drug-related adverse events or the development of drug resistance), but the drug is given anyway, or, alternatively, the test is used to select a specific drug, but the drug is not provided. As companion tests in a co-dependent pairing will often be used to guide drug therapy decisions, this would need to be explicitly addressed. Once listed, these issues could be informed by data that compare the numbers of test 'positive' results and prescriptions*

	filled for the drug.

What is the test-drug effectiveness and safety?

34) (O & D) **Is there evidence available of treatment effect modification or significant interaction between biomarker status and treatment outcomes?** For example, is there evidence of substantial variation in a measure of relative treatment effect between the proposed drug and usual care trial arms after stratifying for biomarker status?	*Treatment effect modification in this setting identifies a relationship between the biomarker and the drug, which is likely to be unique or limited to companion tests assessing a particular biomarker and drugs with a particular mechanism of action (cross refer to the response to Information Request 9). This means that both technologies are required to produce a clinical benefit and the reimbursement decision may need to encompass both technologies.*
35) (O & D) **Is there evidence available of better targeting to patients likely to respond most by using the prognostic impact of the biomarker to determine the baseline risk of disease progression?** For example, is there evidence of minimal variation in a measure of relative treatment effect between the proposed drug and usual care trial arms, but biomarker status helps identify patients at greatest risk of an event which helps maximise the absolute treatment effect?	*If a drug's result is due to better targeting to those patients that are likely to respond most, this identifies a relationship between the biomarker and a potentially broader range of existing and future treatment options (potentially including non-drug treatment options) than is likely to apply for treatment effect modification.* *It is possible for both treatment effect modification and prognostic impact to co-exist. In this case, in order to assess the unique contribution of the drug therapy, an assessment of its effect must be made relative to usual care and adjusted for the background prognostic impact that is operating in both the drug and usual care arms and which is also flagged by that particular biomarker.* *By contrast, if the drug's apparent improvement in result is simply due to the fact that a certain patient subgroup (flagged by a specific biomarker) will always do better, then the level of co-dependency between the technologies is low. This may allow reimbursement of either test or drug or both technologies.*
36) (O & D) **Is the drug effectiveness evidence, as conditioned by the test or biomarker result, obtained in a comprehensive and an unbiased manner?**	*For example, present a systematic review of randomised trials of the proposed drug targeting this biomarker with inclusion/exclusion criteria delineated and a PRISMA flowchart indicating how trials were selected and reasons why any potentially relevant trials were excluded.*
37) (O & D) **Is this drug effectiveness evidence, as conditioned by the test or biomarker result, of good quality?**	*Assess bias, confounding, the impact of chance on results and whether the analyses were pre-specified and/or exploratory. Use an intervention study design critical appraisal checklist to cover all issues likely to affect the internal validity of the presented trial results. Confounding may occur as a consequence of imbalance in biomarker status in the drug and usual care trial arms in the case where biomarker status is also a prognostic factor.*

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

SECTION C

Can the test-drug evidence of effectiveness be translated to an economic model for the Australian clinical setting?

40) (T & D) **Was translation of trial data to the Australian setting conducted appropriately?**	*Corresponds to subsection C.1 of the 2008 PBAC Guidelines†. Identification of: applicability issues (population and circumstances of use), extrapolation issues, transformation issues, other translation issues.*
41) (T & D) **What are the proposed translation analyses?**	*Corresponds to subsections C.2 and C.3 of the 2008 PBAC Guidelines†. Analytical plan addressing applicability - analysis of heterogeneity, subgroup analysis, meta-regression, treatment effect variation.*
42) *(If relevant)* (D) **How are surrogate outcomes transformed to final patient-relevant outcomes?**	*1. Conduct systematic literature review. 2. Identify any randomised trial evidence in related drugs. Quantify effects. 3. Link current drug to evidence in 2. (e.g. by identifying mechanism of action).*

Section D

Is the proposed use of the pair of co-dependent technologies cost-effective?

Is the structure of the model appropriate for the clinical indication being modelled?

43) (O) **Is there an economic evaluation in the broad clinical management setting, initiating before the decision to test or treat?**	*This corresponds to subsections D.1 (type), D.2 (population and circumstances of use), D.3 (structure and rationale - time horizon, outcomes, methods of calculation), and D.4 (variables - costs, outcomes, probabilities, discounting) of the 2008 PBAC Guidelines†.*
44) (O) **Is the economic decision tree consistent with the clinical pathways provided in response to Information Request 12?**	
45) (O & T) **Is there a supplementary analysis of non-health related impacts of diagnostic testing?**	*See p133, subsection D.4 of the 2008 PBAC Guidelines†.*

Were transition probabilities in the model consistent with test and drug performance as determined from the evidence presented for clinical benefit?

46) *(If relevant)* (O & T) **Was the positive predictive value (PPV) of the test calculated and included in the model?** (if **Option 2** 'linked evidence approach' was used in Section B)	*PPV is calculated based on the sensitivity of a test (in the correct population, i.e. no spectrum bias) and prevalence (probability) of the biomarker (e.g. phenotypic expression of mutation) in the target population. It is the probability that a test positive is correct. The PPV is used in a Bayesian manner to condition the model.*
47) *(If relevant)* (O & T) **Was 1-PPV calculated and used in the model?** (if **Option 2** 'linked evidence approach' was used in Section B)	*1-PPV is the probability that a test positive is incorrect (false positive) and predicts the consequence that patients are treated unnecessarily with consequent decrement in effectiveness and increment in harms. It is used in a Bayesian approach to condition the model.*
48) *(If relevant)* (O & T) **Was the negative predictive value (NPV) of the test calculated and included in the model?** (if **Option 2** 'linked evidence approach' was used in Section B)	*NPV is calculated based on the specificity of a test (in the correct population, i.e. no spectrum bias) and 1-prevalence (probability) of the biomarker in the target population and is the probability that a test negative is correct. It is used in a Bayesian manner to condition the model.*
49) *(If relevant)* (O & T) **Was 1-NPV calculated and used in the model?** (if **Option 2** 'linked evidence approach' was	*1-NPV is the probability that a test negative is incorrect (false negative) and predicts the scenario where patients receive usual care instead of the proposed drug with consequent decrement in effectiveness. It is used in a Bayesian manner to condition the model.*

used in Section B)	
50) (O & D) **Were the treatment effects on intended outcomes included appropriately?**	*Where prognostic impact is operating in addition to treatment effect modification, ensure that the model appropriately adjusts for this factor when presenting absolute treatment effects.*
51) (O & D) **Was the incidence of drug-related adverse events included in the model?** **(i) for true positives and false positives?** (if **Option 2** 'linked evidence approach' was used in Section B) **Or** **(ii) from the trial evidence?** (if **Option 1** 'direct evidence' was used in Section B)	*Determine whether biomarker test status predicts or does not predict any comparative treatment effect variation in terms of adverse events. Include the impact of drug-related adverse events on patients with a positive test result.*
52) (O & T) **Was the incidence of test-related adverse events for all those tested included?**	
Were correct resource items and correct costs used, reflecting delivery of the test and drug to patients in Australia?	
53) (O & D) **Were unit drug costs included in the model?**	
54) (O & T) **Were unit test costs included in the model?**	*In estimating the cost of testing (and associated costs), include the cost of tests undertaken on all patients for whom the drug is being considered, not just the cost of the test for those who are found to be suitable. Include all relevant sources of costs (e.g. infrastructure, training, quality assurance) which need to be captured in a MBS fee- for- service for a pathology test.*
55) (O & T) **Were costs of sampling included in the model?**	*For example, taking, storing, retrieving and transporting biopsy samples.*
56) (O & T) **Were costs of test administration included in the model?**	
57) (O & T) **Were unit costs for consultations regarding test results included in the model?**	

58) (O & T) **Were costs of re-testing and non assessable results included in the model?**	
59) (O & T) **Were costs for adverse events associated with testing included in the model?**	
60) (O & T) **Were the costs of additional and further testing as a result of the proposed test included in the model?**	
61) (O & D) **Were costs of drug delivery and administration included in the model?**	
62) (O & D) **Was the cost of drug-related adverse events included in all arms of the model, including those where the test result was false positive?**	
63) (O) **Was the cost of other concomitant drugs included in the model?**	
64) (O) **Were the costs of other relevant healthcare resources (e.g. diagnostic, medical, hospital, allied health) included in the model?**	
What were the results of the economic model?	
65) (O) **Are the results presented in a stepped form?** • Is an Incremental Cost Effectiveness Ratio (ICER) presented?	*This corresponds to subsection D.5 of the 2008 PBAC Guidelines†. Present the economic evaluation in a stepped form (model steps / translation steps), with incremental results and 95% confidence intervals.*
Was uncertainty in the modelled inputs captured appropriately?	
66) (O & D) **Was the uncertainty around the drug's effectiveness assessed?**	*In instances where both treatment effect modification and prognostic impact are operating in the drug/biomarker relationship, assess the uncertainty of the estimated absolute treatment effect and model this uncertainty.*
67) (O & T) **Was uncertainty around test**	

accuracy assessed?	
68) (O) **Was uncertainty around prevalence of the biomarker assessed?**	
69) (O) **Were other variables of uncertainty assessed (e.g. population age, gender)?**	
70) (O) **Was uncertainty around cost inputs assessed?**	
71) (O) **Was a scenario analysis provided concerning the option of PBS listing the drug without the biomarker test pre-requisite?**	

Section E

What is the financial impact of the proposed listing of the pair of co-dependent technologies?

72) (O) **Is a financial impact analysis presented incorporating both MBS and PBS components, with results split by sector (public, private, patient, other)?**	*Corresponds to subsections E.1 (justification of data sources), E.2 (which requires PBS and MBS co-payment calculations, including in and out of hospital rules), E.3, and E.4 of the 2008 PBAC Guidelines†.*
73) (O) **Is an epidemiologic estimate for disease burden presented that is based on the prevalence of the biomarker?** • What is the prevalence of the medical condition? • What is the prevalence of the test result which would mean that the patient is eligible for the drug?	*A market share estimate for this scenario would not be used in the instance of a new biomarker because previous drug utilisation will not have been targeted to this biomarker. Expert epidemiological advice on whether prevalence is expected to remain constant after listing is likely to be needed (see p149 of the 2008 PBAC Guidelines†). First estimate the number(s) of patients likely to be considered for the test (e.g. with the medical condition as defined) and then the proportion of each number likely to receive a positive test result (for use of the drug).*
74) (T) **What is the likely use and overall financial cost of the test?**	*Include the cost of testing all patients who would be considered for the test and the cost of re-testing when unevaluable results are produced or after therapy is initiated (i.e. to monitor therapy or to determine when therapy should cease).*
75) (D) **What is the likely use and overall financial cost of the drug?**	*Include the number of packs (in addition to cost), and disaggregate by PBS beneficiary type.*

76) (D) **Will there be a change in the use of other drugs as a consequence of listing the proposed drug?**	*Consider both concomitant and substituted drugs.*
77) (O) **What other MBS costs would be incurred if the test and drug were listed?**	*Include the change in number of services processed (processing costs), MBS fees incurred (scheduled fee - benefits) particularly considering procedures for administration, consultations for adverse events, consultations and tests related to diagnosis of biomarker status, confirming eligibility, etc.*
78) (O) **What is the estimated financial impact on other health budgets?**	*Corresponds to subsection E.6 of the 2008 PBAC Guidelines†. Include the change in the number of inpatient admissions, accident and emergency attendances, outpatient clinic visits, etc. Also give specific consideration to test, e.g. if it would require any in-hospital advice.*
79) (O) **Is the extent of financial uncertainty estimated?**	*Corresponds to subsection E.6 of the 2008 PBAC Guidelines†.*

RELEVANT GUIDELINES

† Australian Government Department of Health and Ageing. *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee*. Canberra, ACT: Commonwealth of Australia, 2008.

‡ Medical Services Advisory Committee (MSAC). *Guidelines for the assessment of diagnostic technologies*. Canberra, ACT: Commonwealth of Australia, 2005:1-93.

†† Food and Drug Administration Center for Devices and Radiological Health. *Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests*. Diagnostic Devices Branch, Division of Biostatistics, Office of Surveillance and Biometrics, U.S. Department of Health and Human Services, 2007.

TABLE NOTES

MBS = Medicare Benefits Schedule

PBS = Pharmaceutical Benefits Schedule

* ’Direct evidence’ is a trial that compares groups of people receiving either the currently used diagnostic test/test strategy or the proposed diagnostic test/test strategy and measures the differential impact of the diagnostic method on patient health outcomes (MSAC *Guidelines for the assessment of diagnostic technologies*, 2005).

The 'linked evidence approach' was proposed by MSAC whereby evidence of test accuracy comparing the proposed and current test/test strategy could be linked (if considered to be appropriately transferable) to separately sourced evidence of treatment effectiveness in order to approximate the likely clinical effectiveness of the proposed test/test strategy.

^ PRISMA: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *BMJ* 2009;339:b2535, doi: 10.1136/bmj.b2535.

~ QUADAS: Whiting PRA, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Med Res Methodol* 2003, 3(1):25.

NB: This framework is only an important first step and so does not comprehensively engage the full range of possible co-dependent technology submissions for reimbursement. Later steps will address these as more complex scenarios.

FIGURE 1 DOUBLE-RANDOMISED CONTROLLED TRIAL

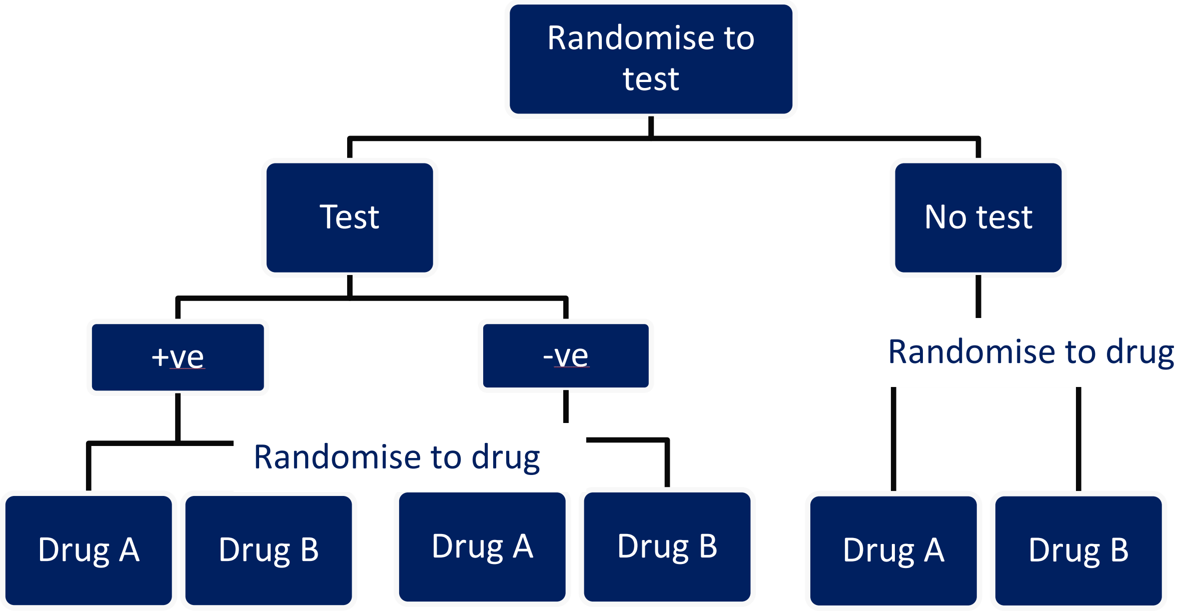
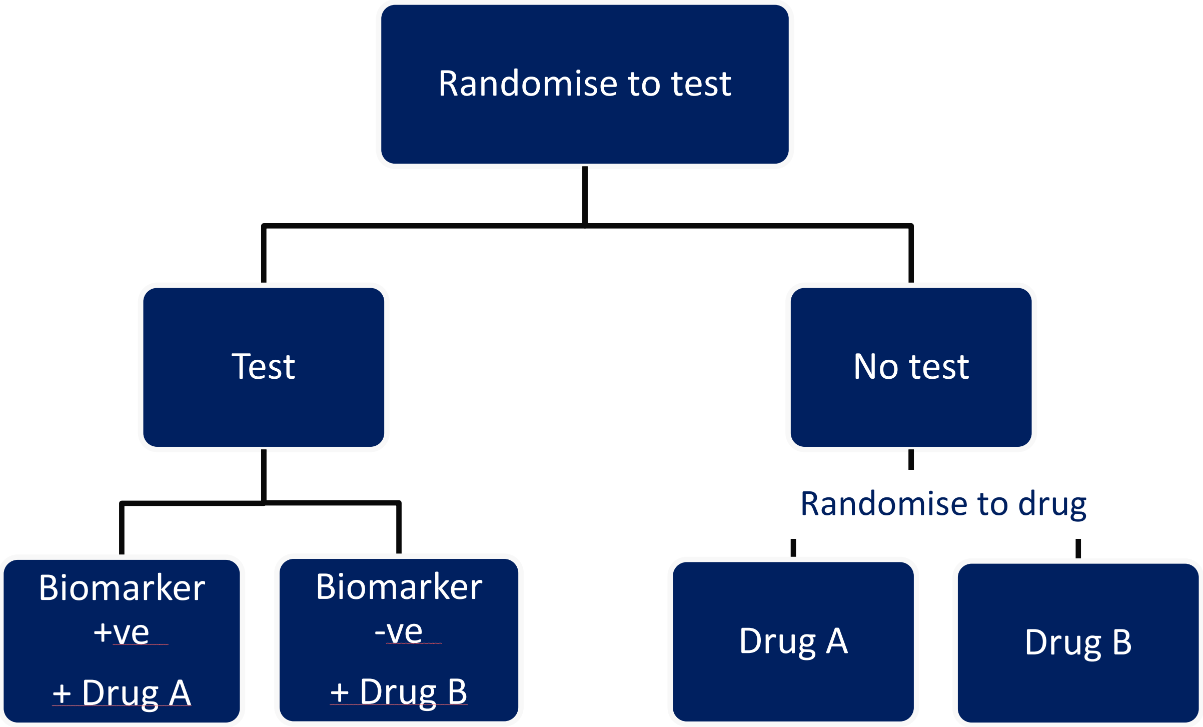
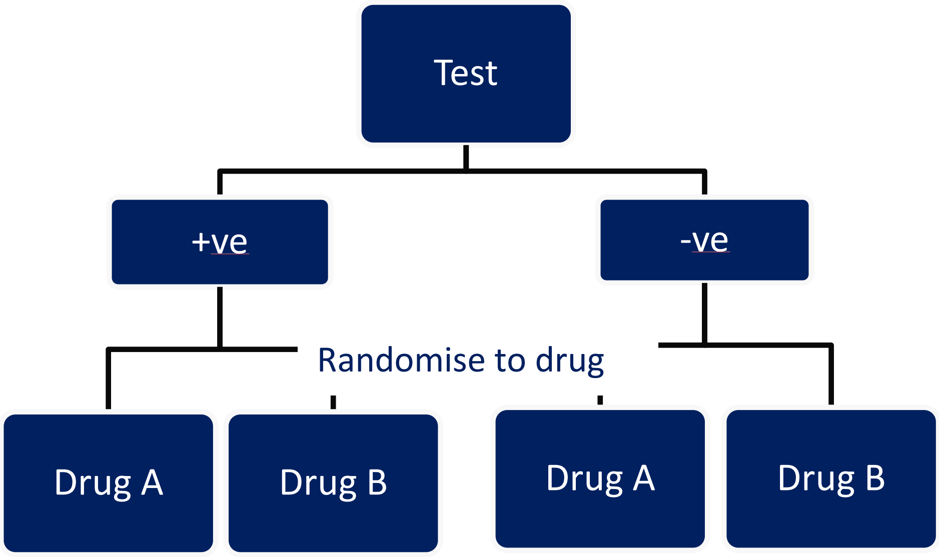


FIGURE 2 SINGLE-RANDOMISED CONTROLLED TRIAL (TARGETED TREATMENT)



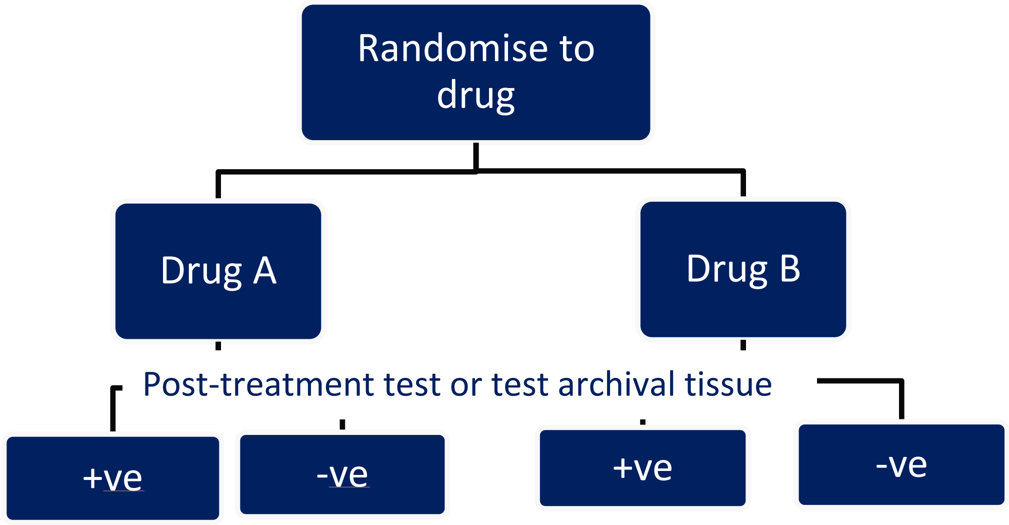
Note: This design cannot explain test (biomarker)-drug relationship. Additional evidence would need to be provided to show whether biomarker is a treatment-effect modifier or a prognostic factor.

FIGURE 3 BIOMARKER-STRATIFIED DESIGN



Note: At drug randomisation (assuming a reasonable sample size) all variables other than biomarker status should be fairly evenly distributed in Drug A and Drug B groups. This explains the likely test (biomarker)-drug relationship but not the incremental benefit of the test – that is there may be uncertainty as to whether the biomarker +ve/-ve is responsible for the differential treatment effect or some other unmeasured variable.

