THE "LINKED EVIDENCE APPROACH" TO ASSESS MEDICAL TESTS: A CRITICAL ANALYSIS

Tracy Merlin, Samuel Lehman
Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, School of Population Health, University of Adelaide

Philip Ryan
Data Management & Analysis Centre (DMAC), Discipline of Public Health, School of Population Health, University of Adelaide

Janet E. Hiller
Faculty of Health Sciences, Australian Catholic University

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 health technology assessment (HTA) reports informing reimbursement decisions by the Medical Services Advisory Committee (August 2005 to March 2012). Eligibility was determined according to predetermined criteria and data were extracted on test characteristics, evaluation methodologies, and reported difficulties. Fifty percent of the evidence-base was independently analyzed by a second reviewer.
Results: Evaluations of medical tests for diagnostic (62 percent), staging (27 percent), and screening (6 percent) purposes were available for eighty-nine different clinical indications. Ninety-six percent 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). Sixty-one percent had the full evidence linkage. Twenty-five percent of test evaluations were considered problematic; all involving LEA ($n = 22$). Problems included: determining test accuracy with an imperfect reference standard (41 percent); assessing likely treatment effectiveness in test positive patients when the new test is more accurate than the comparator (18 percent); and determining probable health benefits in those symptomatic patients ruled out using the test (13 percent). 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.

Keywords: Diagnostic test approval, Evaluation methodology, Systematic review, Reimbursement mechanisms

Guidance on the assessment of medical tests has been produced only recently (2008–11) 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 eighteen agencies in eight 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 randomized controlled trial whereby patients are randomized 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 U.S. and European, test evaluation guidance outlines methods to deal with this type of evidence.
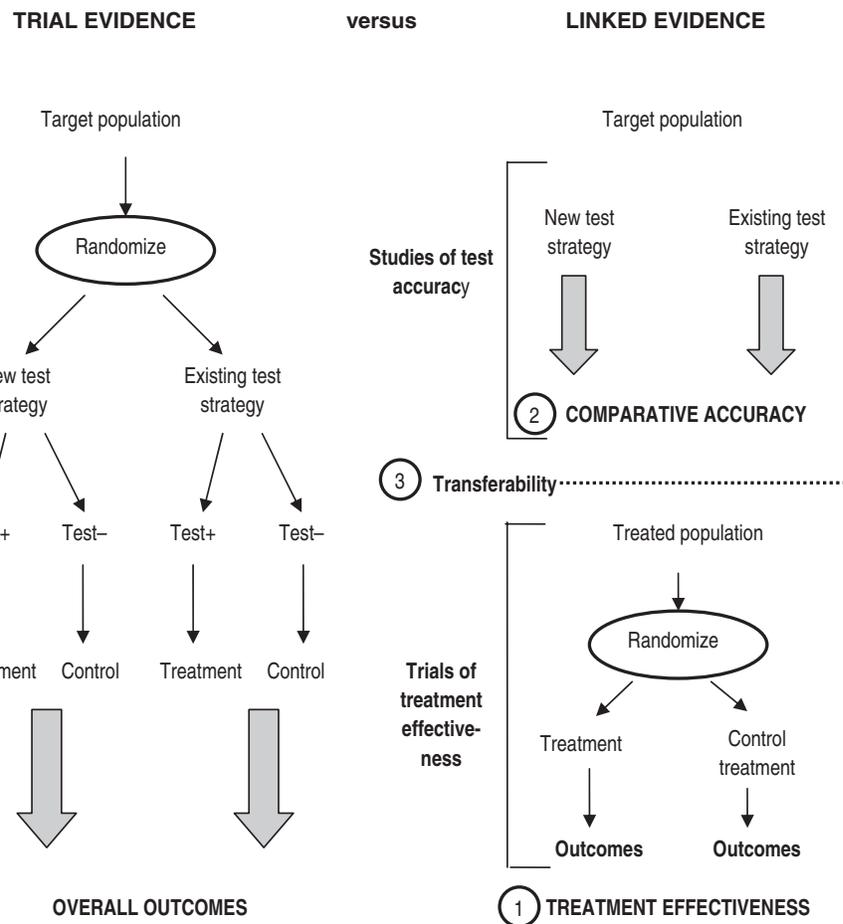
TRIAL EVIDENCE versus LINKED EVIDENCE

Figure 1. The use of direct evidence compared with linked evidence in the evaluation of diagnostic tests in a health technology assessment framework. 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.

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 new test 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).

