DECISION-MAKING ABOUT HEALTHCARE RELATED TESTS AND DIAGNOSTIC STRATEGIES
DECISION-MAKING ABOUT HEALTHCARE RELATED TESTS AND DIAGNOSTIC STRATEGIES

By

REEM ADEL MUSTAFA, MD, MPH

A Thesis Submitted to the School of Graduate Studies in
Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy

McMaster University

© Copyright by Reem Adel Mustafa, 2014
McMaster University

DOCTOR OF PHILOSOPHY (2014)

Hamilton, Ontario (Health Research Methodology)

TITLE: Decision-making about healthcare related tests and diagnostic strategies

AUTHOR: Reem Adel Mustafa, MD, MPH

SUPERVISOR: Professor Holger Schünemann

NUMBER OF PAGES: 375
ABSTRACT

While therapeutic interventions, especially pharmaceuticals, are increasingly subject to critical evaluation of the evidence supporting such interventions and decision-making processes to develop recommendations about them, such approaches and requirements either do not exist for healthcare related tests and diagnostic strategies (HCTDS) or there is a lack of clear regulation on how tests and strategies should be used. Scientists around the world, however, have made important progress on how such methods could be applied by public, policy makers and private funders in order to base decisions in favour or against their use on the best available evidence. These methods call for consideration of the always limited resources that are available to provide the best possible health care to the population. This thesis, based on a mixed-methods approach, provides a systematic evaluation of the available published and grey literature. In this thesis I describe the methodological approach used and the results from the work addressing the following topics: 1. Background about the challenges encountered in the field of diagnosis, 2. How to present evidence from test accuracy systematic reviews and our confidence in this evidence, 3. Summaries and syntheses of evidence about HCTDS. I also provide findings from in-depth interviews with experts and key informants’ experience and views about what is being done in major organizations and societies and how can this field move forward. I then conclude with recommendations on appropriate steps and a framework to support decision makers including guideline panels when making decisions about HCTDS. I also summarize additional products based on our work and future direction and projects of this research program.
ACKNOWLEDGMENTS

The past four years have been a life changing experience. Finishing this work would have not been possible without the help of many people. First, I want to thank my dissertation supervisor, Dr. Holger Schünemann for his outstanding mentorship. His hard work and passion for research are contagious. He sets a standard for dedication. You have helped me grow academically and for this, I am forever grateful.

Thank you to my committee members: Dr. Patrick Bossuyt for his valuable feedback and input, Dr. Jan Brozak for his outstanding support, friendship and great advice, both professionally and personally, and Dr. Amit Garg for his tremendous support and for being the role model nephrologist who inspired me.

Thank you to all the wonderful co-authors that I had the privilege to work with in all my projects. I could not have completed this thesis without your help.

Thank you to all the great people in the HRM program, from the students to the faculty to the administrative team that make McMaster the amazing place it is. I am blessed to be a part of such a great team and to have had the chance to meet you.

Thank you to my sister, my best friend, and my soul mate, Dana, who has supported me through all the milestones in my life. Thanks to my brothers, Mahmoud, Mohammad and Ahmad for always being there for me, for your encouragement, warm hugs and humour that helped me get through the rough ride. Thank you to my dear friends Nancy, Rasha and Lucy for being there when I needed you, for your listening ears and encouragement. You have all made me the person I am today.
DEDICATION

To the loving memory of my father, whose words has always been the guiding light in my journey. Who believed in me and inspired me to reach for the stars.

You will always be missed.

To my mother, the source of unconditional love and endless support in my life.

You have always been the shoulder I lean on.

To my husband, my best friend, and my confidant. To you Fadi, for your patience, understanding, and love. Without your support, I would not have been where I am today.

To my children, Faris and Liane, to your contagious curiosity and infinite questions. You taught me there is always more to be learned and “discovered”.