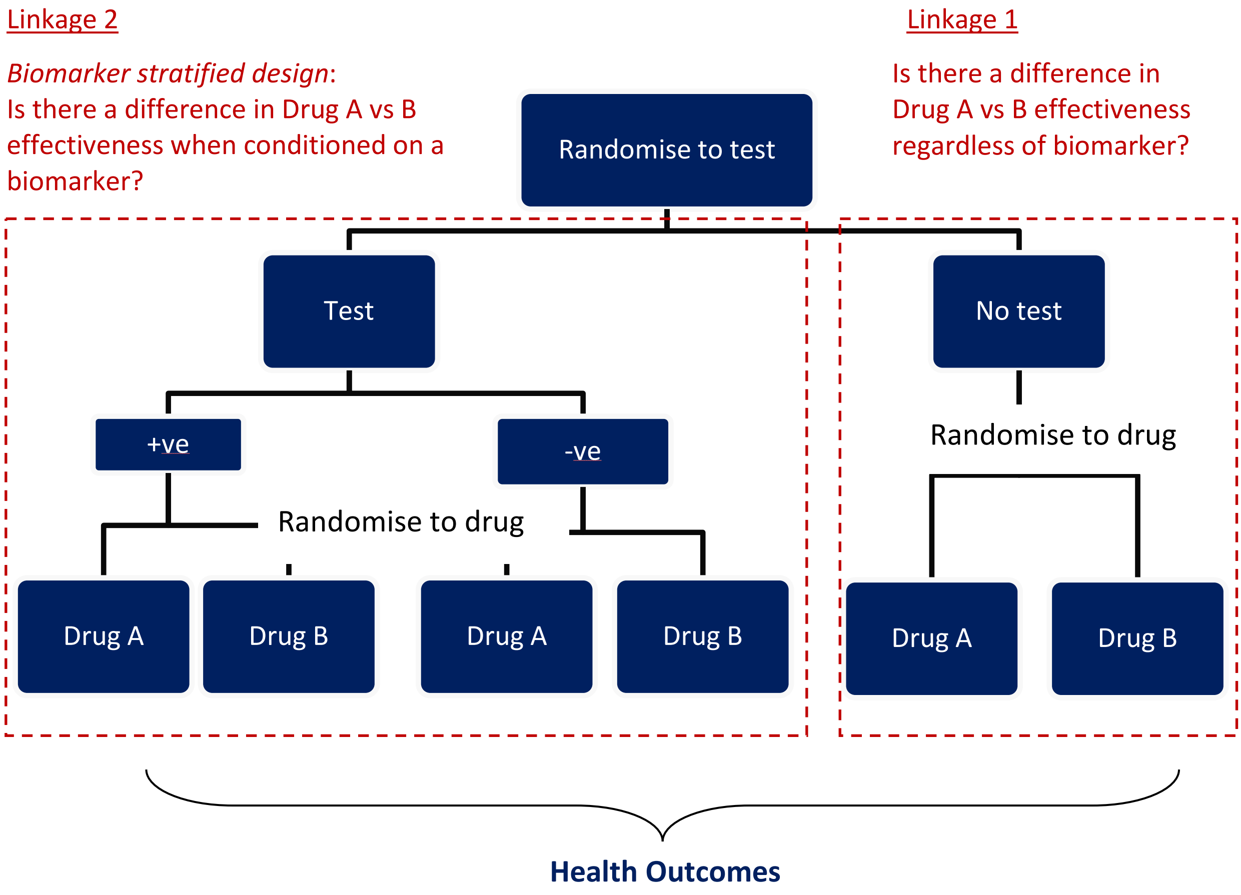
FIGURE 4 BIOMARKER-STRATIFIED DESIGN VIA SUBGROUP ANALYSIS



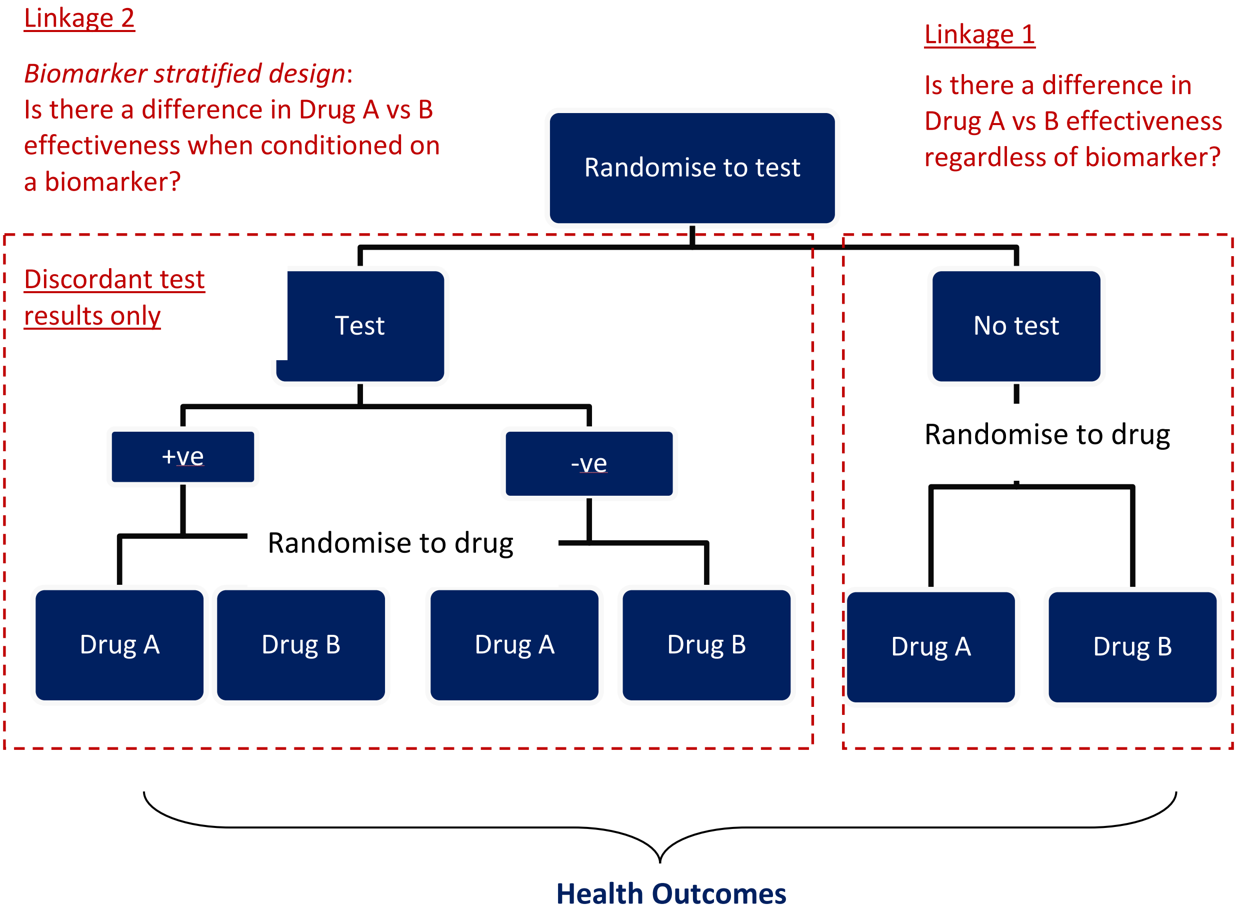
Example of "linked evidence approach"

Some options (not exhaustive) for providing and linking different types of study design to approximate the results of the 'ideal' double-randomised controlled trial (figure 1, additional file 1), assuming there is adequate transferability between populations and interventions in each linkage.

Evidence Linkage Option A


Linkage 2
Biomarker stratified design:
Is there a difference in Drug A vs B effectiveness when conditioned on a biomarker?
Linkage 1
Is there a difference in Drug A vs B effectiveness regardless of biomarker?
Randomise to test
Test
No test
+ve
-ve
Randomise to drug
Randomise to drug
Drug A
Drug B
Drug A
Drug B
Drug A
Drug B
Health Outcomes

Evidence Linkage Option B



Linkage 2

Biomarker stratified design:
Is there a difference in Drug A vs B effectiveness when conditioned on a biomarker?

Linkage 1

Is there a difference in Drug A vs B effectiveness regardless of biomarker?

Randomise to test

Discordant test results only

Test

No test

+ve

-ve

Randomise to drug

Drug A

Drug B

Drug A

Drug B

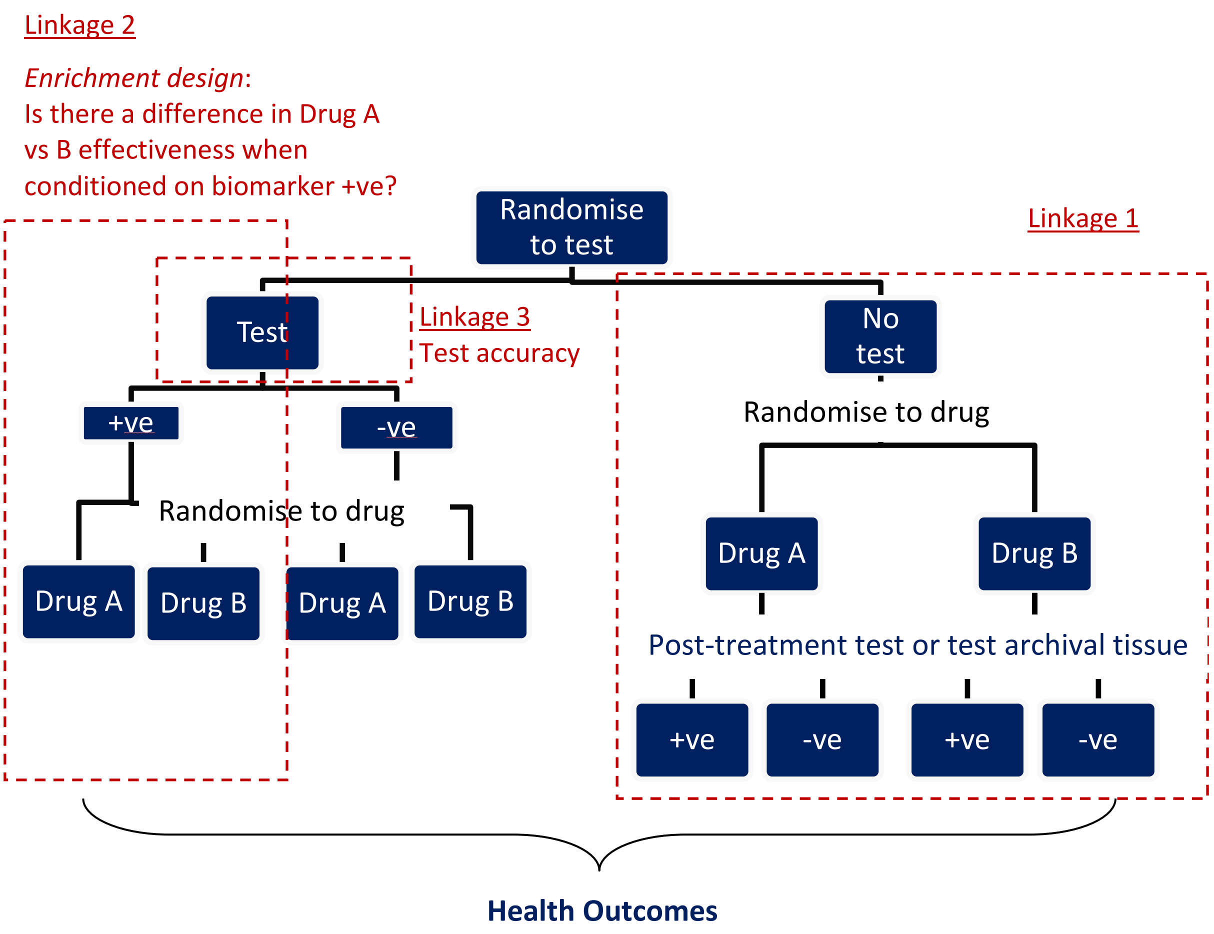
Randomise to drug

Drug A

Drug B

Health Outcomes

Evidence Linkage Option C



Linkage 2

Enrichment design:
Is there a difference in Drug A vs B effectiveness when conditioned on biomarker +ve?

Randomise to test

Test

No test

Linkage 3
Test accuracy

Linkage 1

+ve

-ve

Randomise to drug

Drug A

Drug B

Drug A

Drug B

Randomise to drug

Drug A

Drug B

Post-treatment test or test archival tissue

+ve

-ve

+ve

-ve

Health Outcomes

Example - distinguishing prognostic impact from treatment effect modification

Outcome	Biomarker +ve		Biomarker -ve	
	Drug A N=100	Drug B N=100	Drug A N=100	Drug B N=100
1. Treatment effect modification				
5 year OS	60%	30%	30%	30%
	RR=2.0 [1.4, 2.8]		RR=1.0 [0.7, 1.5]	
			2x	
2. Prognostic impact				
5 year OS	70%	60%	35%	30%
	RR=1.2 [1.0, 1.4]		RR=1.2 [0.8, 1.7]	
3. Treatment effect modification + Prognostic impact				
5 year OS	60%	40%	20%	20%
	RR=1.5 [1.1, 2.0]		RR= 1.0 [0.6, 1.7]	

OS = overall survival; RR = relative risk/rate ratio, [95% Confidence Interval]

REFERENCES

1. AHRQ. Methods Guide for Medical Test Reviews. Rockville, MD: Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services; 2010.

2. CDC. ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing [internet]: Centers for Disease Control and Prevention (CDC), Office of Public Health Genomics; 2010. Available from: http://www.cdc.gov/genomics/gtesting /ACCE/acce_proj.htm.

3. FDA. In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff [internet]. Rockville, Maryland: Food and Drug Administration (FDA), U.S. Department of Health and Human Services; 2011.

4. EUnetHTA. HTA Core Model for Diagnostic Technologies. Work Package 4: European Network for Health Technology Assessment; 2008.

5. NICE. Interim methods statement (pilot). Version 8 [internet]. London: National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme; 2010.

6. MSAC. Guidelines for the assessment of diagnostic technologies. Canberra, ACT: Commonwealth of Australia2005 August 2005.

7. Lord SJ, Irwig L, Bossuyt PM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making. 2009 Sep-Oct;29(5):E1-E12.

8. Staub L, Dyer S, Lord S, Simes RJ. Linking the Evidence: Intermediate Outcomes in Medical Test Assessments. International Journal of Technology Assessment in Health Care. 2012;28(1):52-8.

9. Harris R, Helfand M, Woolf S, Lohr K, Mulrow C, Teutsch S, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med. 2001;20(3 Suppl):21-35.

10. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Medical Decision Making. 1991;11:88-94.

11. di Ruffano L, Davenport C, Eising A, Hyde C, Deeks J. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. Journal of Clinical Epidemiology. 2012;65(3):282-7.

12. Merlin T, Farah C, Schubert C, Mitchell A, Hiller J, Ryan P. Assessing personalized medicines in Australia: A national framework for reviewing codependent technologies. Medical Decision Making 2012 August 22, 2012.

13. Marinovich L. Optical Coherence Tomography [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au.

14. Buckley L, Wang S, Merlin T. Molecular testing for myeloproliferative disease. Part A – Polycythaemia vera, essential thrombocythaemia and primary myelofibrosis. Part B - Systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au.

15. Schoeppe S, Lewis S, Marinovich L, Wortley S. Positron emission tomography for cervical cancer [internet]. Canberra: Commonwealth of Australia; 2010. Available from: www.msac.gov.au.

16. Lord S, Lei W, Griffiths A, Walleser S, Parker S, Thongyoo S, et al. Breast magnetic resonance imaging [internet]. Canberra: Commonwealth of Australia; 2007. Available from: www.msac.gov.au.

17. Merlin T, Moss J, Brooks A, Newton S, Hedayati H, Hiller J. B-type natriuretic peptide assays in the diagnosis of heart failure [internet]. Canberra: Commonwealth of Australia; 2008. Available from: www.msac.gov.au.

18. Marinovich L, Wortley S. Positron emission tomography for glioma. Canberra: Commonwealth of Australia; 2010.

19. Gillespie J, Guarnieri C, Phillips H, Bhatti T. Urinary metabolic profiling for detection of metabolic disorders [internet]. Canberra: Commonwealth of Australia; 2009. Available from: www.msac.gov.au

20. Gillespie J, Smala A, Walters N, Birinyi-Strachan L. Hepatitis B virus DNA testing [internet]. Canberra: Commonwealth of Australia; 2007. Available from: www.msac.gov.au.

21. Craig D, McDaid C, Fonseca T, Stock C, Duffy S, Woolacott N. Are adverse effects incorporated in economic models? A survey of current practice. International Journal of Technology Assessment in Health Care. 2010;26(03):323-9.

22. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. British Medical Journal. 2006;332:1089-92.

23. Micheel C, Ball J, editors. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Washington DC: National Academy of Sciences; 2010.

24. Anderson L, Petticrew M, Rehfuess E, Armstrong R, Ueffing E, Baker P, et al. Using logic models to capture complexity in systematic reviews. Research Synthesis Methods. 2011;2:33-42.

Relevance of Paper 4 to the thesis

Paper 4 has addressed the final research question posed for the thesis, that is:

Can the linked evidence approach (LEA) be feasibly adapted to the evaluation of personalised medicines ie the use of a genetic test to target a pharmaceutical treatment?

Paper 4 describes how the LEA methodology was modified to address the direct evidence requirements of a pharmacogenetic technology, namely, a hypothetical double-randomised (randomised to test and randomised to drug treatment) controlled trial, as opposed to the original theoretical framework of a diagnostic randomised controlled trial. This meant that, in addition to considerations of test accuracy, impact on patient management and treatment effectiveness, evidence is needed on the validity of the biomarker (ie whether it is a prognostic factor, treatment effect modifier or both), and whether testing for the biomarker adds any clinical (and cost) benefit when compared to the usual method of patient selection (eg not testing).

No one approach for linking evidence is mandated in the theoretical framework; different types of studies can be linked to canvass the issues answered by the one hypothetical double randomised controlled trial. This flexibility of approach maximises the use of the varying, and often limited, literature available for each pharmacogenetic technology.

The approach does identify some types of evidence that are more prone to bias than other types of evidence. It is also recognised that linkage pathways that rely on multiple disparate studies are likely to provide less reliable estimates of diagnostic and treatment effectiveness than pathways with fewer linkages, simply because with each linkage there is an assumption that the findings are transferable to the next step in the pathway.

The theoretical framework is an adaption of the MSAC LEA approach. It also addresses the factors considered by Austin Bradford-Hill that are needed to determine causality (in order to determine the validity of the biomarker). It incorporates the approaches already used to evaluate the effectiveness of drug treatments in Australia, and addresses policy maker preferences for the evidence needed to inform reimbursement decision-making.

The framework is the first of its kind, ie an evaluation framework for pharmacogenetic technologies to inform reimbursement decisions. The framework is not just a hypothetical methodological approach but has been successfully translated into practice. The approach was initially piloted at the national level in Australia in 2011. Its use is ongoing as the method used to evaluate pharmacogenetic technologies and to inform MSAC and PBAC decision-making regarding the public funding of tests and drugs, respectively.

The practical experience with this adapted LEA approach is described in Chapter 7.

CHAPTER 7

PRACTICAL EXPERIENCE WITH THE LEA METHOD DEVELOPED TO ASSESS PERSONALISED MEDICINES

In the previous chapter a novel methodology involving the adaptation of LEA was discussed. This methodology underpinned a framework, introduced by the Federal Government in Australia in 2011, that was developed to assess the safety, effectiveness and cost-effectiveness of personalised medicines for reimbursement decisions.

As this was the first such HTA methodology designed for evaluating these types of personalised medicines, additional information has been provided in this thesis to communicate to the reader the international context concerning the evaluation of personal medicines and the evolution of the novel methodology that was developed. It was also important to determine whether the method has been useful in helping policy makers make public funding decisions.

Relevant research question

4. *Can the linked evidence approach (LEA) be feasibly adapted to the evaluation of personalised medicines ie the use of a genetic test to target a pharmaceutical treatment?*

ISSUE IDENTIFIED AND ADDRESSED IN THE PEER-REVIEWED PUBLICATION

Has the adapted LEA approach been translated into evaluation practice and informed policy decisions?

Paper 5 (page 213) describes how this framework fits within the context of related approaches internationally and provides real world examples of when the methodology has been mis-applied and when it has been used correctly to inform reimbursement decision-making. There is unavoidable duplication of some material in Paper 5 with the previous papers in this thesis but it does help to draw the body of work together into a cohesive whole.

Statement of authorship

Merlin T. The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines. *Personalized Medicine*, 2014; 11(4): 435–448. doi: 10.2217/PME.14.28. Available at: http://www.futuremedicine.com/doi/abs/10.2217/pme.14.28

AUTHOR'S CONTRIBUTIONS

Tracy Merlin (Candidate)

Conceived the methodology that is reviewed in this paper, wrote the manuscript plan, reviewed and analysed all of the literature included in the paper, extracted data from policy public summary documents regarding the outcomes of the personalised medicine applications for public funding, developed all figures and tables, wrote the manuscript and revised it after it underwent peer review by four reviewers.

Merlin T. The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines.

Personalized Medicine, 2014; 11(4): 435–448. doi: 10.2217/PME.14.28. Available at: http://www.futuremedicine.com/doi/abs/10.2217/pme.14.28

Paper 5 is reproduced as follows –

The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines

Author name and affiliation

Tracy Merlin

Director, Adelaide Health Technology Assessment (AHTA)

School of Population Health, Mail Drop DX 650 545

University of Adelaide, South Australia, Australia 5005

Email: tracy.merlin@adelaide.edu.au

SUMMARY

It is uncommon to find published clinical trials that measure the health benefits of medical testing. As a consequence, policy makers often have to decide whether access to, or public funding of, medical tests is warranted without knowing the clinical impact of testing on the patient. In the situation where a policy maker is considering a companion genetic test and tailored drug therapy, deficiencies in the evidence base are exacerbated because two technologies need to be assessed and the proposed genetic biomarker needs to be validated.

The Linked Evidence Approach (LEA) is a methodology that was developed in 2005 to cope with inadequacies in the evidence supporting medical test evaluations. In 2010 the approach was adapted to the evaluation of pharmacogenetic interventions. This article describes how LEA and similar analytic frameworks are used internationally, highlights particular challenges with the approach, and proposes ways that LEA might be applied to pharmacogenomic interventions.