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 (by means of 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 the findings from this study to inform the development of a decision framework for 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: (i) Considered by MSAC between August 2005 and March 2012; (ii) Publicly available on the MSAC Web site (www.msac.gov.au) between February and March 2012; (iii) A test requested for reimbursement through government referral, industry application, or an update of a previous assessment; and (iv) A test used for diagnostic, screening or staging purposes.

344

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 characterize the stage of disease in a patient previously diagnosed. Diagnostic tests that may have a therapeutic component were included, for example, a biopsy that happened to capture all of the diseased tissue and so effectively treated the condition.

HTAs were excluded in the following circumstances: (i) The test being assessed was used for monitoring a specific treatment eg titrating a drug according to a biomarker concentration; (ii) The test being assessed was pharmacogenetic, that is, part of a co-dependent technology pairing (12); or (iii) 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 percent 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: (i) "direct evidence only" - reporting only on direct clinical trials, from test to measurement of patient health outcomes; (ii) "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 (e.g., changes in patient management), and the effectiveness of consequent treatment options; (iii) "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; (iv) "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; (v) "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 characterized as having high sensitivity and specificity if, relative to an appropriate reference standard, both parameters were 85 percent 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 analyzed qualitatively.

RESULTS

Figure 2 outlines the process used to select eligible HTAs for the review. We identified test evaluations for eighty-nine clinical indications in thirty-one eligible HTA reports. Testing was reported as being undertaken for diagnostic purposes (62 percent), staging (27 percent), and for screening (6 percent). Four percent of tests were classified as both diagnostic and staging, while 1 percent were jointly diagnostic and therapeutic.

Of the eighty-nine test evaluations, 96 percent used either an abridged (where evidence is linked through to management changes but not patient outcomes) or full LEA methodology, with 61 percent undertaking the full linkage. Overall, 35 percent 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 (i.e., 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 percent in 2007 to 47–50 percent 2 to 3 years later.

In 25 percent (22/89) 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 percent of HTAs that used an approach that synthesized direct evidence alone. In the "problematic" HTAs using LEA, five main challenges were identified.

1. *Imperfect reference standard*: In 34 percent 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
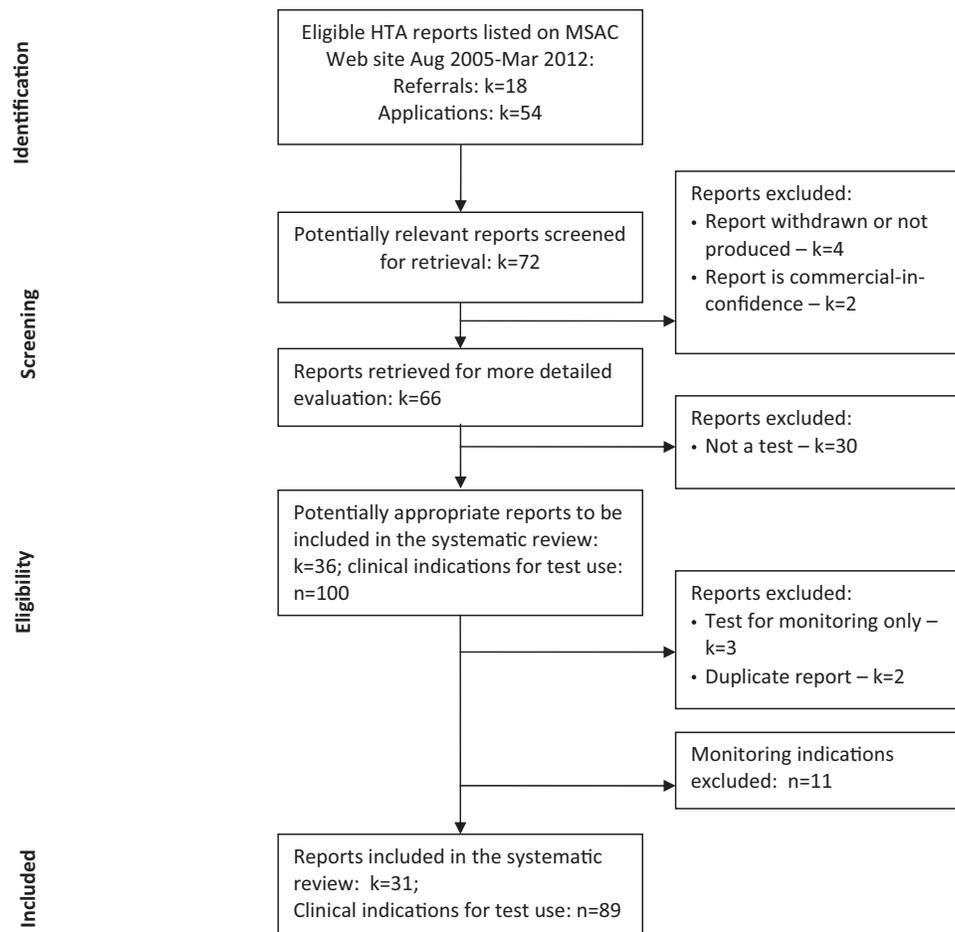
Figure 2. PRISMA flowchart. Adapted from Liberati et al. (2009).