Your endless energy charges me to be the mom you deserve.
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................................. iv

ACKNOWLEDGMENTS ............................................................................................................... v

LIST OF TABLES ....................................................................................................................... ix

LIST OF FIGURES ..................................................................................................................... x

LIST OF APPENDICES .............................................................................................................. xi

LIST OF ABBREVIATIONS ......................................................................................................... xii

DECLARATION OF ACADEMIC ACHIEVEMENT ...................................................................... xiv

CHAPTER 1: INTRODUCTION .................................................................................................. 1

CHAPTER 2: A review of methodological and practical challenges ....................................... 6
  Abstract ................................................................................................................................. 8
  What is new? ......................................................................................................................... 9
  Background............................................................................................................................ 10
  Methods ................................................................................................................................. 11
  Applications and roles of tests in health care ...................................................................... 13
  Diagnostic research designs ............................................................................................... 15
  Challenges in conducting and applying diagnostic research ........................................... 21
  Conclusion ........................................................................................................................... 27

CHAPTER 3: User testing of GRADE evidence tables ............................................................. 40
  Abstract ............................................................................................................................... 43
  What is new? ....................................................................................................................... 45
  Background .......................................................................................................................... 46
  Methods ................................................................................................................................. 48
  Results .................................................................................................................................. 52
  Discussion ............................................................................................................................. 59

Chapter 4: A systematic review of available instruments to assess the quality of evidence and
  strength of recommendations ............................................................................................ 164
  Abstract ............................................................................................................................... 166
  What is new? ....................................................................................................................... 168
  Background .......................................................................................................................... 169
  Methods ................................................................................................................................. 171
  Results .................................................................................................................................. 174
  Discussion ............................................................................................................................. 178

CHAPTER 5: Assessing international guidelines ..................................................................... 272
  Abstract ............................................................................................................................... 274
  What is new? ....................................................................................................................... 275
  Background .......................................................................................................................... 276
  Methods ................................................................................................................................. 277

vii
| Results | .................................................................................................................. | 281 |
| Discussion | ........................................................................................................ | 285 |
| CHAPTER 6: what do experts say? | ........................................................................................................... | 316 |
| Abstract | ........................................................................................................ | 318 |
| What is new | ....................................................................................................... | 320 |
| Background | ....................................................................................................... | 321 |
| Method | .......................................................................................................... | 322 |
| Results | .......................................................................................................... | 327 |
| Discussion | ....................................................................................................... | 335 |
| CHAPTER 7: CONCLUSIONS | ................................................................................................ | 347 |
| Conclusions and Recommendations | .................................................................................... | 348 |
| Products informed by this thesis | ........................................................................................ | 350 |
| Future direction | .................................................................................................... | 361 |
LIST OF TABLES

Table 2.1. Summary of HCTDS applications and their definitions ..................................32
Table 2.2. Roles of HCTDS, their definitions, and examples for each ..............................33
Table 2.3. Different diagnostic RCT designs, their prerequisites, and what questions can and cannot be answered by each of these designs ........................................................................35
Table 3.1: DTA SoF table based on test results (TP, FP and TN, FN) ..............................66
Table 3.2: DTA SoF table based on disease status (TP, FN and TN, FP) .........................67
Table 3.3: DTA SoF table based on pre- and post-test probability ..................................68
Table 4.1: Tools summary with domains in rating QoE and SoR ......................................183
Table 4.2: Main themes of the modification of existing tools to assess QoE and SoR for HCTDS ........................................................................................................................................195
Table 5.1. Summary of results based on AGREE II domains ........................................293
Table 5.2. Linking diagnostic test accuracy to patient-important outcomes .................296
Table 6.1: Outline of topics and subtopics ........................................................................341
Table 6.2: Test accuracy results may be sufficient to extrapolate about overall benefits and harms .......................................................................................................................342
Table 6.3: Essential factors in making decisions about HCTDS .................................343
LIST OF FIGURES

Figure 3.1. Outline of the rounds of feedback and user testing to develop GRADE diagnostic summary tables ................................................................. 69
Figure 3.2. Summary of the domains used for data analysis of user testing and feedback ......................................................................................... 69
Figure 4.1: PRISMA Flow Diagram of Studies ......................................................... 182
LIST OF APPENDICES