Keywords

Individualized medicine; Pharmacogenetics; Health policy; Biological Markers; Economics, Pharmaceutical; Health Services Accessibility/economics; Molecular Targeted Therapy; Outcome Assessment (Health Care)/economics; Outcome Assessment (Health Care)/methods; Technology Assessment, Biomedical; Diagnostic test approval

BACKGROUND

TESTS ARE NOT PERFECT

In the past, tests have not received the same degree of scrutiny by policy makers, when formulating public funding decisions, as therapies have traditionally received. This could be because therapies have immediate and obvious impacts on patient health, whereas the consequences of testing are indirect and not immediately observable. However, tests are far from perfect and may result in immediate harm (associated with the procedure), or harm secondary to inaccurate information. Factors that can affect test accuracy include: poor communication and insufficient understanding of testing procedures; inappropriate test selection/ordering and interpretation of results; patient and/or specimen misidentification; inadequate specimen obtained for testing; specimen collection errors; and specimen contamination [1].

Advances in genetic testing have raised further quality challenges. A survey of all pathology laboratories in Australia, where test regulation and accreditation is fairly stringent, reported that genetic tests ranged considerably in their ability to correctly identify patients who have the target condition [2]. The estimated analytic sensitivities were largely concordant across the 52 laboratories surveyed (Table 1) and suggest that, in the most extreme example, up to 80% of patients with a condition could receive a false negative test result.

Table 1 **Range of genetic test methods offered by 52 Australian laboratories and their estimated analytic sensitivity†**

Method	No. of types of tests based on method	No. of reports	Range of expected analytic sensitivity	No. of discordant reports‡
Diagnostic genetic tests (mutation identification)				
Mutation screening	48	48	60% to >94%	nil
Sequencing	146	174	20% to >94%	4
Sequencing plus MLPA	51	101	60% to >94%	2
Specific assays	108	177	<20% to >94%	2
FISH	25	47	60% to >94%	1
Somatic testing				
Mutation screening	2	2	80% to >94%	nil
Sequencing	3	3	80% to >94%	nil
Specific assays	29	66	40% to >94%	8
FISH	32	51	>94%	nil

† Data taken with permission from [2].

‡ Expected analytic sensitivity for the same method and type of test varying by >20% in different laboratories.

FISH: fluorescent *in situ* hybridisation; MLPA: multiplex ligation-dependent probe amplification.

Even a small amount of imperfection in test accuracy can undermine the commercial viability and cost-effectiveness of a companion diagnostic and therapy [3]. For example, if a test has a high false positive rate, more patients would receive an inappropriate treatment resulting in an increase in treatment costs for no additional health gain. The test/therapy combination would not be cost-effective and unlikely to receive government or insurance subsidization. The impact of test performance on the cost-effectiveness of therapies might have spurred the introduction of methods guidance in the UK [4, 5], Europe [6] and the USA [7-9] (2008-2012) regarding the evaluation of tests for regulation and reimbursement purposes. Australia – perhaps because it had a policy mechanism specific to the evaluation of tests as part of a medical service – produced its guidance on medical test evaluation for reimbursement purposes in 2005 [10].

Evaluation of tests, when performed, is often restricted to test performance only (sensitivity, specificity etc) with little consideration given to the impact on patients of receiving a false negative result – leading to delayed treatment – or of a false positive result – leading to inappropriate treatment (no additional clinical benefit and additional harms from toxicity) [11]. This, in part, may be because many geneticists and laboratory scientists believe that the information from testing has value in and of itself [12]. However, it may also be due to the lack of direct trial evidence assessing the impact of testing on patient health outcomes.

Di Ruffano et al (2012) conducted a capture-recapture analysis using two searches (broad and specific) from the Cochrane CENTRAL hand searched trial database to estimate the number of randomised controlled trials published on diagnostic tests between 2004-2007. Of the 23,888 randomised controlled trials retrieved, 135 were found to be diagnostic randomised controlled trials. The capture-recapture analysis estimated 37 diagnostic trials were published per year for the 4 years [13]. This is in contrast to the 5,938 therapeutic trials per year that were known to have been published.

In general, if a test performs poorly in a diagnostic effectiveness trial, false positive and false negative test results will be reflected in the measured health outcomes of patients. However, as trial evidence of the impact of medical, including genetic, tests on the health outcomes of patients is often scarce, policy makers are faced with making decisions on access to, and reimbursement of, tests on the basis of incomplete and uncertain information.

THE ‘LINKED EVIDENCE APPROACH’ (LEA)

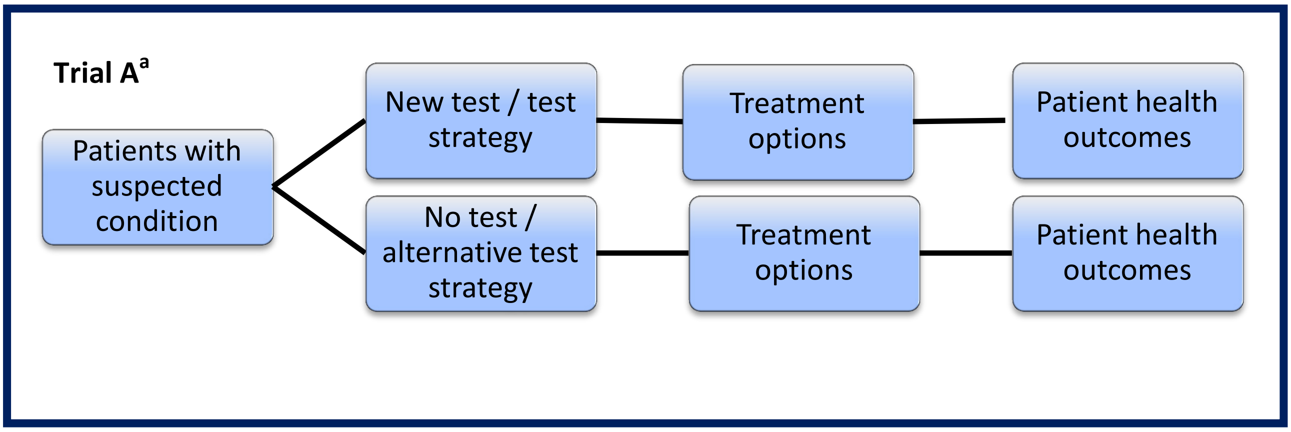
To address this evidence gap, in 2005 a methodology was published that aimed to provide the maximum amount of information on test effectiveness and cost-effectiveness to Australian policy makers [10, 14]. This “linked evidence approach” (LEA) involves the narrative linking of evidence assessing components of a test-treatment pathway in order to predict the likely impact of testing on patient health outcomes. The method was informed by criteria developed by Fryback and Thornbury (1991) to assess the efficacy of diagnostic imaging tests [15]. These criteria include technical efficacy, diagnostic accuracy, diagnostic thinking (change in diagnosis), therapeutic efficacy (change in management) and patient outcome efficacy (change in health outcomes). The method was also informed by analytic frameworks pioneered by the United States Preventive Services Task Force (USPSTF) to identify key questions that guide clinical practice guideline development [16]. These frameworks address both the harms and benefits of medical testing on the patient [17-19].

The Guidelines on LEA [10] recommend the systematic review and narrative linking of evidence under certain conditions. This linking of evidence would occur in instances where direct trial evidence of the effect of testing on patient health outcomes is not available or inadequate for decision making purposes [10]. The evidence linkage is primarily undertaken as a methodological substitute for the ideal hypothetical trial that would be used to measure the health benefits of the test on patients (Figure 1) [14]. The trial design is broken down into its elements and used as the template for the decision analytic model which integrates the information and determines whether the new test is both effective and cost-effective [14]. Evidence addressing each element of the decision analytic model is systematically acquired and rigorously critiqued.

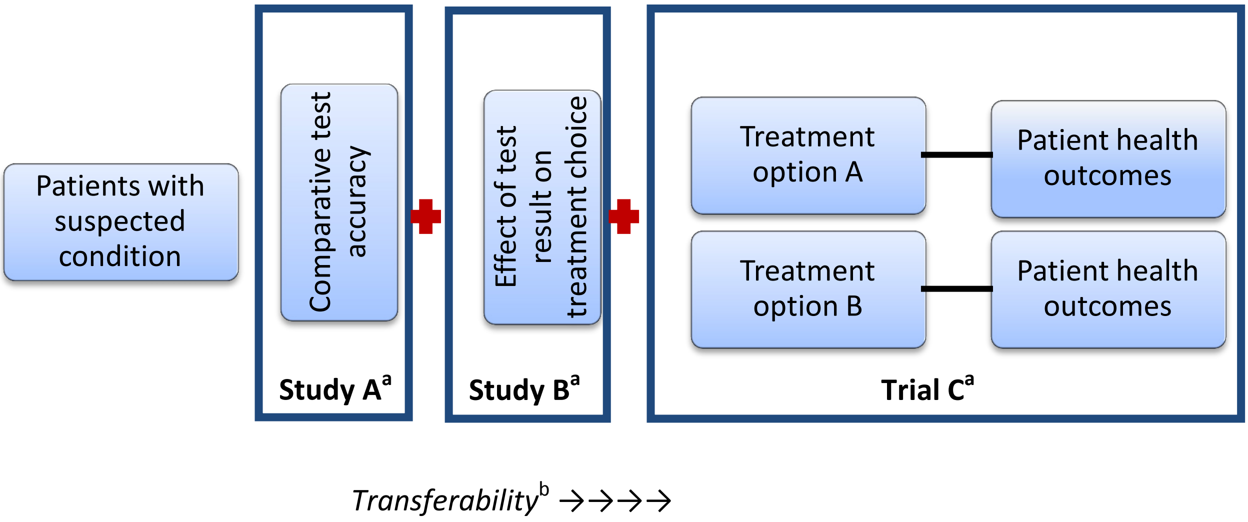
A systematic literature review of Australian health technology assessments (HTAs) [20] reported that the method had been applied to 85 patient indications for testing between 2005 and 2012. The method was used on tests for diagnostic, staging and screening purposes, as well as for genetic tests [21, 22].

Figure 1 Estimating clinical effectiveness of test using linked evidence approach

Direct evidence of test effectiveness



Linked evidence of test effectiveness



[a] The results from one or more of these trials (including meta-analyses) would form the direct evidence base showing test impact on patient health outcomes

[b] The results from one or more of each of these types of studies and trials (including meta-analyses) would form the linked (or indirect) evidence base predicting the test impact on patient health outcomes.

[c] Populations, tests and outcome definitions should be transferable (similar) across linkages.

Adapted from [10]

In the original Guidelines on LEA, it was noted that if there was evidence that the patients eligible for the new test are similar (*transferable*) to those patients currently receiving treatment for the condition, the findings of test accuracy studies could be considered sufficient to determine the clinical utility of the new test [10]. This means that the findings from studies that report the effect of the comparator (current) test on (i) the selection of treatment options for patients, and (ii) the flow-on effects of treatment on the health of these patients, could be used in a decision analytic model to simulate or predict the health benefits associated with the new test.

A key element of this *transferability assumption* is Fryback and Thornbury's criterion on change in management (see Figure 1). If the test, no matter how accurate, does not change the treatment options or management offered by the health professional to a patient, then there will be no impact of the test on the patient's health status. This means that there is no need to evaluate the safety and effectiveness of the treatment options in the linked test-treatment pathway. Put simply, the new test would create an additional cost for no additional patient health benefit and so would be considered cost-ineffective. A decision framework has been developed to assist those applying the linked evidence approach, including providing guidance on the type and extent of evidence needed in a linkage [20].

Staub et al (2012) reviewed the methods used in English-language HTAs of medical tests to assist policy makers with regulation and reimbursement decisions [11]. The review encompassed the work of 18 HTA agencies in eight countries and found that 48% of the 149 HTAs reported only on test accuracy, 11% on test accuracy and the impact of treatment, 24% on test accuracy and impact on patient management, and 17% on all linked evidence elements. Of the 17 HTAs reporting the use of an analytic framework, Fryback and Thornbury's criteria was cited as the integrative framework in five HTAs, while the Australian linked evidence approach was cited in 12 HTAs.

The use of evidence linkage and integrative frameworks in medical test evaluation, in order to inform policy decisions, is increasing and is now recommended by many of the major technology assessment organizations internationally [5-7, 10, 23].

GENETIC TESTING AND METHODS OF EVALUATION

According to a status report published in 2007, approximately 6.8 billion laboratory tests are performed annually in the USA [1]. The revenue, spending, and test volume of the clinical laboratory testing market has grown steadily. More than 4,000 laboratory tests were available for clinical use in 2007, of which 1,162 tests were reimbursed by U.S. Medicare [1].

A key initiative developed specifically for the evaluation of *genetic* tests by the Centers for Disease Control and Prevention in 1997, was the creation of the Office of Public Health Genomics which in turn sponsored the ACCE Model project in 2000 [12]. The ACCE Model was the first publicly-available analytical process for specifically evaluating scientific evidence on emerging genetic tests. The model presents 44 questions that need to be addressed to determine the Analytic validity, Clinical validity, Clinical utility, and Ethical, legal, social implications (ACCE) of a genetic test [8]. It was applied to the assessment of several genetic testing technologies, including – in the first instance - an assessment of prenatal screening for cystic fibrosis via carrier testing for CFTR mutations [24], as well as in mini-reviews of genetic [25] and pharmacogenetic tests [26].

In 2004 the Office of Public Health Genomics established the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative and, in 2005, a Working Group was created to develop an evidence-based process for assessing genetic tests and other clinical applications of genomic technology [27]. The EGAPP initiative commission's systematic reviews of genetic tests that address key questions developed using USPSTF-style analytic frameworks as well as elements of the ACCE model.

The EGAPP Working Group has reported difficulties in generating evidence-based recommendations regarding the clinical utility of different genetic tests because of the scarcity of good quality evidence. The Working Group speculated that randomized controlled trials were lacking because of the constraints imposed by time, recruitment and resources when designing and implementing studies on testing for rare conditions or when the downstream effects of treatment involved relatively small effect sizes compared with usual care. The Working Group methods allow for the use of observational or nonrandomized evidence when high level evidence is not available [12, 27].

This is very similar to the Linked Evidence Approach whereby different hierarchies of evidence and appropriate critical appraisal techniques are used to address different types of questions in the linkage [10]. The evidence hierarchy that is used addresses questions on diagnostic accuracy, interventions (relevant for direct evidence of diagnostic effectiveness and change in management studies), aetiology, prognosis and screening. It was originally produced to assist with clinical practice guideline development [28].

The synthesis of evidence supplied to the EGAPP Working Group is used to formulate recommendations on the use of genetic and genomic tests in clinical practice [29, 30]. EGAPP information is disseminated to various stakeholders but the EGAPP initiative is not directly responsible for the regulation or reimbursement of genetic tests [27].

Molecular testing has seen recent, rapid and escalating growth in developed economies, especially in the fields of infectious diseases and oncology [1, 2]. This may, in part, be because these tests can identify particular genetic biomarkers that predict the therapeutic performance of specific drugs (pharmacogenetic application). Molecular testing methods identify specific sequences of human DNA or RNA to identify errors (mutations) that may or may not be associated with disease ie single nucleotide polymorphism, gene insertion, deletion or rearrangement. In Australia, requests for molecular tests increased 2.8 times from 2006 to 2011 [31]. Somatic genetic tests and diagnostic (mutation identification) tests each increased by 23% from 2006 to 2007, whereas pharmacogenetic tests increased by 101% [2]. Pharmacogenetic interventions involving a companion genetic test and tailored drug treatment have been emerging over the last decade and this has created challenges for existing regulation and reimbursement technology approval mechanisms.

PRACTICAL DIFFICULTIES WITH THE EVALUATION OF PHARMACOGENETIC INTERVENTIONS FOR POLICY DECISIONS

There are some impediments to the evaluation of companion diagnostics and pharmaceuticals for reimbursement decisions. Terasawa et al reported large differences in the way genetic factors are grouped and analyzed within pharmacogenetic studies, making it difficult to combine and interpret findings across studies [32].

Similarly, Laksman and Detsky (2011) noted that *"the huge number of genes available for demonstrated associations and the wealth of information being churned out at an increasing pace leave some with the feeling that we are producing more data than we can analyze or understand"* [33].

Perhaps this is the reason why Holmes et al, when they conducted a systematic review of pharmacogenetic studies in 2009, found that the ratio of commentary/reviews to original research in the available evidence base (4,674 papers spanning 1967-2007) was 25:1 [34]. Researchers and clinicians appear to be struggling to process all of the available information into a cohesive whole.

Holmes et al. also reported that of the original pharmacogenetic studies obtained, the majority focused on candidate genes rather than genome-wide analysis, were of inadequate sample size, provided suboptimal capture of genetic variation and were characterized by 'significance chasing' and reporting bias [34]. These problems lead to a failure to replicate and validate genetic associations [35, 36].

Similarly, systematic literature reviews on selected pharmacogenetic tests for cancer treatment found problems with the evidence-base on *CYP2D6* for tamoxifen in breast cancer, *KRAS* for anti-EGFR antibodies in colorectal cancer, and *BCR-ABL1* for tyrosine kinase inhibitors in chronic myeloid leukemia [32]. Studies had small sample sizes and so could not reliably identify small treatment effects. Additional problems that were observed, irrespective of whether the genetic biomarker was a germline polymorphism (*CYP2D6)* or a somatic mutation (*KRAS*, *BCR-ABL*), included the lack of formal assessment for treatment-by-biomarker interactions (ie treatment effect modification) and the use of surrogate short term outcomes of treatment failure rather than patient-relevant outcomes such as overall survival or progression-free survival. Terasawa et al. noted that adjustments for potential confounding factors were often not based on sound epidemiological principles and that adjustments for multiple comparisons were often not documented [32]. The other limiting factor in the evidence base was that multiple studies on each topic frequently originated from a limited number of specialized centers, meaning that populations could overlap and potentially threaten the generalizability of the findings [32].

There have been recent attempts to strengthen pharmacogenetic research, including conducting post hoc analyses of completed drug trials by genotyping prospectively banked tissue samples from patients prior to them being allocated to a treatment arm. Genotyping of tissue can be performed after the trial has ended, although the benefit of randomization in balancing confounding variables between trial arms is lost. An example of this approach is the pharmacogenetic *KRAS* and *CYP2D6* trials [37-39].

Despite some recent improvements in trial design, the overall evidence base available to inform policy makers on the safety, effectiveness and cost-effectiveness of pharmacogenetic interventions is poor, piecemeal and problematic to evaluate and synthesize.

ADAPTING LEA TO PHARMACOGENETIC INTERVENTIONS

Meckley and Neumann (2010) analysed reimbursement decisions from NICE, AETNA, CIGNA, Premera (Blue Cross), and Centers for Medicare and Medicaid Services with regard to six case studies – namely, *HER2/neu* and trastuzumab; hepatitis C genotyping and ribavirin/pegylated interferon; *UGT1A1* and irinotecan; *VKORC1/CYP2C9* and warfarin; *BRCA1/2* with prophylactic surgical measures; and OncotypeDX with chemotherapy. The authors observed that the strength of evidence available to support the clinical benefits/harms of use of a personalized medicine was the key determinant in predicting positive reimbursement decisions [40]. Similarly, Faulkner et al. (2012) suggested that efforts to develop a coherent system of evaluation of medicines targeted to patients with specific genetic biomarkers have been hampered by the available evidence base which does not fit the established approaches to test and drug evaluation for reimbursement decisions [41].

This was a problem encountered by Terasawa et al. in their systematic reviews of selected pharmacogenetic tests [32]. The lack of a conceptual framework for integrating the disparate pieces of evidence meant that there were difficulties in developing a coherent picture from the mix of genetic association studies, predictive accuracy studies and trials showing treatment effects in patients with a biomarker. It would also be difficult to determine what key evidence, if any, was missing.

Evidence on companion tests and drugs presented for reimbursement decisions have typically concentrated on assessing the clinical benefit of the drug in patients with a particular genetic characteristic, with little attention being paid to: (1) whether the proposed test or combination of tests is accurate at identifying that specific genetic biomarker, or (2) isolating whether that particular genetic biomarker is the target (or effect modifier) for the drug, as opposed to being a consequence of other characteristics that may be defining or responsible for that particular patient group responding to the therapy. These other characteristics include determining whether the genetic biomarker is simply a prognostic factor that predicts improved patient health outcomes irrespective of the treatment offered, or whether the observed effect is due to measured (or unmeasured) confounding factors introduced through a non-randomized (or non-stratified) comparison by biomarker status.

In an attempt to address the deficiencies in the available evidence base for pharmacogenetic interventions and to provide a conceptual framework to incorporate the disparate pieces of evidence, we adapted the linked evidence approach used for test evaluation to personalized medicines [42]. The aim was to develop an approach that was flexible and adaptable to the different types of evidence generated in the research community and yet still provide robust evaluations of the safety, effectiveness and cost-effectiveness of both test and drug. Another key aim was to make areas of clinical risk and cost uncertainty transparent to policy makers.