LEA were reported ($N = 22$), 41 percent 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 percent; $N = 22$). Current treatment options would have only been trialed in populations with a spectrum of disease identified by the less accurate comparator test. Overall, 33 percent ($N = 15$) 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 percent) 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, that is, 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
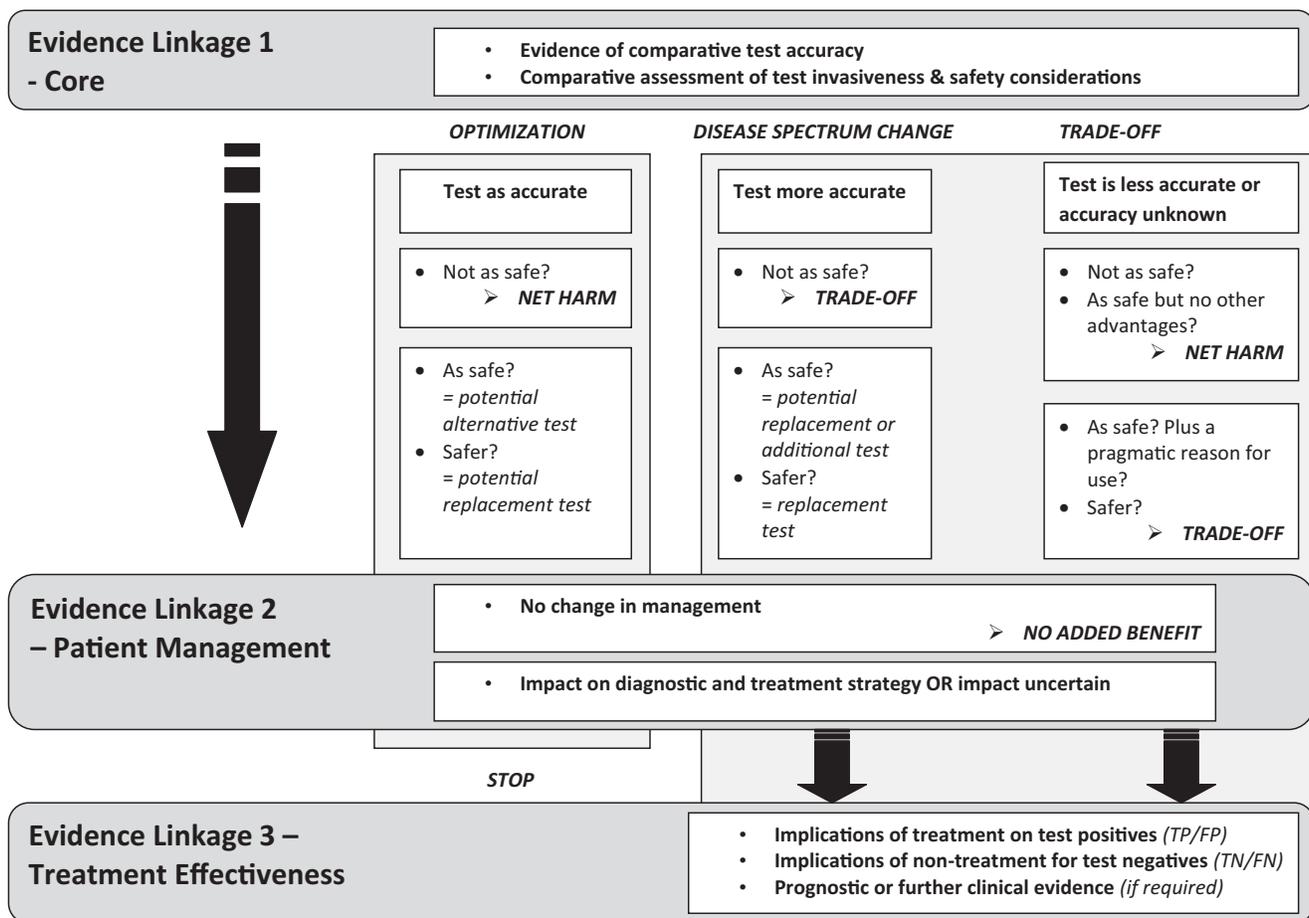
Evidence Linkage 1 - Core
- Evidence of comparative test accuracy
- Comparative assessment of test invasiveness & safety considerations

OPTIMIZATION

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)

Figure 3. Decision framework to implement the linked evidence approach when evaluating medical tests.