THE CO-DEPENDENT TECHNOLOGY EVALUATION FRAMEWORK

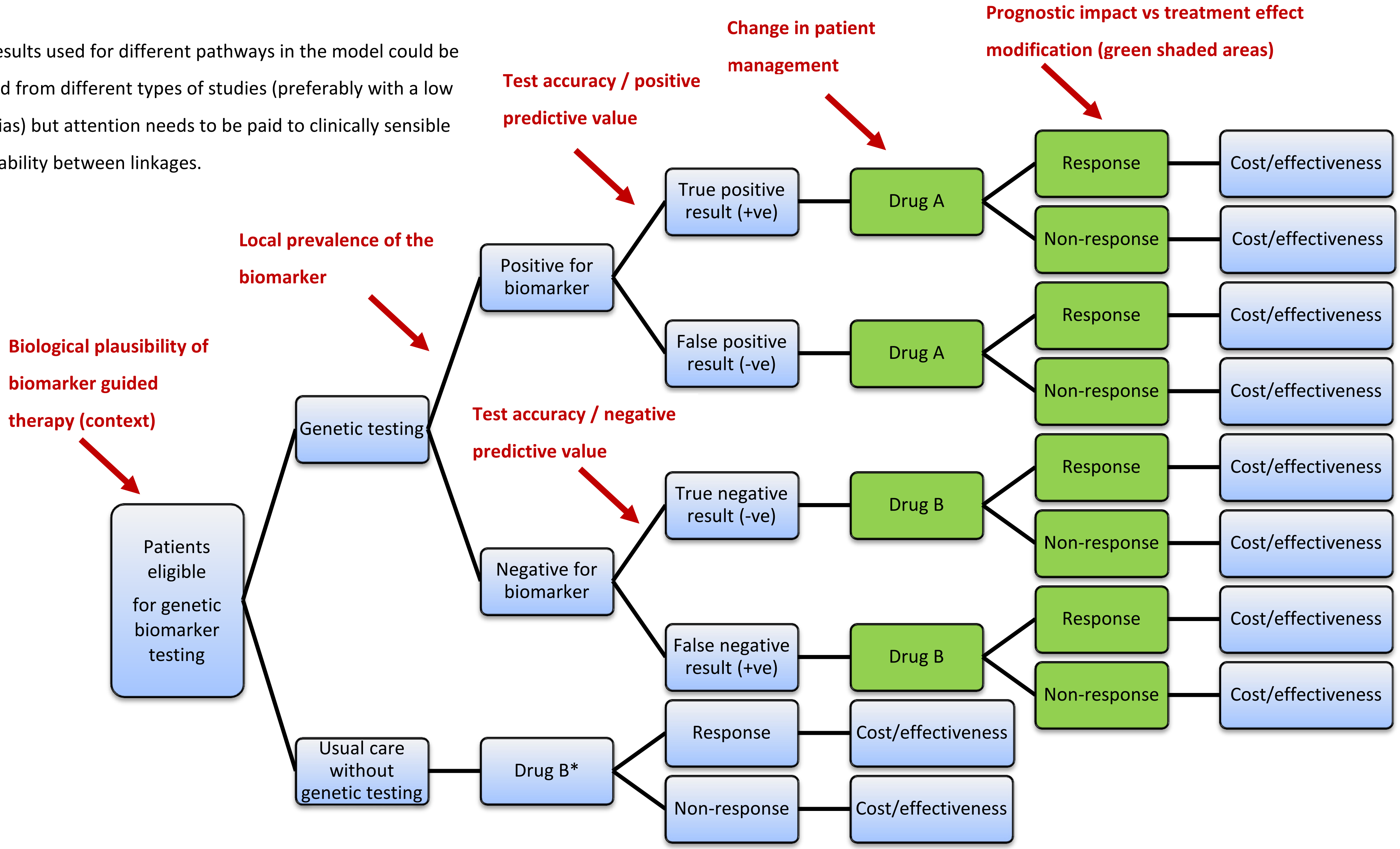
The framework for assessing pharmacogenetic interventions in Australia has been reported elsewhere [42]. In summary, the approach uses the hypothetical framework of a double randomized controlled trial as a template for determining what information elements are needed when linking evidence – this trial design is unlikely in practice but consists of all the elements needed to evaluate the test, drug and the interaction between the two. The use of a hypothetical framework had been suggested previously for undertaking test evaluations in LEA [14]. Consistent with this, our framework was informed by an awareness of the importance of maintaining transferability across evidence linkages and a need to define the likely biases if the transferability assumptions could not be fulfilled. Complementary to this approach, given the largely observational evidence base associated with pharmacogenetic interventions, elements suggested by Bradford-Hill to determine causation [43] were incorporated within the framework. This included a biological plausibility (or 'rationale') criterion to justify, using molecular biological or pharmacological principles, the plausibility of treatment effect modification (or interaction) between the biomarker itself and the drug, or alternatively between the drug and another factor for which the biomarker is a proxy. A criterion was also included to ascertain whether there is any other validated biomarker which predicts variation in the comparative treatment effect (between using the drug and not using the drug) (Additional File 1 of [42]).

In addition, information was requested to ascertain whether the proposed genetic biomarker is a prognostic factor or a treatment effect modifier and to determine the strength of any treatment effect modification (Additional File 1 and Additional File 3 of [42]).

The end product in 2010 was 79 information requests that have been incorporated into Government guidance for applicants seeking reimbursement of companion tests and drugs (Additional File 1 of [42]). The optimal study designs needed to address some of the 79 items and methods for presenting information and reducing bias were also outlined, often with reference to current methodological norms. Key information requested as part of this evaluation framework is outlined in Figure 2, as it relates to a simple decision analytic model.

One of the advantages of the adaption of LEA to pharmacogenetic applications as described in Figure 2, is that the clinical evidence is systematically acquired and critically appraised for internal and external validity. This occurs prior to use as inputs and transition probabilities in the decision analytic modeling underpinning the economic evaluation [44]. Cost-effectiveness estimates are therefore likely to be more realistic and arguably less biased and sensitivity analyses can be used to vary key clinical (eg harms from inappropriate treatment) and cost inputs over which there is uncertainty.

Note: Results used for different pathways in the model could be extracted from different types of studies (preferably with a low risk of bias) but attention needs to be paid to clinically sensible transferability between linkages.


Prognostic impact vs treatment effect modification (green shaded areas)
Change in patient management
Test accuracy / positive predictive value
Local prevalence of the biomarker
Biological plausibility of biomarker guided therapy (context)
Test accuracy / negative predictive value
Patients eligible for genetic biomarker testing
Genetic testing
Positive for biomarker
True positive result (+ve)
Drug A
Response
Cost/effectiveness
Non-response
Cost/effectiveness
False positive result (-ve)
Drug A
Response
Cost/effectiveness
Non-response
Cost/effectiveness
Negative for biomarker
True negative result (-ve)
Drug B
Response
Cost/effectiveness
Non-response
Cost/effectiveness
False negative result (+ve)
Drug B
Response
Cost/effectiveness
Non-response
Cost/effectiveness
Usual care without genetic testing
Drug B*
Response
Cost/effectiveness
Non-response
Cost/effectiveness

* Drug B is usual care but a scenario could be tested whereby Drug A (the proposed pharmacogenetic drug) is offered without the companion genetic test

Figure 2 Using the results from linked evidence as inputs in a simple decision analytic model to estimate the comparative costs and effectiveness of a pharmacogenetic intervention

INTERNATIONAL EXPERIENCE

The method used by the Expert Working Group of EGAPP when evaluating genetic tests – including pharmacogenetics [45-48] - and developing practice recommendations, involves the synthesis of a chain of evidence, along with consideration of ACCE model criteria [12, 27]. As of February 2014, EGAPP had recommended 36 tier 1 pharmacogenetic interventions – 30 of which are used to guide cancer treatment - with a base of synthesized evidence supporting implementation into clinical practice.[23] In 2013, the Working Group started investigating basic modeling techniques to deal with the lack of available evidence on genetic tests [12]. The Working Group supports the need for additional approaches and methods for evidence generation and innovative modeling strategies. They also recognize that basing recommendations on evidence from poorer quality studies (risk of bias) will require accepting a higher risk of providing no net health benefit or introducing patient harms. As such, consistent with the Australian approach, they indicate that there needs to be a careful consideration of the risk of harm balanced with the opportunity for benefit when considering a genetic test, and to develop a plan for addressing evidence gaps [12].

The evaluation of pharmacogenetics in Europe is lagging a little behind other major developed health systems and part of the reason relates to structural impediments. In the United Kingdom, the National Institute of Health and Care Excellence (NICE) is making some headway in evaluating products that have dissimilar regulatory evidence requirements, such as CE-marked and in-house laboratory tests, as well as pharmaceuticals. Other European countries, however, often have completely unlinked processes, which makes it particularly difficult to evaluate both the test and the drug in a coherent manner [49]. Health technology assessments conducted in Europe again exemplify the learning curve associated with identifying evidence that can properly populate economic models to inform policy makers of the cost-effectiveness of pharmacogenetic interventions [50].

To date, NICE has completed one pharmacogenetic test appraisal under their new diagnostic assessment program [51], with a further four underway. Although several pharmacogenetic interventions have been evaluated using the NICE Technology Appraisal process, the required evidence base mainly pertains to the clinical effectiveness and cost-effectiveness of

[23] http://www.cdc.gov/genomics/gtesting/tier.htm

the drug component of the technology rather than determining whether the genetic test is effective in identifying the eligible patient population [50]. This means that until recently there has been minimal scrutiny as to whether the drug is being appropriately targeted and/or inappropriately replacing effective treatments as a consequence of incorrect test results.

The NICE diagnostic test assessment program recognizes the utility of the linked evidence approach when modeling the effects of testing. The methods manual states

> *"If data on the final patient outcomes of a diagnostic technology are not available, it may be necessary to combine the evidence from different parts of the care pathway. In this case the linkages between diagnosis, treatment and final outcomes need to be specified, and relevant data about those linkages needs to be obtained and reviewed. Data about test accuracy and the nature of the care pathway and its outcomes can be used to create an assessment comparing the effect of different testing approaches."* [5]

The experience in Australia of assessing companion test-drug combinations for reimbursement decisions has accelerated since the introduction of the co-dependent technology evaluation framework. Apart from the five pharmacogenetic interventions that were used as case studies for the development of the framework in late 2010, there have been a further nine companion test-drug evaluations conducted since the evaluation framework was finalized. Five of these interventions were reimbursed, two were rejected and two have been deferred (as of March 2014, see Table 2). The evaluation framework appears to be working well in terms of the technical requirements being met and evaluated in a fairly timely way. Given the lack of available direct evidence, many of the applications have had to provide linked evidence to address the 79 information requests [42]. Areas of uncertainty are made clear and the use of specific inputs in the models can be critically appraised. This has allowed Government to negotiate reduced prices as a consequence of the uncertainties identified [52-54]. It has also allowed subsidized market access for products that would not previously have been considered as having an acceptable evidence base because of the lack of direct evidence [42], and perhaps rejected for public funding.

There have, however, been lengthy delays in some instances because of the need to coordinate policy processes – perhaps similar to the European situation. As Australia has independent committees evaluating each of the test and drug [55], there have been several deferrals in order to seek advice from the other Committee and so there have been delays before coordinated advice could be provided to Government.

Table 2 Pharmacogenetic interventions submitted for a reimbursement decision in Australia after introduction of the co-dependent evaluation framework

Condition	**Genetic test/ biomarker**	**Drug**	**Current status (March 2014)**
Locally advanced or metastatic melanoma	*BRAF* V600 mutation test	Dabrafenib	Funding of test and drug recommended. Prospective data collection on test utilization to be undertaken to inform the risk-share arrangement. [52, 59]
HIV infection	Genotype test for HIV tropism	Maraviroc	Funding of test rejected on basis of insufficient evidence that test adequately distinguishes between HIV-infected individuals who should and should not receive Maraviroc. [60, 61]
Gastric cancer	*HER2* gene amplification	Trastuzumab	Recommendation deferred as further consolidation of information required between committees assessing test and drug. [62, 63]
Locally advanced or metastatic NSCLC	*EGFR* mutation test	Gefitinib	Funding of test and drug recommended. [53, 64]
Locally advanced or metastatic melanoma	*BRAF* V600 mutation test	Vemurafenib	Not considered cost-effective. Recommendation deferred pending further negotiation with the sponsor. Sponsor indicated it is unlikely to re-submit an application. [65, 66]
Locally advanced or metastatic NSCLC	*EGFR* mutation test	Erlotinib	Funding of test and drug recommended. [54, 67]
Breast cancer	*HER2* IHC test	Neoadjuvant trastuzumab	Funding of test and drug recommended. [68, 69]
NSCLC	*ALK* test	Crizotinib	Recommendation deferred. Acceptable comparative

			effectiveness. Unacceptable cost-effectiveness and so negotiation with sponsor commenced. Awaiting decision by other committee with regard to test listing. [70]
Metastatic colorectal cancer	*KRAS* testing	Panitumumab	Funding of drug recommended. Test currently listed but additional information being sought with regard to role of wider *RAS* testing. [71]

NSCLC: non-small cell lung cancer; HIV: human immunodeficiency virus; IHC: immunohistochemistry

LIMITATIONS OF LEA

One of the limitations of LEA is finding evidence to support all areas of the linkage. This does not mean that the companion test and drug are not beneficial, only that there is insufficient evidence to make a determination either way. The main area of difficulty is identifying evidence of the likely treatment effect of the co-dependent pharmaceutical in patients without the biomarker or in an untested population.

Some researchers suggest that in the biomarker development phase, once a specific treatment is established, it is unethical to randomize patients to a control arm of no therapy until there are sufficient data on the biomarker’s clinical validity [56]. Industry researchers have also suggested that if a pharmaceutical has been developed to target a particular biomarker it would be unethical to randomize patients without the biomarker to receive that pharmaceutical. Studies are less likely to be mounted when there is a risk of harm.

Either way, when there is a lack of information to support those aspects of the linkage, modeling can be undertaken to determine whether the conservative assumptions of clinical benefit/harm that necessarily need to be made are likely to have an impact on the overall clinical and cost-effectiveness of the pharmacogenetic intervention. However, modeling cannot substitute for good trial data and trials should be performed when there is equipoise regarding likely benefits and harms; such that even if a pharmaceutical has been developed to address a particular biomarker it needs to be confirmed through robust trial evidence that the biomarker is relevant. There have been instances in the past where researchers have wrongly attributed a clinical benefit or harm to an interaction between genetic biomarker and drug [27, 56].

An example of the limitations associated with LEA can be drawn from the first pharmacogenetic test evaluation conducted by an external assessment group commissioned by NICE under the diagnostic assessment program [51]. Three modeling methods were provided to estimate the cost-effectiveness of different EGFR tyrosine kinase tests in adults with locally advanced or metastatic lung cancer experience. These included: 1) a ‘comparative effectiveness’ analysis which only used direct evidence of testing on final health outcomes; 2) a ‘linked evidence’ analysis which included evidence of test accuracy for predicting response to tyrosine kinase inhibitors, according to EGFR mutation status, and the clinical effect was estimated from other trials; and 3) an 'assumption of equal prognostic value’ analysis when no data were available on either the comparative effectiveness or the accuracy of EGFR mutation tests for predicting response to tyrosine kinase inhibitors. The incremental cost effectiveness ratio (ICER) produced using either the direct or linked evidence approaches were very similar (the ICER for the prognostic value analysis could not be calculated). However, the test accuracy estimates used in the linked evidence model were considered unreliable as they were sampled from different populations, using different test methods and different definitions of resistance mutations. As a consequence of this, the cost-effectiveness analysis was not considered robust by NICE. Despite these uncertainties, a decision was made to recommend EGFR testing with Sanger sequencing based methods, the Cobas EGFR mutation test and the TheraScreen EGFR PCR kit [51].

This example highlights how, even using linkage methods to extract the most out of the pharmacogenetic data available, it is critical to ensure that the evidence used to derive inputs for modeling is internally consistent, clinically meaningful and *transferable* across the linkages ie similar populations, tests, biomarker definitions and outcome criteria are used.

FUTURE PERSPECTIVE

Despite the difficulties associated with the evaluation of a test to identify a single genetic biomarker that could guide treatment with a single pharmaceutical (sometimes over a background of usual care), pharmacogenetics is actually a simple example of a personalized (or stratified) medicine. Regulatory and reimbursement consideration of the use of genome testing paired with targeted prophylactic and symptomatic treatment raises further complexities – technical, legal, ethical and social.

The *genome* is the entirety of an organism's hereditary information. The introduction of DNA microarray platforms and projects like IT Future Of Medicine (ITFOM)[24] – a consortium of partners whose role it is to construct computational models of the molecular and anatomical biological processes that occur in every human – means that integrated maps of human genomes across diverse populations are being developed and can be used to predict and validate genome wide association studies [57]. In the future, genomic regions associated with human disease will be able to be isolated and both prophylactic and symptomatic treatments will be able to be targeted to each individual’s genomic profile [33].

Proponents of this type of research have indicated that in these circumstances randomized controlled trials will become obsolete and that n-of-1 trials would be the only possible alternative for determining the clinical benefits and harms of individualized therapies. However, it is unlikely that payers would subsidize population-based testing and treatments on the basis of n-of-1 studies [58]. An alternative could be the use of complementary evaluation processes. Linked evidence approaches addressing stratified personalized medicines (pharmacogenetics) could be used to inform policy making and at the same time inform the prediction models developed for individualized medicines. The rigorous approaches used to inform reimbursement decisions would mean that there is some assurance that the genetic association underpinning a stratified personalized medicine is valid. The individualized risk prediction models could then be assessed in studies that compare treatment/prophylaxis guided by ITFOM (or other) genome risk prediction models versus treatment guided by clinical judgment (or a previous version of a risk prediction model). Randomized controlled trials could be used to assess short term benefits/harms of the two types of treatment targeting models and/or prospective cohort studies, registries or comparative effectiveness research [56] could be used to assess long term benefits/harms. The risk prediction models would likely need review and re-specification on an ongoing basis as new developments and understandings occur in the personalized medicine evidence base.

CONCLUSION

Momentum is gaining in the use of linked evidence approaches to identify relevant data on pharmacogenetic interventions and to synthesize the findings in a robust and yet flexible

[24] http://www.itfom.eu/images/downloads/ITFoM_fet11_setting%20the%20scene%2017_05_2011.pdf

manner for reimbursement decisions. However, the approach should be used and appraised cautiously as the body of accumulated evidence needs to maintain internal clinical coherence. The approach is meant to be a proxy for an ideal trial design, it is not meant to be a 'Frankenstein creation' for incorporating disparate or biased pieces of evidence. The approach, if used well, can explicate the patient risks and benefits from pharmacogenetic interventions, enable value for money determinations to be made, and assist policy makers to formulate informed reimbursement decisions.

FINANCIAL DISCLOSURE / ACKNOWLEDGEMENTS

The author is regularly contracted by the Australian Government Department of Health to undertake evaluations of genetic tests and personalized medicines to inform reimbursement decisions. The author worked with Government to develop the co-dependent technology evaluation framework.

With thanks to Claude Farah, David Tamblyn and Skye Newton for their comments on the manuscript.

The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

EXECUTIVE SUMMARY

Tests are not perfect

- Inaccurate tests can lead to delayed, inappropriate or harmful treatment.
- Trial evidence of the direct impact of medical tests on the health outcomes of patients is scarce.

The ‘Linked Evidence Approach’ (LEA)

- The ‘linked evidence approach’ (LEA) is an integrative framework, developed in Australia, that narratively links evidence addressing key elements of the test-treatment pathway. The framework used to link the disparate pieces of evidence is a hypothetical randomized controlled trial designed to determine the diagnostic effectiveness of the new test. As direct trial evidence is often absent, this linkage approach maximizes the available information for policy makers so that the likely impact of the new test on patient health outcomes can be determined.
- The findings from these evidence linkages can be used as inputs in decision analytic modeling to predict whether the new test provides good value for money when compared to existing diagnostic approaches.
- LEA was informed by methods pioneered by the United States Preventive Services Task Force (USPSTF) and the efficacy criteria proposed by Fryback and Thornbury (1991).

Genetic testing and methods of evaluation

- The Office of Public Health Genomics in the United States has produced two key programs for the evaluation of emerging genetic tests - the ACCE Model Project and the Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative.
- Pharmacogenetic interventions involving a companion genetic test and tailored drug treatment have been emerging over the last decade and this has created challenges for existing regulation and reimbursement technology approval mechanisms.

Adapting LEA to pharmacogenetic interventions for policy decisions

- Pharmacogenetic interventions often have a poorer evidence base than other testing interventions. However, research suggests that the strength of evidence available to support the clinical benefits/harms of these interventions is the key predictor of positive reimbursement decisions.
- In order to address the deficiencies in the typical evidence base for pharmacogenetic interventions, an evaluation framework to inform reimbursement decisions was developed using LEA.

The co-dependent technology evaluation framework

- The approach uses the hypothetical framework of a double randomized controlled trial as a template for determining what information elements are needed for the evaluation. Information is also elicited on the biological plausibility of the genetic biomarker, as well as whether the biomarker is a prognostic factor or an effect modifier for the accompanying drug therapy. All of this information is linked narratively and then the findings are integrated using the medium of decision analytic modeling.
- The framework includes 79 information requests and these have been incorporated into Government guidance for applicants seeking reimbursement of these companion tests and drugs.

International experience

- Methods of linking evidence to inform test reimbursement decisions is gaining momentum but the largest application to pharmacogenetic interventions is in Australia. Since 2011, nine pharmacogenetic test-drug evaluations have been conducted. Five of these personalized medicines were publicly funded, two were rejected and two have been deferred. To date, the National Institute of Health and Care Excellence (NICE), UK, has evaluated one pharmacogenetic intervention but has a further four underway.

Limitations of LEA

- The main limitation of LEA is finding evidence to support all areas of the linkage and to ensure that there is transferability of populations, genetic tests, biomarker definitions and outcome criteria between each linked piece of evidence, particularly when used in economic modeling.

Future perspective

- Regulatory and reimbursement consideration of the use of genome testing to guide targeted prophylactic and symptomatic treatments raises further complexities – technical, legal, ethical and social.
- The suggestion that the current evidence-based paradigm is too inflexible to address individualized medicine – through genomics - is premature. Complementary assessment methods can be used, including the use of LEA to inform genomic prediction models and then the validation of genomic prediction models to guide therapies using standard empirical methods.

REFERENCES

1. Wolcott J, Schwartz A, Goodman C: *Laboratory Medicine: A National Status Report.* Division of Laboratory Systems, Centers for Disease Control and Prevention (CDC) , 385 (2008). Available at: https://www.futurelabmedicine.org/our_findings/

2. Suthers G: *Report of the Australian Genetic Testing Survey 2006.* Royal College of Pathologists of Australasia, Adelaide, Australia, 117 (2008).

3. Garrison LP, Austin MJF: The Economics of Personalized Medicine: A Model of Incentives for Value Creation and Capture*. Drug Inf J* 41, 501-509 (2007).

4. National Institute for Health and Clinical Excellence: *Interim Methods Statement (pilot). Version 8 Final*. National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme, Manchester, UK, 46 (2010). Available at:

   http://www.nice.org.uk/media/09E/D5/DiagnosticsAssessmentProgrammeInterimMethodsStatement.pdf

5. National Institute for Health and Care Excellence: *Diagnostics Assessment Programme Manual*. National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme, Manchester, UK, 130 (2011).

6. European Network for Health Technology Assessment (EUnetHTA): *HTA Core Model for Diagnostic Technologies. Work Package 4.* FinOHTA, Finnish Office for HTA, Finland, 176 (2008).

7. Agency for Healthcare Research and Quality (AHRQ): *Methods Guide for Medical Test Reviews*. AHRQ Publication No. 12-EC017. Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services Rockville, MD, 188 (2012).

8. Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics: *ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing* (2010). Available at:

   http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm

*** The first comprehensive model for undertaking a review of genetic tests in terms of Analytic validity, Clinical validity, Clinical utility, and associated Ethical, legal and social implications.**

9. Food and Drug Administration (FDA): *In Vitro Companion Diagnostic Devices: Draft Guidance for Industry and Food and Drug Administration Staff*. Food and Drug Administration (FDA), U.S. Department of Health and Human Services, Rockville, MD, 12 (2011).

10. Medical Services Advisory Committee (MSAC): *Guidelines for the Assessment of Diagnostic Technologies.* Commonwealth of Australia, Canberra, ACT, 1-93 (2005).

*** The first description of using a linked evidence approach to evaluate medical tests to inform policy decisions concerning test reimbursement.**

11. Staub L, Dyer S, Lord S, Simes Rj: Linking the Evidence: Intermediate Outcomes in Medical Test Assessments*. Int J Technol Assess Health Care* 28(1), 52-58 (2012).

12. Evaluation of Genomic Applications in Practice and Prevention Working Group: The EGAPP initiative: lessons learned*. Genet Med*, advance online publication 8 August 2013, doi:10.1038/gim.2013.110 (2013).

*** An updated and useful discussion of the processes and methods used by the EGAPP Working Group in the evaluation of genetic tests.**

13. Di Ruffano L, Davenport C, Eising A, Hyde C, Deeks J: A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare*. J Clin Epidemiol* 65(3), 282-287 (2012).

14. Lord SJ, Irwig L, Bossuyt PM: Using the principles of randomized controlled trial design to guide test evaluation*. Med Decis Making* 29(5), E1-E12 (2009).

**** A key paper on the concepts behind use of the linked evidence approach in the evaluation medical tests.**

15. Fryback DG, Thornbury JR: The efficacy of diagnostic imaging*. Med Decis Making* 11, 88-94 (1991).

16. Harris R, Helfand M, Woolf S *et al.*: Current methods of the US Preventive Services Task Force: a review of the process*. Am J Prev Med* 20(3 Suppl), 21-35 (2001).

17. Nelson H, Huffman L, Fu R, Harris E: Genetic Risk Assessment and BRCA Mutation Testing for Breast and Ovarian Cancer Susceptibility: Systematic Evidence Review for the U.S. Preventive Services Task Force. *Ann Intern Med* 143, 362-379 (2005).

18. Nelson H, Pappas M, Zakher B, Mitchell J, Okinaka-Hu L, Fu R: Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer in Women: A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation. *Ann Intern Med* 160, 255-266 (2014).

19. Whitlock E, Garlitz B, Harris E, Bell T, Smith P: Screening for Hereditary Hemochromatosis: A Systematic Review for the U.S. Preventive Services Task Force. *Ann Intern Med* 145, 209-223 (2006).

20. Merlin T, Lehman S, Hiller JE, Ryan P: The "linked evidence approach" to assess medical tests: a critical analysis. *Int J Technol Assess Health Care* 29(3), 343-350 (2013).

**** A systematic review of the use of the linked evidence approach in Australia to evaluate medical tests for policy purposes.**

21. Buckley L, Wang S, Merlin T: *Molecular testing for myeloproliferative disease. Part A – Polycythaemia vera, essential thrombocythaemia and primary myelofibrosis. Part B - Systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia.* MSAC application 1125 Assessment report, Commonwealth of Australia, Canberra, ACT, 273 (2009).

22. Gillespie J, Guarnieri C, Phillips H, Bhatti T: *Urinary metabolic profiling for detection of metabolic disorders*. MSAC application 1114 Assessment report, Commonwealth of Australia, Canberra, ACT, 168 (2009).

23. Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F: The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. *J Clin Epidemiol* (11), 1116-1122 (2007).

24. Haddow J, Palomaki G: *Population based prenatal screening for cystic fibrosis via carrier testing: ACCE review*. Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, 252 (2002). Available at: http://www.cdc.gov/genomics/gtesting/ACCE/ACCE.htm

25. Rowley P, Haddow JE, Palomaki GE: *DNA testing strategies aimed at reducing morbidity and mortality from hereditary non-polyposis colorectal cancer (HNPCC): An ACCE mini-review*. Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia, 42 (2003). Available at: http://www.cdc.gov/genomics/gtesting/ACCE/ACCE.htm

26. McClain M, Palomaki G, Piper M, Haddow J: A rapid-ACCE review of CYP2C9 and VKORC1 alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding. *Genet Med* 10(2), 89-98 (2008).

27. Teutsch SM, Bradley LA, Palomaki GE et al.: The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group. *Genet Med* 11(1), 3-14 (2009).

*** The original guidance on the processes and methods used by the EGAPP Working Group in the evaluation of genetic tests.**

28. Merlin T, Weston A, Tooher R: Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Med Res Methodol* 9, 34 (2009).

29. Bonis P, Trikalinos T, Chung M et al.: *Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications*. Evidence Report/Technology Assessment No. 150 (Prepared by Tufts-New England Medical Center Evidence-based Practice Center under Contract No. 290-02-0022). AHRQ Publication No. 07-E008, Agency for Healthcare Research and Quality, Rockville, MD, 781 (2007).

30. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group: Recommendations from the EGAPP Working Group: genetic testing strategies in newly diagnosed individuals with colorectal cancer aimed at reducing morbidity and mortality from Lynch syndrome in relatives. *Genet Med* 11(1), 35-41 (2009).

31. Mina K: *Report of the RCPA Genetic Testing Survey* 2011. Royal College of Pathologists of Australasia, 35 (2012). Available at:

http://www.rcpa.edu.au/Library/Practising-Pathology/RCPA-Genetic-Testing/Docs/RCPA-Genetic-Testing-Survey-Report

32. Terasawa T, Dahabreh I, Castaldi P, Trikalinos T: *Systematic Reviews on Selected Pharmacogenetic Tests for Cancer Treatment: CYP2D6 for Tamoxifen in Breast*

Cancer, KRAS for anti-EGFR antibodies in Colorectal Cancer, and BCR-ABL1 for Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia. AHRQ Technology Assessment Report, Agency for Healthcare Research and Quality, Rockville, MD, 184 (2010).

33. Laksman Z, Detsky As: Personalized medicine: understanding probabilities and managing expectations*. J Gen Intern Med* 26(2), 204-206 (2011).

34. Holmes MV, Shah T, Vickery C, Smeeth L, Hingorani AD, Casas JP: Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. *PLoS One* 4(12), e7960 (2009).

35. Ioannidis J, Trikalinos T, Ntzani E, Contopoulos-Ioannidis D: Genetic associations in large versus small studies: an empirical assessment. *Lancet* (361), 567-571 (2003).

36. Colhoun H, Mckeigue P, Davey Smith G: Problems of reporting genetic associations with complex outcomes. *Lancet* 361, 865-872 (2003).

37. Van Cutsem E, Kohne CH, Hitre E et al.: Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer. *N Engl J Med* 360(14), 1408-1417 (2009).

38. Douillard JY, Oliner KS, Siena S et al.: Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer. *N Engl J Med* 369(11), 1023-1034 (2013).

39. Regan MM, Leyland-Jones B, Bouzyk M et al.: CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the breast international group 1-98 trial. *J Natl Cancer Inst* 104(6), 441-451 (2012).

40. Meckley LM, Neumann PJ: Personalized medicine: factors influencing reimbursement. *Health Policy* 94(2), 91-100 (2010).

*** A key study of the factors that predict and affect reimbursement of personalized medicines.**

41. Faulkner E, Annemans L, Garrison L et al.: Challenges in the development and reimbursement of personalized medicine-payer and manufacturer perspectives and implications for health economics and outcomes research: a report of the ISPOR personalized medicine special interest group. *Value Health* 15(8), 1162-1171 (2012).

42. Merlin T, Farah C, Schubert C, Mitchell A, Hiller JE, Ryan P: Assessing personalized medicines in Australia: a national framework for reviewing codependent technologies. *Med Decis Making* 33(3), 333-342 (2013).

**** Detail on the framework developed - using a linked evidence approach - to evaluate companion diagnostics and drug therapies for government reimbursement decisions.**

43. Bradford Hill A: The environment and disease: association or causation? *Proc R Soc Med* 58, 295-300 (1965).

44. Craig D, Mcdaid C, Fonseca T, Stock C, Duffy S, Woolacott N: Are adverse effects incorporated in economic models? A survey of current practice. *Int J Technol Assess Health Care* 26(03), 323-329 (2010).

45. Evaluation of Genomic Applications in Practice and Prevention (Egapp) Working Group: Recommendations from the EGAPP Working Group: testing for cytochrome P450 polymorphisms in adults with nonpsychotic depression treated with selective serotonin reuptake inhibitors. *Genet Med* 9(12), 819-825 (2007).

46. Evaluation of Genomic Applications in Practice and Prevention (Egapp) Working Group: Recommendations from the EGAPP Working Group: can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan? *Genet Med* 11(1), 15-20 (2009).

47. Evaluation of Genomic Applications in Practice and Prevention (Egapp) Working Group: Recommendations from the EGAPP Working Group: routine testing for Factor V Leiden (R506Q) and prothrombin (20210G>A) mutations in adults with a history of idiopathic venous thromboembolism and their adult family members. *Genet Med* 13(1), 67-76 (2011).

48. Evaluation of Genomic Applications in Practice and Prevention (Egapp) Working Group: Recommendations from the EGAPP Working Group: can testing of tumor tissue for mutations in EGFR pathway downstream effector genes in patients with metastatic colorectal cancer improve health outcomes by guiding decisions regarding anti-EGFR therapy? *Genet Med* 15(7), 517-527 (2013).

49. Payne K, Annemans L: Reflections on market access for personalized medicine: recommendations for Europe. *Value Health* 16(6 Suppl), S32-38 (2013).

50. Fleeman N, Martin Saborido C, Payne K et al.: The clinical effectiveness and cost-effectiveness of genotyping for CYP2D6 for the management of women with breast cancer treated with tamoxifen: a systematic review. *Health Technol Assess* 15(33), 1-102 (2011).

51. National Institute for Health and Care Excellence: *EGFR-TK mutation testing in adults with locally advanced or metastatic non-small-cell lung cancer*. NICE diagnostics guidance 9, National Institute for Health and Care Excellence, Manchester, 46 (2013). Available at: http://guidance.nice.org.uk/DG9/Guidance/pdf/English

52. Pharmaceutical Benefits Advisory Committee: *ADDENDUM – July 2013. Product: Dabrafenib, capsules, 50 mg and 75 mg Tafinlar®*. Commonwealth of Australia, Canberra, ACT, 3 (2013).

53. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Gefitinib, tablet, 250mg, Iressa®*. Commonwealth of Australia, Canberra, ACT, 13 (2013).

54. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Erlotinib, tablets, 25 mg, 100 mg, 150 mg (as hydrochloride), Tarceva®*. Commonwealth of Australia, Canberra, ACT, 21 (2013).

55. Hailey D: The history of health technology assessment in Australia. *Int J Technol Assess Health Care* 25 Suppl 1, 61-67 (2009).

56. Ginsburg GS, Kuderer NM: Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond. *J Clin Oncol* 30(34), 4233-4242 (2012).

57. Abecasis GR, Auton A, Brooks LD et al.: An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422), 56-65 (2012).

58. Becla L, Lunshof JE, Gurwitz D et al.: Health technology assessment in the era of personalized health care. *Int J Technol Assess Health Care* 27(2), 118–126 (2011).

59. Medical Services Advisory Committee: *Public Summary Document, Application No. 1207 – Testing for V600 status in patients with locally advanced or metastatic melanoma for eligibility for dabrafenib treatment*. Commonwealth of Australia, Canberra, ACT, 8 (2013).

60. Medical Services Advisory Committee: *Public Summary Document, Application No. 1174. Assessment of viral tropism testing of HIV to inform treatment with maraviroc*. Commonwealth of Australia, Canberra, ACT, 15 (2012).

61. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Maraviroc, tablets, 150 mg and 300 mg, Celsentri®*. Commonwealth of Australia, Canberra, ACT, 10 (2012).

62. Medical Services Advisory Committee: *Public Summary Document, Application No. 1163 - Assessment of HER2 gene amplification for use of trastuzumab in gastric cancer*. Commonwealth of Australia, Canberra, ACT, 17 (2012).

63. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Trastuzumab, powder for I.V. infusion, 60 mg and 150 mg, Herceptin®*. Commonwealth of Australia, Canberra, ACT, 7 (2012).

64. Medical Services Advisory Committee: *Public Summary Document, Application 1161 - Gefitinib first line testing for mutations of epidermal growth factor receptor (EGFR) in patients with locally advanced or metastatic non-small cell lung cancer (NSCLC*). Commonwealth of Australia, Canberra, ACT, 14 (2013).

65. Medical Services Advisory Committee: *Public Summary Document, Application No. 1172 – BRAF genetic testing in patients with melanoma for access to proposed PBS-funded vemurafenib*. Commonwealth of Australia, Canberra, ACT, 8 (2013).

66. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Vemurafenib; tablet, 240 mg, Zelboraf®*. Commonwealth of Australia, Canberra, ACT, 7 (2013).

67. Medical Services Advisory Committee: *Public Summary Document, Application 1173 - Testing for epidermal growth factor receptor (EGFR) status in patients with locally advanced (stage IIIB) or metastatic (stage IV) non-small cell lung cancer (NSCLC) for access to erlotinib*. Commonwealth of Australia, Canberra, ACT, 15 (2013).

68. Medical Services Advisory Committee: *Public Summary Document, Application No. 1230 – HER2 ISH testing for access to trastuzumab for neoadjuvant breast cancer*. Commonwealth of Australia, Canberra, ACT, 4 (2012a).

69. Pharmaceutical Benefits Advisory Committee: *Public Summary Document, Product: Trastuzumab, powder for I.V. infusion, 60 mg and 150 mg, Herceptin®*. Commonwealth of Australia, Canberra, ACT, 11 (2012a).

70. Pharmaceutical Benefits Advisory Committee: *November 2013 PBAC Meeting Outcomes - Deferrals*. Australian Government Department of Health, Canberra, ACT, 4 (2013). Available at: http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/pbac-outcomes/2013-11

71. Pharmaceutical Benefits Advisory Committee: *November 2013 PBAC Meeting Outcomes - Positive Recommendations*. Australian Government Department of Health, Canberra, ACT, 41 (2013). Available at:

http://www.pbs.gov.au/info/industry/listing/elements/pbac-meetings/pbac-outcomes/2013-11

Relevance of Paper 5 to the thesis

Like Paper 4**, Paper 5** has addressed the final research question posed for the thesis, that is:

Can the linked evidence approach (LEA) be feasibly adapted to the evaluation of personalised medicines ie the use of a genetic test to target a pharmaceutical treatment?

Paper 5 expands on information provided in papers 3 and 4 by describing the evolution of the LEA method and discussing some of the practicalities in its application. This paper discusses briefly how LEA has been used internationally for the development of clinical practice guidelines and more recently by NICE for their diagnostic HTA program. The paper highlights how important it is to ensure that there is reasonable transferability or clinical applicability of populations between linkages when discussing the clinical utility of the test. This is particularly important when attempting to quantify the clinical and cost impact of the test in the economic models used in the HTA process; not least because the policy-makers will distrust the results of a model if there is no face validity in having similar populations, tests and healthcare contexts of the studies providing inputs into the model.

Paper 5 describes how methods of linking evidence to inform test reimbursement decisions are gaining momentum but that Australia is currently making the greatest use of these methods to evaluate pharmacogenetic interventions. Up until the paper's publication, nine pharmacogenetic test-drug evaluations had been conducted. Five of these personalised medicines were publicly funded, two were rejected and two were deferred. Similarly, the National Institute of Health and Care Excellence (NICE), in the UK, had evaluated one pharmacogenetic intervention using the method – with some learning curve issues - but had a further four underway. The LEA approach has been adapted to the evaluation of personalised medicines – its use has had significant policy impact in Australia and the application of the method is ongoing.

Given the highly technical nature of the methodology, particularly to people unfamiliar with diagnostic tests, genetics and/or decision analytic modelling, a full-day workshop was held in Washington DC at the 2014 Health Technology Assessment international (HTAi) conference to guide participants through the logic and application of the process. The response to the workshop was encouraging. There were requests – including from the Australian Government – to have the workshop repeated. A higher level discussion of the method and how it can be applied internationally (particularly given the constraints of individual health systems) is scheduled for the HTAi conference in Oslo, Norway, in June 2015 – with myself and a representative from CADTH in Canada, the Ludwig Boltzmann Institute in Austria, and the European Diagnostic Manufacturers' Association presenting a Panel on the topic. During the same month, I have been invited by the European Diagnostic Manufacturers' Association to present at a workshop on companion diagnostics in Brussels, Belgium.

CHAPTER 8

CONCLUSION

The evaluation of medical tests has been an under-developed methodological area in health technology assessment. This PhD thesis has attempted to redress this lack by providing several methodological approaches to support the assessment of these technologies so that policy makers have the requisite information to make considered public funding decisions.

Tests are not benign technologies, although that has often been the view (Roberts 2006). There are ethical issues associated with testing when subsequent treatment options are futile. As many tests are non-invasive or minimally invasive (and thus considered ‘safe’), in the past the health community had given little consideration to the downstream consequences of diagnostic testing and investigative procedures. These include, if the test results are inaccurate, false reassurance that the patient does not have the condition tested for (resulting in delays in effective treatment) or false positive diagnoses that result in more invasive testing and/or potentially harmful treatment that is not needed.

When I began this PhD the Cochrane Collaboration had barely started its methodological work on systematic reviews of diagnostic test accuracy and the concept of linking evidence to determine test effectiveness had just emerged in Australia. Reviews of test accuracy are now common and LEA is recommended for determining test effectiveness in HTA in Europe, the UK and the USA. Experience with the approach, however, is still largest in Australia so it is in a good position to provide guidance on LEA use and adaptation to different testing circumstances. The publications in this thesis have kept pace with the rapid methodological developments occurring in medical test evaluation over the last eight years and have gone some way towards informing progress in this area. Although some of the developments associated with Paper 1 have been superseded by the GRADE methodology, the approach – as it relates to the study design-related bias of test accuracy studies – is still sound.

Research Answers

This PhD thesis attempted to answer the following questions:

1. **Is the use of the linked evidence approach feasible when assessing medical tests for public funding decisions?**

The first linkage in LEA relates to evidence of test accuracy. To determine the internal validity of test accuracy studies these studies need to be appraised in terms of the likelihood of bias as a consequence of the study design and then judged as to whether bias has been introduced through the way the study has been executed. The hierarchy of study designs originally used in Australia was primarily relevant when direct evidence of test effectiveness was available (NHMRC (1999)). However, as discussed previously, direct evidence of *diagnostic* effectiveness (ie the impact of a test on health outcomes) is rare, thus the existing hierarchy could not be used – and was being inappropriately used for diagnostic accuracy studies. As part of the research for this thesis a hierarchy of evidence was created that enabled diagnostic accuracy studies to be appropriately ranked in terms of the likelihood that bias has been introduced by the design of the study (Merlin, T., Weston & Tooher 2009). This allowed the first linkage in LEA to be addressed properly and facilitated use of the methodology. A glossary of the terms used in the hierarchy was also produced. The hierarchy is now standard when evaluating test accuracy studies in Australian HTA (as well as in Australian clinical practice guideline development).

The systematic review of Australian HTAs conducted for this thesis was clear in demonstrating that LEA is feasible (Merlin, T., Lehman, et al. 2013). The method was used in the vast majority (96%) of medical test HTAs produced on behalf of the Medical Services Advisory Committee following the introduction of guidance on the approach in 2005 (MSAC 2005a).

2. **What effect (if any) has the linked evidence approach methodology had on Australian policy makers' decisions to publicly fund medical tests?**

   I had hypothesised that the mandated use of LEA in Australia would change the way that medical tests were evaluated for their clinical utility. The results of our study confirmed this. LEA methodology was the most common method for presenting data to policy makers between 2005 and 2014. Decision-making was, therefore, based on the linkage of systematically reviewed evidence on test performance, relative to an accepted reference standard, to evidence on the impact of the test on treatment decisions and through careful consideration of the likely impact of the test on patient health outcomes – including the impact of false positive and false negative test results. At the least, it is likely that this could have led to more informed decision-making.

The results of prediction models indicated that the choice of methodological approach is unlikely to affect the direction (positive or negative) of funding decisions but that the use of the linked evidence approach is strongly *negatively* associated with a medical test obtaining interim funding. The additional information provided with LEA may reduce the uncertainty associated with decision-making and therefore reduce the need to make interim funding decisions. There was a concomitant increase in more definitive positive or negative public funding recommendations perhaps because of the increased quantity and coherence (prediction of clinical utility) of the information provided to decision-makers.

Whether use of the method in HTA has also resulted in *better* decision-making with regards to the funding of medical tests is a matter for speculation.

3. Are there any specific situations where the use of the linked evidence approach is inadequate? If so, are there ways that the approach can be improved?

The systematic review of Australian HTAs indicated that the majority of assessments undertaken on behalf of the Medical Services Advisory Committee since 2005 used the LEA method but one quarter of these reported problems using the approach. None of the HTAs reporting on direct evidence alone reported difficulties in conducting the evaluation (Merlin, T., Lehman, et al. 2013).

My research confirmed the methodological inadequacies of LEA that were anticipated in the *MSAC Diagnostic Guidelines*, namely situations when the tested and treated populations differed in terms of spectrum of disease (eg if the test identified cases earlier in their disease course than would normally be the case) and when the reference standard for determining test performance was imperfect (MSAC 2005a).

There were, however, situations identified where inadequacies with LEA were unanticipated. This included circumstances when a triage test was being used to rule out disease, in instances when only surrogate outcomes for measuring health impact were available—resulting in a need for additional prognostic data in the linkage—and when the test being evaluated was already an established technology—leading to difficulty in determining the correct comparator.