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. *Optimization*: 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 usage (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 optimized.

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 (e.g., 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.

C. *Disease spectrum change*: Of all the scenarios, the one where a randomized 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. As noted in scenario B, 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. Furthermore, because the data have been systematically acquired, it can then be used as inputs in the decision analytic modeling 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 (i.e., 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 randomized 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 (e.g., 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 randomized 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 before 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 conceptualizing *a priori* the overall theoretical basis for linking evidence (7), as well as the optimal study de-

signs needed to address or measure assumptions inherent in the synthesis plan (8).

Where this study differs from previous research is by proposing that any *a priori* conceptualization 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 recognizes the necessary preconditions 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 preconditions 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 standardized use of the methodology.

CONTACT INFORMATION

Tracy Merlin BA(Hons), AdvDipPM, MPH, Associate Professor, Public Health; Director, Adelaide Health Technology Assessment (AHTA), Discipline of Public Health, School of Population Health, University of Adelaide, Adelaide, South Australia, Australia

Samuel Lehman BHlthSc, Research Assistant, AHTA, Discipline of Public Health, School of Population Health, University of Adelaide, Adelaide, South Australia, Australia

Janet E. Hiller MPH, PhD, FPHAA, Executive Dean, Faculty of Health Sciences, Australian Catholic University, Fitzroy, Victoria, Australia; Adjunct Professor, School of Population Health, Faculty of Health Sciences, University of Adelaide, Adelaide, South Australia, Australia

Philip Ryan MBBS, FAFPHM, Emeritus Professor, Public Health, Data Management & Analysis Centre (DMAC), Discipline of Public Health, School of Population Health, University of Adelaide, Adelaide, South Australia, Australia

CONFLICTS OF INTEREST

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]*. Atlanta, GA: Centers for Disease Control and Prevention (CDC), Office of Public Health Genomics;

2010. http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm. (accessed April 26, 2012).

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. Copenhagen: 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 Australia; 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;29:E1-E12.

8. Staub L, Dyer S, Lord S, Simes RJ. Linking the evidence: Intermediate outcomes in medical test assessments. *Int J Technol Assess Health Care*. 2012;28:52-58.

9. 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*. 2001;20(Suppl):21-35.

10. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis 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. *J Clin Epidemiol*. 2012;65:282-287.

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. *Med Decis Making*. 2013;33:333-342.

13. Marinovich L. *Optical coherence tomography [internet]*. Canberra: Commonwealth of Australia; 2009. www.msac.gov.au. (accessed April 2, 2012).

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. www.msac.gov.au. (accessed April 2, 2012).

15. Schoeppe S, Lewis S, Marinovich L, Wortley S. *Positron emission tomography for cervical cancer [internet]*. Canberra: Commonwealth of Australia; 2010. www.msac.gov.au. (accessed April 2, 2012).

16. Lord S, Lei W, Griffiths A, et al. *Breast magnetic resonance imaging [internet]*. Canberra: Commonwealth of Australia; 2007. www.msac.gov.au. (accessed April 2, 2012).

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. www.msac.gov.au. (accessed April 26, 2012).

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. www.msac.gov.au. (accessed April 2, 2012).

20. Gillespie J, Smala A, Walters N, Birinyi-Strachan L. *Hepatitis B virus DNA testing [internet]*. Canberra: Commonwealth of Australia; 2007. www.msac.gov.au. (accessed March 30, 2012).

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. *Int J Technol Assess Health Care*. 2010;26:323-329.

22. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: Assessing new tests against existing diagnostic pathways. *BMJ*. 2006;332:1089-1092.

23. Micheel C, Ball J, eds. *Evaluation of biomarkers and surrogate endpoints in chronic disease*. Washington, DC: National Academy of Sciences; 2010.

24. Anderson L, Petticrew M, Rehfuess E, et al. Using logic models to capture complexity in systematic reviews. *Res Synth Methods*. 2011;2:33-42.