The identification of these inadequacies with LEA led to the formulation of a decision framework that used specific strategies to circumvent test evaluations that may be problematic. Three scenarios were constructed that reflected all of the probable results of the first evidence linkage (test accuracy studies and the assessment of safety/invasiveness) and indicated when subsequent evidence linkages could be used to address evidence gaps and thus reduce the predicted decision-maker uncertainty. The framework is Bayesian in that the evaluation technique and evidence linkage approach to be used is informed by findings from evidence that is accumulated earlier in the linkage.

The development of this decision framework for applying LEA means that:

- there is a standardised methodology that can be used across different test evaluations, maintaining consistency in evidence selection (less risk of bias) and transparency of approach;
- a parsimonious evidence collation approach can be justified, such that where the policy decision will be a foregone conclusion (ie the new test has no additional benefit over comparator test in terms of accuracy and safety/invasiveness) there is no need to undertake unnecessary additional evidence collation. The evidence-based advice is tailored to the circumstances (ie the likely uncertainties) so that information that may obfuscate the main message and confuse the policy maker is not presented. With prior agreement from policy makers with regard to the approach to be taken, the policy objective can be met and in less time and at less cost; and
- there is a systematised approach to finding relevant inputs for decision-modelling and economic evaluations. Decision modelling is needed to determine the likely cost-effectiveness of the test in the local health system (Briggs, A, Claxton & Sculpher, 2006). The LEA method basically reflects the use of the test in the test-treatment pathway and this test-treatment pathway is usually reflected in the decision analytic upon which an economic model is structured. The systematically and comprehensively acquired evidence that is obtained for LEA can be used to derive unbiased (and factual, as opposed to hypothetical) transition probabilities or estimates—and ranges of estimates—to use as inputs at chance nodes in an economic model assessing the cost-effectiveness of the new test (Craig et al. 2010; Caro, JJ, Briggs et al. 2012). These inputs are derived from the collated evidence on

test accuracy, changes in management and the health impacts of treating on the basis of accurate test results, as well as inaccurate test results (ie consequences of false positive and false negative results).

4. **Can the linked evidence approach be feasibly adapted to the evaluation of personalised medicines?**

As LEA is a 'proxy' for the information obtained from a direct trial or study of test effectiveness—the impact of testing on patient health outcomes—the design of the hypothetical direct study is used to guide what evidence could or should be collated. The hypothetical best design would be a randomised controlled trial, such that eligible patients with suspected disease would be allocated to the new test or the likely comparator (eg existing test or no test) and their health outcomes (subsequent to treatment) would be monitored (Lord, SJ, Irwig & Bossuyt 2009).

When applying LEA to a genetic test paired with a drug, the hypothetical direct evidence that is needed is a double randomised trial which is even more unlikely to be available than the single randomised controlled trial mentioned above. Patients would be allocated to use of the new test or an alternative/no test and then patients in each of those test conditions, no matter whether positive for the genetic biomarker or not, would be randomised to subsequent use of the targeted new drug or usual care. Such a design would capture all of the possible consequences of using the companion test and drug or their available alternatives but would also be likely to be considered unethical, expensive to mount and resource intensive.

The use of an extended LEA enables the lack of directly available evidence for these personalised medicines to be addressed. The LEA method used for determining test effectiveness (as described above) is still used but is also linked to evidence of alternative treatment effectiveness according to biomarker status. The hypothetical design (a double randomised controlled trial), if used as a template for the evidence linkage, assists with determining what information is missing from the collated evidence and, thus, where the potential policy uncertainties might lie (Merlin, T, Farah, et al. 2013).

The presented research demonstrates that LEA can be feasibly adapted to the evaluation of personalised medicines consisting of a companion diagnostic test and a pharmaceutical. The method requires many evidence linkages to explore the relationship between biomarker and treatment. This is needed as, if there is actually no relationship (or interaction) between the biomarker and the drug, there needs to be recognition of the potential for separate policy decisions for subsidising or not subsidising the test and/or the drug.

This research has met the stated objectives of the thesis which were to:

- Determine whether the use of LEA in the evaluation of medical tests is appropriate for informing health policy;
- Assess the strengths and weaknesses of the methodology;
- Analyse whether the method can be applied to different tests and testing situations;
- Provide recommendations on how the methodology could be improved; and
- Provide guidance and tools to assist researchers and evaluators with the application of LEA and its adaptation to the HTA of pharmacogenetic tests.

Challenges

There were several problems encountered while conducting the research for this thesis. Most of these related to time. This PhD was undertaken half-time, to fit around my work as an HTA practitioner, and so what would normally take four years full-time has taken eight years part-time. During these eight years there have been some major developments in HTA in Australia.

The most significant development occurred in 2009 when a review of health technology assessment in Australia was commissioned by the Australian Government (Australian Government Department of Health and Ageing 2009a). This review was a response to the recommendation from the Banks review on "Rethinking Regulation – Report of the Taskforce on Reducing Regulatory Burdens on Business, January 2006" that:

"The Australian Government should undertake a system-wide, independent and public review of health technology assessment, with the objective of reducing fragmentation, duplication and unnecessary complexity, which can delay the introduction of beneficial new medical technologies. Health technology assessment processes and decisions should also be made more transparent, in line with good regulatory practice."

The terms of reference of the HTA Review were to report on (Australian Government Department of Health and Ageing 2009):

1. Simplification and better co-ordination between the Commonwealth HTA processes (as identified in the Review scope), which includes:
    a. consideration of a single entry point and tracking system for applications for market registration and funding;
    b. making time to affordable access as short as possible for new technologies while maintaining or improving the rigour of evaluation processes; and
    c. examination of the feasibility of conducting concurrent assessments for market registration and funding and coverage purposes, noting current work in this area.
2. Improving role clarity and addressing duplication between processes, where it exists, including consideration of consolidating functions with the Australian HTA system.
3. Reviewing post marketing surveillance mechanisms to ensure the ongoing safety, and efficacy of medical devices.
4. Strengthening transparency and procedural fairness in the assessment, decision making and fee negotiation arrangements for processes (as outlined in the Review scope) through improved communication with stakeholders about process, methodologies, outcomes and performance against key indicators.
5. Enhanced arrangements for assessment of co-dependent and hybrid technologies.

One of the proposals suggested by the review was for the Australian Government to provide 'improved guidance on methodologies and methodological processes'. The entire HTA process underpinning evaluations of technologies for Medicare funding therefore underwent a transformation (Australian Government Department of Health and Ageing, 2009b). A risk-based ("fit-for-purpose") evaluation system was implemented, meaning that the evaluation process and the evidence requirements were tailored to the level of health/financial/societal risk associated with introducing the technology.

In addition, a new step was included in the evaluation process. This step allowed the likely place of the technology in the health system (ie population/s using the technology, acceptable comparator/s and outcomes) to be determined, with stakeholder and public input, prior to the formal evaluation taking place. A further change then allowed the agency involved in submitting the technology for a public funding decision to have the option of providing an application—involving a systematic literature review and economic evaluation—to be critiqued by an independent evaluator contracted by the Government. Those agencies that did not wish to exercise this option were still able to follow the original evaluation process which involved an independent contracted HTA of the technology. The application and critique process was introduced in order to expedite the time taken between submission and a public funding decision.

The change in evaluation process affected the inclusion criteria for the systematic reviews of MSAC HTAs included in this thesis. The primary reason for this was the need to maintain consistency in process both before and after the introduction of the *MSAC Diagnostic Guidelines* in order to minimise confounding. There was a strong likelihood that applicants for public funding would not—at least initially—be experts in HTA, would not have a methodological background or familiarity with MSAC processes and Guidelines—as would be the case with the independent academic evaluators that traditionally did the contracted HTAs for Government. It was decided to restrict the included HTAs to those contracted by the Government.

The decision to develop a decision framework for applying the linked evidence approach (Merlin, T., Lehman, et al. 2013) was, in part, a response to the HTA Review proposal indicating a need for methodological guidance. With new agencies submitting applications for public funding, often with little familiarity with evaluation processes let alone the more complex evaluations like diagnostic tests, it was thought that explicit guidance on evaluating diagnostic tests would assist both applicants and independent evaluators in providing a common evaluation approach and understanding of the evidence requirements. This would also facilitate consistent decision-making by MSAC.

One of the other outcomes from the HTA Review was to create a system for evaluating co-dependent technologies. As the Australian Government had siloed processes for assessing drugs and services involving tests/devices/procedures, it was recognised that inconsistent decision-making was occurring between the Committees responsible for one of the technologies in a co-dependent pairing. In addition there was no guidance available at the time, internationally, on how to evaluate co-dependent technologies. This was the impetus for undertaking the development of the co-dependent technology evidentiary framework (Merlin, T., Farah, et al. 2013) for this thesis. However, the concurrent structural changes to the system for processing co-dependent technologies in Australia did impact on the timely development of the methodological framework as the requirements for stakeholder input changed over the course of the project, ie from different parts of government, decision-makers and the public (industry).

Translation of the Research

A new methodology is only really relevant if it is used. The methodologies presented in this thesis have all been translated into policy and practice in Australia.

The research included in three of the five papers contributing to this thesis has translated into national health policy and technology evaluation practice. Two of the five papers in this thesis also provide information on whether and how these methodologies have informed health policy in Australia.

The evidence hierarchy produced by Merlin, T., Weston & Tooher (2009) was recommended for inclusion in part of a system, known as FORM, that was mandated by the National Health and Medical Research Council—the peak clinical practice guideline development agency in Australia—to be used by clinicians and methodologists to develop and grade recommendations in national evidence-based clinical practice guidelines. In addition, the evidence hierarchy was used by health technology assessors contracted by the Australian Government to evaluate medical services for public funding decisions.

The translational impact for the paper by Merlin, T., Lehman, et al. (2013) has not yet been fully established, given it was only published in 2013. However, the decision framework that was formulated for applying the linked evidence approach in the evaluation of medical tests is being used in both protocols and evaluations of health technologies in Australia for public funding decisions by MSAC. It has also informed the development of new guidance for the assessment of investigational services to be released by the Australian Government later this year.

The final publication in this thesis presented the first process available, internationally, for evaluating co-dependent technologies, specifically personalised medicines (pharmacogenetics). This co-dependent technology evidentiary framework was pilot-tested, revised, and is now in use by the Australian agencies responsible for the public funding of tests and drugs nationally – the MSAC and the PBAC, respectively. Industry submissions of co-dependent technologies, requesting funding on the Medicare Benefits Schedule (for the test) and the Pharmaceutical Benefits Schedule (for the drug) provide information on each of the items requested by the framework. The method is now recommended in MSAC's *Technical Guidelines for the Assessment of Investigative Technologies* (Medical Services Advisory Committee 2015, In Press), the *Guidelines for Submissions to the Pharmaceutical Benefits Advisory Committee* (Pharmaceutical Benefits Advisory Committee 2013), and is followed by both those who submit applications for co-dependent technologies for public funding in Australia and independent evaluators of these submissions. In addition, the approach has informed the methods undertaken to evaluate personalised medicines in Austria (Kisser, 2014), and there is interest in adapting the methodology to other reimbursement systems internationally.

Future Directions

The original methods and tools created as part of this thesis may become the basis for future research enquiry. One of the spin-off projects that I am currently undertaking is a web-based survey of health technology assessors, industry applicants and policy makers to determine whether the co-dependent technology evaluation framework is understandable, facilitates decision-making processes and is able to be populated with evidence. The survey requests input on each individual section of the framework in order to determine whether slight modification and/or training in use of the method would be beneficial.

Avenues of additional investigation might include:

- Assessing whether the linked evidence approach can be used as part of other commonly used evidence grading systems such as GRADE (The GRADE* working group 2005) to assess the clinical utility of a medical test
- Determining whether the decision framework for applying LEA to medical tests is acceptable to policy-makers in Australia and internationally ie given the potential abbreviated presentation of evidence
- Investigating the experience of Australian health technology assessors, and the newer industry applicants, with the decision framework for applying LEA to medical tests and gauging the utility of this method in other health systems
- Determining whether the co-dependent technology evidentiary framework can be adapted to other regulatory and reimbursement systems worldwide

- Applying and/or modifying the co-dependent technology evidentiary framework to co-dependent technologies that do not involve a personalised medicine eg a device and a drug (such as drug-eluting stents), as well as to scenarios that involve multiple biomarkers (in a test-drug pairing) or to scenarios where the biomarker has been accepted but the drug has not (or vice versa) or to comparisons between two or more co-dependent technologies
- Translating the decision framework for medical tests and the co-dependent technology evaluation framework into primary research, such that the evidence quality and type of evidence required by these frameworks can be reasonably produced. Several models for facilitating this type of collaboration between different stakeholders, to develop new research on genetic or genomic tests, have been developed (Deverka, PA, Lavallee et al. 2012; Lal, JA, Schulte in den Bäumen et al. 2011) but have yet to be used for co-dependent technologies specifically
- Developing processes to assess pharmacogenomic medicines (truly individualised medicine), that rely on testing the human genome of each individual and tailoring prophylactic and therapeutic options accordingly, for public funding decisions.

Final Thoughts

This thesis provides a thorough evaluation of LEA as a method for assessing medical tests in the context of providing evidence-based advice to policy makers. Tools have been created to ameliorate several identified weaknesses with the methodology, including the development of an evidence hierarchy for diagnostic accuracy studies and the development of a decision framework for applying LEA. The methodology has also been adapted and extended for use in the evaluation of pharmacogenetic technologies for reimbursement decisions – an area where no previous methodology has been developed. The methods developed as part of this thesis are an original contribution to the evaluation of the safety, effectiveness and cost-effectiveness of medical tests. These methods have been adopted as part of national HTA processes and practices in Australia and have wider application to the global HTA community.

BIBLIOGRAPHY

Abecasis, GR, Auton, A, Brooks, LD, DePristo, MA, Durbin, RM, Handsaker, RE, Kang, HM, Marth, GT & McVean, GA 2012, 'An integrated map of genetic variation from 1,092 human genomes', *Nature*, vol. 491, no. 7422, Nov 1, pp. 56-65.

Agency for Healthcare Research and Quality (AHRQ) 2010, *Draft Methods Guide for Medical Test Reviews*, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Rockville, MD.

Agency for Healthcare Research and Quality (AHRQ) 2012, *Comprehensive Overview of Methods and Reporting of Meta-Analyses of Test Accuracy*, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Rockville, MD.

Agency for Healthcare Research and Quality (AHRQ) 2012, *Methods Guide for Medical Test Reviews*, Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services, Rockville, MD.

Akaike, H 1974, 'A new look at the statistical model identification ', *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716-723.

Alter, DA, Stukel, TA & Newman, A 2006, 'Proliferation of cardiac technology in Canada: a challenge to the sustainability of Medicare', *Circulation*, vol. 113, no. 3, Jan 24, pp. 380-387.

Anderson, L, Petticrew, M, Rehfuess, E, Armstrong, R, Ueffing, E, Baker, P, Francis, D & Tugwell, P 2011, 'Using logic models to capture complexity in systematic reviews', *Research Synthesis Methods*, vol. 2, pp. 33-42.

Andradas, E, Blasco, JA, Valentin, B, Lopez-Pedraza, MJ & Gracia, FJ 2008, 'Defining products for a new health technology assessment agency in Madrid, Spain: a survey of decision makers', *International Journal of Technology Assessment in Health Care*, vol. 24, no. 1, Winter, pp. 60-69.

Australian Government Department of Health and Ageing 2008, *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee*, Commonwealth of Australia, Canberra, ACT.

Australian Government Department of Health and Ageing 2009, *Review of Health Technology Assessment in Australia - A discussion paper*, Commonwealth of Australia, Canberra, ACT.

Australian Government Department of Health and Ageing 2010, *Draft information requests for assessing a pair of co-dependent technologies*, Australian Government Department of Health and Ageing, Canberra, ACT.

Australian Government Department of Health and Ageing 2012a, *Medicare Benefits Schedule Book*, Commonwealth of Australia, Canberra. Available at: <http://www.health.gov.au/internet/mbsonline/publishing.nsf/Content/Downloads-201303>.

Australian Government Department of Health and Ageing 2012b, 'An Overview of the New Arrangements for Listing on the Medicare Benefits Schedule '. Available at: <http://www.msac.gov.au/internet/msac/publishing.nsf/Content/B8E1F7C44BE7E25BCA257A7D002477C3/$File/Overview-new-arrangements-MSAC.pdf>.

Becla, L, Lunshof, JE, Gurwitz, D, den Baumen, TS, Westerhoff, HV, Lange, BMH & Brand, A 2011, 'Health technology assessment in the era of personalized health care', *International Journal of Technology Assessment in Health Care*, vol. 27, no. 2, pp. 118–126.

Bellomo, R & Bagshaw, SM 2006, 'Evidence-based medicine: classifying the evidence from clinical trials - the need to consider other dimensions', *Critical Care*, vol. 10, p. 232.

Benson, K & Hartz, AJ 2000, 'A comparison of observational studies and randomized, controlled trials', *New England Journal of Medicine*, vol. 342, no. 25, pp. 1878-1886.

Biesheuvel, CJ, Grobbee, DE & Moons, KG 2006, 'Distraction from randomization in diagnostic research', *Annals of Epidemiology*, vol. 16, no. 7, Jul, pp. 540-544.

Black, N 1996, 'Why we need observational studies to evaluate the effectiveness of health care', *British Medical Journal*, vol. 312, pp. 1215-1218.

Bleyer A & Gilbert Welch H 2012, 'Effect of three decades of screening mammography on breast-cancer incidence ', *New England Journal of Medicine*, vol. 367, pp. 1998-2005.

Bonis, P, Trikalinos, T, Chung, M, Chew, P, Ip, S, DeVine, D & Lau, J 2007, *Hereditary nonpolyposis colorectal cancer: diagnostic strategies and their implications*, Agency for Healthcare Research and Quality, Rockville, MD. Available at: <http://archive.ahrq.gov/downloads/pub/evidence/pdf/hnpcc/hnpcc.pdf>.

Bossuyt, PM, Irwig, L, Craig, J & Glasziou, P 2006, 'Comparative accuracy: assessing new tests against existing diagnostic pathways', *British Medical Journal*, vol. 332, pp. 1089-1092.

Bossuyt, PM, Reitsma, JB, Bruns, DE, Gatsonis, CA, Glasziou, PP, Irwig, LM, Lijmer, JG, Moher, D, Rennie, D, de Vet, HCW & for the STARD group 2003, 'Towards Complete and Accurate Reporting of Studies of Diagnostic Accuracy: The STARD Initiative', *American Journal of Radiology*, vol. 181, pp. 51-56.

Bradford Hill, A 1965, 'The environment and disease: association or causation?', *Proceedings of the Royal Society of Medicine*, vol. 58, pp. 295-300.

Brozek, JL, Akl, EA, Alonso-Coello, P, Lang, D, Jaeschke, R, Williams, JW, Phillips, B, Lelgemann, M, Lethaby, A, Bousquet, J, Guyatt, GH & Schunemann, HJ 2009, 'Grading quality of evidence and strength of recommendations in clinical practice guidelines. Part 1 of 3. An overview of the GRADE approach and grading quality of evidence about interventions', *Allergy*, vol. 64, no. 5, May, pp. 669-677.

Brozek, JL, Akl, EA, Jaeschke, R, Lang, DM, Bossuyt, P, Glasziou, P, Helfand, M, Ueffing, E, Alonso-Coello, P, Meerpohl, J, Phillips, B, Horvath, AR, Bousquet, J, Guyatt, GH & Schunemann, HJ 2009, 'Grading quality of evidence and strength of recommendations in clinical practice guidelines: Part 2 of 3. The GRADE approach to grading quality of evidence about diagnostic tests and strategies', *Allergy*, vol. 64, no. 8, Aug, pp. 1109-1116.

Bruel, A, Cleemput, I, Aertgeerts, B, Ramaekers, D & Buntinx, F 2007, 'The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed', *Journal of Clinical Epidemiology*, no. 11, pp. 1116-1122. Available at: <http://onlinelibrary.wiley.com/o/cochrane/clcmr/articles/CMR-11774/frame.html>.

Buckley, E & Merlin, T 2010, *Molecular testing for the diagnosis of systematic mast cell disease, hypereosinophilic syndromes and chronic eosinophilic leukaemia.*, Australian Government, Canberra, ACT. Available at: <http://www.msac.gov.au/internet/msac/publishing.nsf/Content/MSACCompletedAssessments1120-1140>.

Buckley, L, Wang, S & Merlin, T 2009, *Molecular testing for myeloproliferative disease. Part A – Polycythaemia vera, essential thrombocythaemia and primary myelofibrosis. Part B - Systemic mast cell disease, hypereosinophilic syndrome and chronic eosinophilic leukaemia*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Busse, R, Orvain, J, Velasco, M, Perleth, M, Drummond, M, Gurtner, F, Jorgensen, T, Jovell, A, Malone, J, Ruther, A & Wild, C 2002, 'Best practice in undertaking and reporting health technology assessments. Working Group 4 Report', *International Journal of Technology Assessment in Health Care*, vol. 18, no. 2, pp. 361-422.

Centers for Disease Control and Prevention (CDC) 2010, *ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing*, Centers for Disease Control and Prevention (CDC), Office of Public Health Genomics. Available at: <http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm>.

Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics 2010, *ACCE Model List of 44 Targeted Questions Aimed at a Comprehensive Review of Genetic Testing*. Available at: <http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm>.

Clark, GM, Zborowski, D.M., Culbertson, J.L., Whitehead, M., Savoie, M., Seymour, L., Shepherd, F.A., 2006, 'Clinical utility of epidermal growth factor receptor expression for selecting patients with advanced non-small cell lung cancer for treatment with erlotinib', *Journal of Thoracic Oncology*, vol. 1, pp. 837-846.

Clarke, M & Oxman, AD 2003, *Cochrane reviewers' handbook 4.2.1*, The Cochrane Collaboration, Oxford, viewed 8/03/04 2004. Available at: <http://www.cochrane.org/cochrane/hbook.htm>.

Coleman, K, Standfield, L & Weston, A 2004, *The utilisation of established frameworks in assessing and applying non-intervention/non-randomised evidence [Internal report]*, Health Advisory Committee, National Health and Medical Research Council (NHMRC), Canberra, ACT.

Colhoun, H, McKeigue, P & Davey Smith, G 2003, ' Problems of reporting genetic associations with complex outcomes.', *Lancet*, vol. 361, pp. 865-872.

Concato, J, Shah, N & Horwitz, RI 2000, 'Randomized, controlled trials, observational studies, and the hierarchy of research designs', *New England Journal of Medicine*, vol. 342, no. 25, pp. 1887-1892.

Conti, R, Veenstra, DL, Armstrong, K, Lesko, LJ & Grosse, SD 2010, 'Personalized Medicine and Genomics: Challenges and Opportunities in Assessing Effectiveness, Cost-Effectiveness, and Future Research Priorities', *Medical Decision Making*, vol. 30, no. 3, Jan 4, pp. 328-340.

Cooper, H & Hedges, LV (eds) 1994, *The handbook of research synthesis*, Russell Sage Foundation, New York.

Craig, D, McDaid, C, Fonseca, T, Stock, C, Duffy, S & Woolacott, N 2010, 'Are adverse effects incorporated in economic models? A survey of current practice', *International Journal of Technology Assessment in Health care*, vol. 26, no. 03, pp. 323-329.

Daly, J, Willis, K, Small, R, Green, J, Welch, N, Kealy, M & Hughes, E 2007, 'A hierarchy of evidence for assessing qualitative health research', *Journal of Clinical Epidemiology*, vol. 60, pp. 43-49.

Deeks, JJ 2001a, 'Systematic review of evaluations of diagnostic and screening tests', in M Egger, GD Smith & DG Altman (eds), *Systematic reviews in health care: meta-analysis in context.*, Second edn, BMJ Publishing Group., London., pp. 248-282.

Deeks, JJ 2001b, 'Systematic reviews of evaluations of diagnostic and screening tests', *British Medical Journal*, vol. 323, no. 21 July, pp. 157-162.

Denny, E & Khan, KS 2006, 'Systematic reviews of qualitative evidence: What are the experiences of women with endometriosis?', *Journal of Obstetrics and Gynaecology*, vol. 26, no. 6, pp. 501-506.

Department of Health and Human Services 2005, *Drug-diagnostic co-development: concept paper*, Food and Drug Administration (FDA).

di Ruffano, L, Davenport, C, Eising, A, Hyde, C & Deeks, J 2012, 'A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare', *Journal of Clinical Epidemiology*, vol. 65, no. 3, pp. 282-287.

Dickson, M, Hurst, J & Jacobzone, S 2003, *OECD Health Working Papers No. 4: Survey of pharmacoeconomic assessment activity in eleven countries*, no. DELSA/ELSA/WD/HEA(2003)4, Organisation for Economic Cooperation and Development, Paris, France.

Douillard, JY, Oliner, KS, Siena, S, Tabernero, J, Burkes, R, Barugel, M, Humblet, Y, Bodoky, G, Cunningham, D, Jassem, J, Rivera, F, Kocakova, I, Ruff, P, Blasinska-Morawiec, M, Smakal, M, Canon, JL, Rother, M, Williams, R, Rong, A, Wiezorek, J, Sidhu, R & Patterson, SD 2013, 'Panitumumab-FOLFOX4 treatment and RAS mutations in colorectal cancer', *New England Journal of Medicine*, vol. 369, no. 11, Sep 12, pp. 1023-1034.

Downs, SH & Black, N 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*, vol. 52, no. 6, pp. 377-384.

Draborg, E & Gyrd-Hansen, D 2005, 'Time-trends in health technology assessments: an analysis of developments in composition of international health technology assessments from 1989 to 2002', *International Journal of Technology Assessment in Health Care*, vol. 21, no. 4, Fall, pp. 492-498.

Draborg, E, Gyrd-Hansen, D, Poulsen, PB & Horder, M 2005, 'International comparison of the definition and the practical application of health technology assessment', *International Journal of Technology Assessment in Health Care*, vol. 21, no. 1, Winter, pp. 89-95.

Edward, SJ, Stevens, AJ, Braunholtz, DA, Lilford, RJ & Swift, T 2005, 'The ethics of placebo-controlled trials: a comparison of inert and active placebo controls.', *World Journal of Surgery*, vol. 29, no. 5, pp. 610-614.

Egger, M, Ebrahim, J & Davey Smith, G 2002, ' Where now for metaanalysis?', *International Journal of Epidemiology*, vol. 31, pp. 1-5.

Eikelboom, JW, Mehta, SR, Pogue, J & Yusuf, S 2001, 'Safety outcomes in meta-analyses of phase 2 vs phase 3 randomized trials: Intracranial hemorrhage in trials of bolus thrombolytic therapy', *Journal of the American Medical Association*, vol. 285, no. 4, pp. 444-450.

Elwood, JM 1998, *Critical appraisal of epidemiological studies and clinical trials*, Second edition, Oxford University Press, Oxford.

Essers, BA, Seferina, SC, Tjan-Heijnen, VC, Severens, JL, Novak, A, Pompen, M, Oron, UH & Joore, MA 2010, 'Transferability of Model-Based Economic Evaluations: The Case of Trastuzumab for the Adjuvant Treatment of HER2-Positive Early Breast Cancer in the Netherlands', *Value in Health*, vol. 13, no. 4, Jan 15, pp. 375-380.

European Network for Health Technology Assessment (EUnetHTA) 2008, *HTA Core Model for Diagnostic Technologies. Work Package 4*, European Network for Health Technology Assessment.

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group 2007, 'Recommendations from the EGAPP Working Group: testing for cytochrome P450 polymorphisms in adults with nonpsychotic depression treated with selective serotonin reuptake inhibitors', *Genetics in Medicine*, vol. 9, no. 12, Dec, pp. 819-825.

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group 2009a, 'Recommendations from the EGAPP Working Group: can UGT1A1 genotyping reduce morbidity and mortality in patients with metastatic colorectal cancer treated with irinotecan?', *Genetics in Medicine*, vol. 11, no. 1, Jan, pp. 15-20.

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group 2009b, 'Recommendations from the EGAPP Working Group: genetic testing strategies in newly diagnosed individuals with colorectal cancer aimed at reducing morbidity and mortality from Lynch syndrome in relatives', *Genetics in Medicine*, vol. 11, no. 1, pp. 35-41.

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group 2011, 'Recommendations from the EGAPP Working Group: routine testing for Factor V Leiden (R506Q) and prothrombin (20210G>A) mutations in adults with a history of idiopathic venous thromboembolism and their adult family members', *Genetics in Medicine*, vol. 13, no. 1, Jan, pp. 67-76.

Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group 2013, 'Recommendations from the EGAPP Working Group: can testing of tumor tissue for mutations in EGFR pathway downstream effector genes in patients with metastatic colorectal cancer improve health outcomes by guiding decisions regarding anti-EGFR therapy?', *Genetics in Medicine*, vol. 15, no. 7, Jul, pp. 517-527.

Evaluation of Genomic Applications in Practice and Prevention Working Group 2013, 'The EGAPP initiative: lessons learned', *Genetics in Medicine*, Aug 8.

Faulkner, E, Annemans, L, Garrison, L, Helfand, M, Holtorf, AP, Hornberger, J, Hughes, D, Li, T, Malone, D, Payne, K, Siebert, U, Towse, A, Veenstra, D & Watkins, J 2012, 'Challenges in the development and reimbursement of personalized medicine-payer and manufacturer perspectives and implications for health economics and outcomes research: a report of the ISPOR personalized medicine special interest group', *Value in Health*, vol. 15, no. 8, Dec, pp. 1162-1171.

Finkelstein, Y, Bournissen, FG, Hutson, JR & Shannon, M 2009, 'Polymorphism of the ADRB2 gene and response to inhaled beta- agonists in children with asthma: a meta-analysis', *Journal of Asthma*, vol. 46, no. 9, Nov, pp. 900-905.

Fleeman, N, Martin Saborido, C, Payne, K, Boland, A, Dickson, R, Dundar, Y, Fernandez Santander, A, Howell, S, Newman, W, Oyee, J & Walley, T 2011, 'The clinical effectiveness and cost-effectiveness of genotyping for CYP2D6 for the management of women with breast cancer treated with tamoxifen: a systematic review', *Health Technology Assessment*, vol. 15, no. 33, Sep, pp. 1-102.

Food and Drug Administration (FDA) 2011, *In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff*, Food and Drug Administration (FDA), U.S. Department of Health and Human Services, Rockville, Maryland.

Food and Drug Administration (FDA) 2011a, *In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff*, U.S. Department of Health and Human Services, Rockville, Maryland.

Food and Drug Administration (FDA) 2011b, *In Vitro Companion Diagnostic Devices: Draft guidance for industry and Food and Drug Administration staff*, Food and Drug Administration (FDA), U.S. Department of Health and Human Services, Rockville, Maryland.

Food and Drug Administration. Expert Working Group (Efficacy) of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH) *Guidance for Industry: E15 Definitions for Genomic Biomarkers,* *Pharmacogenomics, Pharmacogenetics, Genomic Data and Sample Coding Categories.* U.S. Department of Health and Human Services.

Fronsdal, KB, Facey, K, Klemp, M, Norderhaug, IN, Morland, B & Rottingen, JA 2010, 'Health technology assessment to optimize health technology utilization: using implementation initiatives and monitoring processes', *International Journal of Technology Assessment in Health Care*, vol. 26, no. 3, Jul, pp. 309-316.

Fryback, DG & Thornbury, JR 1991, 'The efficacy of diagnostic imaging', *Medical Decision Making*, vol. 11, pp. 88-94.

Garau, M, Towse, A, Garrison, L, Housman, L & Ossa, D 2012, 'Can and should value-based pricing be applied to molecular diagnostics?', *Personalized Medicine*, vol. 10, no. 1, 2013/01/01, pp. 61-72.

Garrison, LP & Austin, MJF 2007, 'The Economics of Personalized Medicine: A Model of Incentives for Value Creation and Capture', *Drug Information Journal*, vol. 41, pp. 501-509.

Gillespie, J, Guarnieri, C, Phillips H & Bhatti, T 2009, *Urinary metabolic profiling for detection of metabolic disorders*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au >.

Gillespie, J, Smala, A, Walters, N & Birinyi-Strachan, L 2007, *Hepatitis B virus DNA testing*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Ginsburg, GS & Kuderer, NM 2012, 'Comparative effectiveness research, genomics-enabled personalized medicine, and rapid learning health care: a common bond', *Journal of Clinical Oncology*, vol. 30, no. 34, Dec 1, pp. 4233-4242.

Gopalakrishna, G, Mustafa, RA, Davenport, C, Scholten, RJ, Hyde, C, Brozek, J, Schunemann, HJ, Bossuyt, PM, Leeflang, MM & Langendam, MW 2014, 'Applying Grading of Recommendations Assessment, Development and Evaluation (GRADE) to diagnostic tests was challenging but doable', *Journal of Clinical Epidemiology*, vol. 67, no. 7, Jul, pp. 760-768.

Guyatt, GH, Oxman, AD, Kunz, R, Falck-Ytter, Y, Vist, GE, Liberati, A & Schunemann, HJ 2008, 'Going from evidence to recommendations', *British Medical Journal*, vol. 336, no. 7652, May 10, pp. 1049-1051.

Haddow, J & Palomaki, G 2002, *Population based prenatal screening for cystic fibrosis via carrier testing: ACCE review*, Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention Atlanta, Georgia. Available at:
<www.cdc.gov/genomics/activities/FBR/ACCE.htm>.

Hailey, D 2003, *HTA Initiative #9. Elements of effectiveness for Health Technology Assessment Programs*, Alberta Heritage Foundation for Medical Research, Edmonton, Alberta.

Hailey, D 2009, 'The history of health technology assessment in Australia', *International Journal of Technology Assessment in Health Care*, vol. 25 Suppl 1, Jul, pp. 61-67.

Harbord, R, Bachmann, L, Shang, A, Whiting, P, Deeks, J, Egger, M & Sterne, J 2005, 'An empirical comparison of methods for meta-analysis of studies of diagnostic accuracy', *XIII Cochrane Colloquium,* Melbourne, Australia.

Harris, R, Helfand, M, Woolf, S, Lohr, K, Mulrow, C, Teutsch, S, Atkins, D & for the Methods Work Group Third US Preventive Services Task Force 2001, 'Current methods of the US Preventive Services Task Force: a review of the process', *American Journal of Preventive Medicine*, vol. 20, no. 3 Suppl, pp. 21-35.

Henshall, C et al. 1997, 'Priority setting for health technology assessment: theoretical considerations and practical approaches', *International Journal of Technology Assessment in Health Care*, vol. 13, pp. 144-185.

Higgins, J & Green, S (eds) 2005, *Cochrane Handbook for Systematic Reviews of Interventions 4.2.5 [updated May 2005].* The Cochrane Library, vol. Issue 3, John Wiley & Sons, Ltd., Chichester, UK.

Hivon, M, LeHoux, P, Denis, J-L & Tailliez S. 2005, 'Use of health technology assessment in decision making: Coresponsibility of users and producers?', *International Journal of Technology Assessment in Health Care*, vol. 21, no. 2, pp. 268-275.

Holmes, MV, Shah, T, Vickery, C, Smeeth, L, Hingorani, AD & Casas, JP 2009, 'Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies', *PLoS One*, vol. 4, no. 12, p. e7960.

HTAi and INAHTA 2007, *Resources for health technology assessment*, Health Technology Assessment international and the International Network of Agencies for Health Technology Assessment", viewed 19/10/2007 2007. Available at:
<http://www.inahta.org/upload/HTA_resources/AboutHTA_Resources_for_HTA.pdf>.

Ioannidis, J, Trikalinos, T, Ntzani, E & Contopoulos-Ioannidis, D 2003, 'Genetic associations in large versus small studies: an empirical assessment.', *Lancet*, no. 361, pp. 567-571.

Jackson, BR 2008, 'The dangers of false-positive and false-negative test results: false-positive results as a function of pretest probability', *Clinical Laboratory Medicine*, vol. 28, no. 2, pp. 305-319, vii.

Jackson, JB & Balfour, HH, Jr. 1988, 'Practical diagnostic testing for human immunodeficiency virus', *Clinical Microbiology Reviews*, vol. 1, no. 1, Jan, pp. 124-138.

Jackson, T 2007, 'Health technology assessment in Australia: challenges ahead', *Medical Journal of Australia*, vol. 187, no. 5, pp. 262-264.

Karapetis, CS, Khambata-Ford, S, Jonker, DJ, O'Callaghan, CJ, Tu, D, Tebbutt, NC, Simes, RJ, Chalchal, H, Shapiro, JD, Robitaille, S, Price, TJ, Shepherd, L, Au, HJ, Langer, C, Moore, MJ & Zalcberg, JR 2008, 'K-ras mutations and benefit from cetuximab in advanced colorectal cancer', *New England Journal of Medicine*, vol. 359, no. 17, Oct 23, pp. 1757-1765.

Khan, KS, Ter Riet, G, Glanville, JM, Sowden, AJ & Kleijnen, J 2001, *Undertaking systematic reviews of research on effectiveness. CRD's guidance for those carrying out or commissioning reviews*, CRD Report, no. CRD Report Number 4 (second edition), NHS Centre for Reviews and Dissemination, University of York, York.

Kisser, A & Zechmeister, I 2014, *Procedural Guidance for the systematic evaluation of biomarker tests*, Decision Support Document Nr. 77, Ludwig Boltzmann Institute, Vienna, Austria.

Kunz, R & Oxman, AD 1998, 'The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials', *British Medical Journal*, vol. 317, no. 7167, pp. 1185-1190.

Laksman Z & Detsky AS 2011, 'Personalized medicine: understanding probabilities and managing expectations', *Journal of Genetics in Internal Medicine*, vol. 26, no. 2, pp. 204-206.

Lancet Editorial 1983, 'Opren scandal', *Lancet*, vol. 1, pp. 219-220.

Lee, CK, Lord, SJ, Coates, AS & Simes, RJ 2009, 'Molecular biomarkers to individualise treatment: assessing the evidence', *Medical Journal of Australia*, vol. 190, no. 11, Jun 1, pp. 631-636.

Leeflang, MM, Scholten, RJ, Rutjes, AW, Reitsma, JB & Bossuyt, PM 2006, 'Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies', *Journal of Clinical Epidemiology*, vol. 59, no. 3, Mar, pp. 234-240.

Leung, GM, Woo, PP, Cowling, BJ, Tsang, CS, Cheung, AN, Ngan, HY, Galbraith, K & Lam, TH 2008, 'Who receives, benefits from and is harmed by cervical and breast cancer screening among Hong Kong Chinese?', *Journal of Public Health (Oxf).* vol. 30, no. 3, pp. 282-292. Epub 2008 May 2014.

Lijmer, JG, Mol, BW, Heisterkamp, S, Bonsel, GJ, Prins, MH, van der Meulen, JH & Bossuyt, PM 1999, 'Empirical evidence of design-related bias in studies of diagnostic tests.', *Journal of the American Medical Association*, vol. 282, no. 11, pp. 1061 - 1066.

Lord, S, Lei, W, Griffiths, A, Walleser, S, Parker, S, Thongyoo, S & Eckermann, S 2007, *Breast magnetic resonance imaging*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Lord, SJ, Irwig, L & Bossuyt, PM 2009, 'Using the principles of randomized controlled trial design to guide test evaluation', *Medical Decision Making*, vol. 29, no. 5, Sep-Oct, pp. E1-E12.

Luce, BR, Drummond, M, Jonsson, B, Neumann, PJ, Schwartz, JS, Siebert, U & Sullivan, SD 2010, 'EBM, HTA, and CER: clearing the confusion', *Milbank Quarterly*, vol. 88, no. 2, Jun, pp. 256-276.

Lurje, G & Lenz, HJ 'EGFR Signaling and Drug Discovery', *Oncology*, vol. 77, no. 6, Feb 2, pp. 400-410.

Mallett, S, Deeks, J, Halligan, S, Hopewell, S, Cornelius, V & Altman, D 2006, ' Systematic review of diagnostic tests in cancer: review of methods and reporting', *British Medical Journal*, vol. 333, p. 413.

Marinovich, L 2009, *Optical Coherence Tomography*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Marinovich, L & Wortley, S 2010, *Positron emission tomography for glioma*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

McClain, M, Palomaki, G, Piper, M & Haddow, J 2008, 'A rapid-ACCE review of CYP2C9 and VKORC1 alleles testing to inform warfarin dosing in adults at elevated risk for thrombotic events to avoid serious bleeding.', *Genetics in Medicine*, vol. 10, no. 2, pp. 89-98.

McEwen, J 2007, *A history of therapeutic goods regulation in Australia*, Commonwealth of Australia, Canberra, ACT.

Meckley, LM, Gudgeon, JM, Anderson, JL, Williams, MS & Veenstra, DL 'A policy model to evaluate the benefits, risks and costs of warfarin pharmacogenomic testing', *Pharmacoeconomics*, vol. 28, no. 1, pp. 61-74.

Meckley, LM & Neumann, PJ 2010, 'Personalized medicine: factors influencing reimbursement', *Health Policy*, vol. 94, no. 2, Feb, pp. 91-100.

Medical Services Advisory Committee (MSAC) 2000, *Funding for new medical technologies and procedures: application and assessment guidelines*, Commonwealth of Australia, Canberra, ACT.

Medical Services Advisory Committee (MSAC) 2005a, *Guidelines for the assessment of diagnostic technologies*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2005b, *Report of a review of the Medical Services Advisory Committee*, Medical Services Advisory Committee, Canberra, ACT.

Medical Services Advisory Committee (MSAC) 2012a, *Public Summary Document Application No. 1163 - Assessment of HER2 gene amplification for use of trastuzumab in gastric cancer*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2012b, *Public Summary Document Application No. 1174. Assessment of viral tropism testing of HIV to inform treatment with maraviroc*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2012c, *Public Summary Document. Application No. 1230 – HER2 ISH testing for access to trastuzumab for neoadjuvant breast cancer*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2013a, *Out-of-Session MSAC Consideration – October 2013. Application 1161 - Gefitinib first line testing for mutations of epidermal growth factor receptor (EGFR) in patients with locally advanced or metastatic non-small cell* *lung cancer (NSCLC),* Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2013b, *Out-of-Session MSAC Consideration – September 2013. Application 1173 - Testing for epidermal growth factor receptor (EGFR) status in patients with locally advanced (stage IIIB) or metastatic (stage IV) non-small cell lung cancer (NSCLC) for access to erlotinib,* Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2013c, *Public Summary Document Application No. 1172 – BRAF genetic testing in patients with melanoma for access to proposed PBS-funded vemurafenib*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2013d, *Public Summary Document Application No. 1207 – Testing for V600 status in patients with locally advanced or metastatic melanoma for eligibility for dabrafenib treatment*, Commonwealth of Australia, Canberra, ACT. Available at: <www.msac.gov.au>.

Medical Services Advisory Committee (MSAC) 2015, *Technical Guidelines for the Assessment of Investigative Technologies*, Commonwealth of Australia, Canberra, ACT. In press.

Merlin, T 2014, 'The use of the 'linked evidence approach' to guide policy on the reimbursement of personalized medicines', *Personalized Medicine*, vol. 11, no. 4, 2014/06/01, pp. 435-448.

Merlin, T, Farah, C, Schubert, C, Mitchell, A, Hiller, JE & Ryan, P 2013, 'Assessing personalized medicines in Australia: a national framework for reviewing codependent technologies', *Medical Decision Making*, vol. 33, no. 3, Apr, pp. 333-342.

Merlin, T, Lehman, S, Hiller, JE & Ryan, P 2013, 'The "linked evidence approach" to assess medical tests: a critical analysis', *International Journal of Technology Assessment in Health Care*, vol. 29, no. 3, Jul, pp. 343-350.

Merlin, T, Middleton, P, Salisbury, J & Weston, A 2005, 'Ways to ensure evidence-based clinical practice guidelines are of high quality', *XIII Cochrane Colloquium,* Melbourne, Australia.

Merlin, T, Moss, J, Brooks, A, Newton, S, Hedayati, H & Hiller, J 2008, *B-type natriuretic peptide assays in the diagnosis of heart failure*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Merlin, T, Weston, A & Tooher, R 2005a, 'Re-assessing and revising "levels of evidence" in the critical appraisal process', *XIII Cochrane Colloquium,* Melbourne, Australia.

Merlin, T, Weston, A & Tooher, R 2005b, 'Revising a national standard: redevelopment of the Australian NHMRC evidence hierarchy', *Italian Journal of Public Health (Supplement 1)*, vol. 2, no. 2, p. 156.

Merlin, T, Weston, A & Tooher, R 2009, 'Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'', *BMC Medical Research Methodology*, vol. 9, p. 34.

Micheel, C & Ball, J (eds) 2010, *Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease*, National Academy of Sciences, Washington DC.

Middleton, P, Tooher, R, Salisbury, J, Coleman, K, Norris, S, Grimmer, K & Hillier, S 2005, 'Assessing the body of evidence and grading recommendations in evidence-based clinical practice guidelines', *XIII Cochrane Colloquium,* Melbourne, Australia.

Mina, K 2012, *Report of the RCPA Genetic Testing Survey 2011*, Royal College of Pathologists of Australasia, RCoPo Australasia.

Moher, D, Schulz, K & Altman, D 2001, 'The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials', *Journal of the American Medical Association*, vol. 285, pp. 1987-1991.

Moynihan, R 2013, 'What is disease? And why it's a healthy question', *British Medical Journal*, vol. 346, p. f107.

Muir Gray, J 2001, 'Tests', *Evidence-based health care. How to make health policy and management decisions.*, 2nd edition edn, Churchill-Livingstone, Edinburgh, pp. 72-84.

Mulrow, CD, Cook, DJ & Davidoff, F 1997, 'Systematic reviews: Critical links in the great chain of evidence', *Annals of Internal Medicine*, no. 126, 1 March 1997, pp. 389 - 391.

National Institute for Health and Care Excellence (NICE) 2011, *Diagnostics Assessment Programme Manual*, Manchester, UK.

National Institute for Health and Care Excellence (NICE) August 2013, 'EGFR-TK mutation testing in adults with locally advanced or metastatic non-small-cell lung cancer', *NICE diagnostics guidance 9*, National Institute for Health and Care Excellence, United Kingdom. Available at: <www.nice.org.uk/dg9>.

National Institute for Health and Clinical Excellence (NICE) 2007, *The guidelines manual*, National Institute for Health and Clinical Excellence, London.

National Institute for Health and Clinical Excellence (NICE) 2010, *Interim methods statement (pilot). Version 8*, National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme, London.

National Institute for Health and Clinical Excellence (NICE) 2011, *Diagnostics Assessment Programme Manual*, National Institute of Health and Clinical Excellence Centre for Health Technology Evaluation, Diagnostics Assessment Programme, London.

Nelson, H, Huffman, L, Fu, R & Harris, E 2005, 'Genetic Risk Assessment and BRCA Mutation Testing for Breast and Ovarian Cancer Susceptibility: Systematic Evidence Review for the U.S. Preventive Services Task Force', *Annals of Internal Medicine*, vol. 143, pp. 362-379.

Nelson, H, Pappas, M, Zakher, B, Mitchell, J, Okinaka-Hu, L & Fu, R 2014, 'Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer in Women: A Systematic Review to Update the U.S. Preventive Services Task Force Recommendation', *Annals of Internal Medicine*, vol. 160, pp. 255-266.

NHMRC 1999a, *A guide to the development, implementation and evaluation of clinical practice guidelines*, National Health and Medical Research Council, Commonwealth of Australia, Canberra, ACT.

NHMRC 1999b, *How to present the evidence for consumers: preparation of consumer publications*, National Health and Medical Research Council, Canberra.

NHMRC 2000a, *How to put the evidence into practice: implementation and dissemination strategies.*, National Health and Medical Research Council, Canberra.

NHMRC 2000b, *How to review the evidence: systematic identification and review of the scientific literature*, National Health and Medical Research Council, Canberra.

NHMRC 2000c, *How to use the evidence: assessment and application of scientific evidence*, National Health and Medical Research Council, Canberra.

NHMRC 2001, *How to compare the costs and benefits: evaluation of the economic evidence*, National Health and Medical Research Council, Canberra.

NHMRC 2003, *Using socioeconomic evidence in clinical practice guidelines*, Commonwealth of Australia, Canberra, ACT.

NHMRC 2008, *NHMRC additional levels of evidence and grades for recommendations for developers of guidelines. Stage 2 consultation. Early 2008 - end June 2009*, National Health and Medical Research Council, Canberra, ACT, viewed 6/8/08 2008. Available at: <http://www.nhmrc.gov.au/guidelines/consult/consultations/add_levels_grades_dev_guidelines2.htm>.

Noyes, J, Popay, J, Pearson, A, Hannes, K & Booth, A 2008, 'Chapter 20: Qualitative research and Cochrane reviews', in J Higgins & S Green (eds), *Cochrane Handbook of Systematic Reviews of Interventions, Version 5.0.1*, Version 5.0.1 (updated September 2008) edn, The Cochrane Collaboration.

OECD 2005, 'Chapter 4. Decision-making and implementation: an analysis of survey results.', *Health Technologies and Decision Making. The OECD Health project*, OECD, Paris, France, pp. 71-94.

Parliamentary Office of Science and Technology 2009, *Postnote: Personalised medicine*, vol. 329, London, UK.

Payne, K & Annemans, L 2013, 'Reflections on market access for personalized medicine: recommendations for Europe', *Value in Health*, vol. 16, no. 6 Suppl, Sep-Oct, pp. S32-38.

Petherick, ES, Villaneuva, EV, Dumville, J, Bryan, EJ & Dharmage, S 2007, 'An evaluation of methods used in health technology assessments produced for the Medical Services Advisory Committee', *Medical Journal of Australia*, vol. 187, no. 5, pp. 289-292.

Pharmaceutical Benefits Advisory Committee (PBAC) 2013, *Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.4)*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2012a, *Public Summary Document, Product: Trastuzumab, powder for I.V. infusion, 60 mg and 150 mg, Herceptin®*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2012b, *Public Summary Document, Product: Maraviroc, tablets, 150 mg and 300 mg, Celsentri®*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2013a, *ADDENDUM – July 2013. Product: Dabrafenib, capsules, 50 mg and 75 mg Tafinlar®*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2013b, *Public Summary Document, Product: Vemurafenib; tablet, 240 mg, Zelboraf®*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2013c, *Public Summary Document, Product: Erlotinib, tablets, 25 mg, 100 mg, 150 mg (as hydrochloride), Tarceva®*, Commonwealth of Australia, Canberra, ACT.

Pharmaceutical Benefits Advisory Committee (PBAC) 2013d, *Public Summary Document, Product: Gefitinib, tablet, 250mg, Iressa®*, Commonwealth of Australia, Canberra, ACT.

Phillips, B, Ball, C, Sackett, D, Badenoch, D, Straus, S, Haynes, B & Dawes, M 2001, *Oxford Centre for Evidence-Based Medicine Levels of Evidence (May 2001)*, Centre for Evidence-Based Medicine, Oxford.

Popay, J (ed.) 2006, *Moving beyond effectiveness in evidence synthesis: methodological issues in the synthesis of diverse sources of evidence*, National Institute for Health and Clinical Excellence, London.

Productivity Commission 2005, *Impacts of advances in medical technology in Australia. Research report*, Research report, Productivity Commission, Australian Government, Melbourne, Victoria.

Regan, MM, Leyland-Jones, B, Bouzyk, M, Pagani, O, Tang, W, Kammler, R, Dell'orto, P, Biasi, MO, Thurlimann, B, Lyng, MB, Ditzel, HJ, Neven, P, Debled, M, Maibach, R, Price, KN, Gelber, RD, Coates, AS, Goldhirsch, A, Rae, JM & Viale, G 2012, 'CYP2D6 genotype and tamoxifen response in postmenopausal women with endocrine-responsive breast cancer: the breast international group 1-98 trial', *Journal of the National Cancer Institute*, vol. 104, no. 6, Mar 21, pp. 441-451.

Roberts, MS 2006, 'The Use of Decision Analysis for Understanding the Impact of Diagnostic Testing Errors in Pathology', *American Journal of Clinical Pathology*, vol. 126, pp. S36-S43.

Roden D.M., Altman R.B., Benowitz M.D. & et al. 2006, 'Pharmacogenomics: challenges and opportunities', *Annals of Internal Medicine*, vol. 145, pp. 749-757.

Rowley, P, Haddow, JE & Palomaki, GE *DNA testing strategies aimed at reducing morbidity and mortality from hereditary non-polyposis colorectal cancer (HNPCC): An ACCE mini-review* Office of Genomics and Disease Prevention, Centers for Disease Control and Prevention, Atlanta, Georgia. Available at: <http://www.cdc.gov/genomics/gtesting/ACCE/FBR/>.

Sackett, DL & Haynes, RB 2002, 'The architecture of diagnostic research', *British Medical Journal*, vol. 324, pp. 539-541.

Sanderson, S, Tatt, ID & Higgins, JP 2007, 'Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography', *International Journal of Epidemiology*, vol. 36, no. 3, Jun, pp. 666-676.

Scaf-Klomp, W, Sanderman, R, van de Wiel, HB, Otter, R & van den Heuvel, WJ 1997, 'Distressed or relieved? Psychological side effects of breast cancer screening in The Netherlands', *Journal of Epidemiology and Community Health.*, vol. 51, no. 6, pp. 705-710.

Scartozzi, M, Bearzi, I, Mandolesi, A, Pierantoni, C, Loupakis, F, Zaniboni, A, Negri, F, Quadri, A, Zorzi, F, Galizia, E, Berardi, R, Biscotti, T, Labianca, R, Masi, G, Falcone, A & Cascinu, S 2009, 'Epidermal Growth Factor Receptor (EGFR) gene copy number (GCN) correlates with clinical activity of irinotecan-cetuximab in K-RAS wild-type colorectal cancer: a fluorescence in situ (FISH) and chromogenic in situ hybridization (CISH) analysis', *BMC Cancer*, vol. 9, p. 303.

Schmitt, F 2009, 'HER2+ breast cancer: how to evaluate?', *Advanced Therapeutics*, vol. 26 Suppl 1, Jul, pp. S1-8.

Schoeppe, S, Lewis, S, Marinovich, L & Wortley, S 2010, *Positron emission tomography for cervical cancer*, Commonwealth of Australia, Canberra. Available at: <www.msac.gov.au>.

Schunemann, HJ, Oxman, AD, Brozek, J, Glasziou, P, Bossuyt, P, Chang, S, Muti, P, Jaeschke, R & Guyatt, GH 2008, 'GRADE: assessing the quality of evidence for diagnostic recommendations', *Evidence Based Medicine*, vol. 13, no. 6, Dec, pp. 162-163.

Schunemann, HJ, Oxman, AD, Brozek, J, Glasziou, P, Jaeschke, R, Vist, GE, Williams, JW, Kunz, R, Craig, J, Montori, VM, Bossuyt, PM & Guyatt, GH 2008, 'Grading quality of evidence and strength of recommendations for diagnostic tests and strategies', *British Medical Journal*, vol. 336, pp. 1106-1110.

Scottish Intercollegiate Guidelines' Network (SIGN) 2008, *SIGN 50: a guideline developer's handbook*, SIGN, Edinburgh.

Shak, S 1999, 'Overview of the trastuzumab (Herceptin) anti-HER2 monoclonal antibody clinical program in HER2-overexpressing metastatic breast cancer. Herceptin Multinational Investigator Study Group', *Seminars in Oncology*, vol. 26, no. 4 Suppl 12, Aug, pp. 71-77.

Shickle, D & Chadwick, R 1994, 'The ethics of screening: is 'screeningitis' an incurable disease?', *Journal of Medical Ethics.*, vol. 20, no. 1, pp. 12-18.

Staub, L, Dyer, S, Lord, S & Simes RJ 2012, 'Linking the Evidence: Intermediate Outcomes in Medical Test Assessments', *International Journal of Technology Assessment in Health Care*, vol. 28, no. 1, pp. 52-58.

Steinberg EP, Tunis S & Shapiro D 1995, 'Insurance coverage for experimental technologies.', *Health Affairs*, vol. 14, pp. 143-158.

Suthers, G 2008, *Report of the Australian Genetic Testing Survey 2006*, Royal College of Pathologists of Australasia, Adelaide.

Tamura, K, Okamoto, I, Kashii, T, Negoro, S, Hirashima, T, Kudoh, S, Ichinose, Y, Ebi, N, Shibata, K, Nishimura, T, Katakami, N, Sawa, T, Shimizu, E, Fukuoka, J, Satoh, T & Fukuoka, M 2008, 'Multicentre prospective phase II trial of gefitinib for advanced non-small cell lung cancer with epidermal growth factor receptor mutations: results of the West Japan Thoracic Oncology Group trial (WJTOG0403)', *British Journal of Cancer*, vol. 98, no. 5, Mar 11, pp. 907-914.

Terasawa, T, Dahabreh, I, Castaldi, P & Trikalinos, T 2009, *Systematic Reviews on Selected Pharmacogenetic Tests for Cancer Treatment: CYP2D6 for Tamoxifen in Breast Cancer, KRAS for anti-EGFR antibodies in Colorectal Cancer, and BCR-ABL1 for Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia. Draft.*, AHRQ Technology Assessment Report, Agency for Healthcare Research and Quality, Rockville, MD.

Teutsch, SM, Bradley, LA, Palomaki, GE, Haddow, JE, Piper, M, Calonge, N, Dotson, WD, Douglas, MP & Berg, AO 2009, 'The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative: methods of the EGAPP Working Group', *Genetics in Medicine*, vol. 11, no. 1, Jan, pp. 3-14.

The GRADE working group 2004, 'Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches', *BMC Health Services Research*, vol. 4, no. 1, p. 38.

The GRADE working group 2005, 'Systems for grading the quality of evidence and the strength of recommendations II: A pilot study of a new system for grading the quality of evidence and the strength of recommendations.', *BMC Health Services Research*, vol. 5, no. 1, p. 25.

The GRADE working group 2008, 'GRADE: what is quality of evidence and why is it important to clinicians?', *British Medical Journal*, vol. 336, pp. 995-998.

Therapeutic Goods Administration 2015, *Therapeutic Goods Administration: An introduction to the work of Australia's regulator of therapeutic goods*. Australian Government

Department of Health, Canberra, ACT. Available at: <https://www.tga.gov.au/introduction-tga>

Tikkinen, K, Leinonen, J, Guyatt, G & et al 2012, 'What is a disease? Perspectives of the public, health professionals and legislators.', *BMJ Open*, vol. 2, p. e001632.

Trueman, P, Grainger, D & Downs, KE 2010, 'Coverage with Evidence Development: Applications and issues', *International Journal of Technology Assessment in Health Care*, vol. 26, no. 1, pp. 79-85.

U.S. Department of Health and Human Services Food and Drug Administration 2005, 'Guidance for Industry: Pharmacogenomic Data Submissions. Procedural', Internet, FDA, Rockville, Maryland. Available at: <http://www.fda.gov/cder/guidance/index.htm>.

Van Cutsem, E, Kohne, CH, Hitre, E, Zaluski, J, Chang Chien, CR, Makhson, A, D'Haens, G, Pinter, T, Lim, R, Bodoky, G, Roh, JK, Folprecht, G, Ruff, P, Stroh, C, Tejpar, S, Schlichting, M, Nippgen, J & Rougier, P 2009, 'Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer', *New England Journal of Medicine*, vol. 360, no. 14, Apr 2, pp. 1408-1417.

Velasco-Garrido, M & Busse, R 2005, *Health technology assessment: An introduction to objectives, role of evidence, and structure in Europe. Policy Brief*, World Health Organization, on behalf of European Observatory on Health Systems and Policies, Copenhagen, Denmark.

Wald, A 1943, 'Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large', *Transactions of the American Mathematical Society*, vol. 54, pp. 426-482.

Wang, B, Wang, J, Huang, SQ, Su, HH & Zhou, SF 2009, 'Genetic Polymorphism of the Human Cytochrome P450 2C9 Gene and Its Clinical Significance', *Current Drug Metabolism*, vol. 10, no. 7, Sep, pp. 781-834.

Whiting P, RA, Reitsma JB, Bossuyt PM, Kleijnen J. 2003, 'The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews.', *BMC Medical Research Methodology*, vol. 3, no. 1, p. 25.

Whiting, PF, Rutjes, AWS, Westwood, ME, Mallett, S, Deeks, JJ, Reitsma, JB, Leeflang, MMG, Sterne, JAC & Bossuyt, PMM 2011, 'QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies', *Annals of Internal Medicine*, vol. 155, no. 8, pp. 529-536.

Whitlock, E, Garlitz, B, Harris, E, Bell, T & Smith, P 2006, 'Screening for Hereditary Hemochromatosis: A Systematic Review for the U.S. Preventive Services Task Force', *Annals of Internal Medicine*, vol. 145, pp. 209-223.

Wolcott, J, Schwartz, A & Goodman, C 2008, *Laboratory Medicine: A National Status Report*, The Lewin Group, Centers for Disease Control and Prevention (CDC) Office of Public Health Genomics.

Wong, W, Carlson, J, Thariani, R & Veenstra, D 2010, 'Cost Effectiveness of Pharmacogenomics: A Critical and Systematic Review', *Pharmacoeconomics*, vol. 28, no. 11, pp. 1001-1013.

Xie H-G & Frueh FW 2005, 'Pharmacogenomics steps toward personalized medicine', *Future Medicine*, vol. 2, no. 4, pp. 325-337.

APPENDIX: TRANSLATION INTO POLICY AND PRACTICE

Paper 1

This paper arose from voluntary methodological work I was doing to revise the National Health and Medical Research Council's (NHMRC) evidence hierarchy so that it could be used by evidence-based clinical practice guideline developers when answering questions concerning the diagnostic accuracy of a test. Preliminary work on this project began in 2005, the hierarchy was piloted in 2007 – when I enrolled in my PhD part-time – and it was eventually finalised and published in 2009. This evidence hierarchy formed part of one of the systems, known as FORM[25], mandated at the time by the NHMRC – which is the peak guideline development agency in Australia - to be used for the development and grading of national evidence-based clinical practice guidelines. Relevant publications describing or mandating use of the methodology include:

- *NHMRC additional levels of evidence and grades for recommendations for developers of guidelines*. Available at: https://www.nhmrc.gov.au/_files_nhmrc/file/guidelines/developers/nhmrc_levels_grades_evidence_120423.pdf
- Merlin T, Weston A, Tooher R. Extending an evidence hierarchy to include topics other than treatment: revising the Australian 'levels of evidence'. *BMC Medical Research Methodology*, 2009, 9:34 doi:10.1186/1471-2288-9-34. Available at: http://www.biomedcentral.com/1471-2288/9/34
- Hillier S, Grimmer-Somers K, Merlin T, Middleton P, Salisbury J, Tooher R, Weston A. FORM: An Australian method for grading recommendations in evidence-based clinical guidelines. *BMC Medical Research Methodology* 2011, 11:23 doi:10.1186/1471-2288-11-23. Available at: http://www.biomedcentral.com/1471-2288/11/23
- National Health and Medical Research Council. *Procedures and requirements for meeting the 2011 NHMRC standard for clinical practice guidelines*. Melbourne: National Health and Medical Research Council; 2011. (Requirement C.7, p18).

[25] Formulating Optimal Recommendations Methodology (FORM)

Available at: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/cp133_nhmrc_procedures_requirements_guidelines_v1.1_120125.pdf

The paper has a "Highly Accessed" designation from *BMC Medical Research Methodology*, with the article being accessed in BioMed Central 21,970 times up until April 2015. It has also been cited 103 times, according to Google Scholar.

The Australian Government has required that health technology assessment processes follow, where appropriate, the evidence-based standards and methodologies promulgated by the NHMRC. This evidence hierarchy became embedded in health technology assessment processes that are used to inform policy makers in Australia. To date, the evidence hierarchy has been used when evaluating new medical tests for 96 clinical indications in order to inform public subsidy decisions (through the Medical Benefits Schedule) by the Medical Services Advisory Committee (MSAC)[26].

Paper 2

This paper is not yet in the public domain. However, the paper itself evaluates the impact of the linked evidence approach on public funding decisions. The results indicate that the methodology has directly affected policy decisions.

Paper 3

This paper was published in 2013 and so its translational impact has yet to be fully established. However, the decision framework developed to use the linked evidence approach has been incorporated into health technology assessments evaluating medical tests for a public funding decision by the Australian Medical Services Advisory Committee (MSAC). As a consequence, several medical tests have been evaluated using this LEA decision framework (see case studies 10-13, page xvi). The approach has also informed the MSAC Guidelines on the assessment of investigational medical services, which is due for release later this year.

[26] http://www.msac.gov.au

Paper 4

This publication also affected the health technology assessment processes in Australia. It is the first framework developed internationally for the evaluation of personalised medicines (pharmacogenetic technologies), and was pilot-tested, revised, and is now in use by the Australian agencies responsible for the public funding of tests and drugs nationally – the Medical Services Advisory Committee (MSAC) and the Pharmaceutical Benefits Advisory Committee (PBAC), respectively. Industry submissions requesting funding through the Medicare Benefits Schedule (for a genetic test) and the Pharmaceutical Benefits Schedule (for the companion drug) provide information to the Government that is based on the items proposed in the evaluation framework. The method is recommended in MSAC's *Technical Guidelines for preparing assessment reports for the Medical Services Advisory Committee – Service Type: Investigative* (MSAC 2015) and in the *Guidelines for Preparing Submissions to the Pharmaceutical Benefits Advisory Committee* (Pharmaceutical Benefits Advisory Committee 2013). The approach is applied by both those who submit applications for pharmacogenetic technologies for public funding in Australia and the independent evaluators of these submissions. The policy impact of the methodology is described in Paper 5 of the thesis.

The evaluation framework has been considered and adapted for evaluations conducted by the Ludwig Boltzmann Institute, Austria (Kisser & Zechmeister 2014) and is currently being assessed for feasibility by key staff in the Canadian Agency for Drugs and Technology in Health (CADTH). It has also been the subject of presentations to the European Medicines Agency (EMA) and the National Institute for Health and Care Excellence (NICE) when developing processes for evaluating medical tests in the UK (Garau et al. 2012). Recent personalised medicine evaluations by NICE have used the method (National Institute for Health and Care Excellence August 2013).

Paper 5

This paper was published in 2014 and so its translational impact is yet to be established. The paper itself evaluates the impact of the linked evidence approach, as applied to personalised medicines, on public funding decision-making in Australia. The results indicate that the methodology has affected policy decisions.