Appendix 3.1: Workshop package Grading the quality of evidence and preparing Summary of Findings tables about diagnostic tests .................................................. 70
Appendix 3.2.1a: Interview Package 6A – Full 30-Min Interview .................................. 91
Appendix 3.2.1b: Galactomannan ELISA for the diagnosis of invasive aspergillosis 102
Appendix 3.2.2A: Interview Package 4A – Full Interview ........................................... 109
Appendix 3.2.2b: Enhancing the usability and usefulness of Summary of Findings tables for decision making about diagnostic tests – Tables for Evaluation ....... 136
Appendix 3.3: Evidence Profile with Individual DTA Outcomes .................................. 155
Appendix 3.4 .................................................................................................................. 158
Appendix 4.1: Search strategies for published literature on evidence grading systems .................................................................................................................. 196
Appendix 4.2: Quality Rating and Evidence Grading Systems Identified in AHRQ and CADTH Reports ........................................................................................................... 200
Appendix 4.3: List of tools included with references per tool ...................................... 210
Appendix 4.4: Excluded studies with reasons for exclusion and references ............... 211
Appendix 5.1: Guidelines Data Abstraction Form ......................................................... 297
Appendix 5.2: Overview of guidelines we reviewed on healthcare related tests and their applications .................................................................................................................. 303
Appendix 5.3: Details about Guidelines making recommendations about the same test for the same condition: comparison of guideline methods and recommendations ....................................................................................................... 310
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AACC</td>
<td>American Association for Clinical Chemistry</td>
</tr>
<tr>
<td>AACE</td>
<td>American Association of Clinical Endocrinologists</td>
</tr>
<tr>
<td>AAN</td>
<td>American Academy of Neurology</td>
</tr>
<tr>
<td>AAO-HNS</td>
<td>American Academy of Otolaryngology - Head and Neck Surgery</td>
</tr>
<tr>
<td>AAP</td>
<td>American Academy of Pediatrics</td>
</tr>
<tr>
<td>AASM</td>
<td>American Academy of Sleep Medicine</td>
</tr>
<tr>
<td>ACCP</td>
<td>American College of Chest Physicians</td>
</tr>
<tr>
<td>ACP</td>
<td>American College of Physicians</td>
</tr>
<tr>
<td>ACR</td>
<td>American College of Radiology</td>
</tr>
<tr>
<td>AGREE</td>
<td>Appraisal of Guidelines for Research and Evaluation</td>
</tr>
<tr>
<td>AHA</td>
<td>American Heart Association</td>
</tr>
<tr>
<td>AHRQ</td>
<td>Agency for Health Care Research and Quality</td>
</tr>
<tr>
<td>AWMF</td>
<td>Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften</td>
</tr>
<tr>
<td>CADTH</td>
<td>Canadian Agency for Drugs and Technologies in Health</td>
</tr>
<tr>
<td>CBO</td>
<td>Dutch Institute for Healthcare Improvement (Centraal Begeleidings Orgaan)</td>
</tr>
<tr>
<td>CCO</td>
<td>Cancer Care Ontario</td>
</tr>
<tr>
<td>CCOPEBC</td>
<td>Cancer Care Ontario’s Program in Evidence-Based Care</td>
</tr>
<tr>
<td>CDA</td>
<td>Canadian Diabetes Association</td>
</tr>
<tr>
<td>CeVEAS</td>
<td>Centro per la valutazione della efficacia della assistenza sanitaria</td>
</tr>
<tr>
<td>CHERG</td>
<td>Child Health Epidemiology Reference Group</td>
</tr>
<tr>
<td>CMA Infobase</td>
<td>Canadian Medical Association Infobase</td>
</tr>
<tr>
<td>COMPUS</td>
<td>Canadian Optimal Medication Prescribing and Utilization Services</td>
</tr>
<tr>
<td>CONSORT</td>
<td>Consolidated Standards of Reporting Trials</td>
</tr>
<tr>
<td>CPG</td>
<td>Clinical Practice Guideline</td>
</tr>
<tr>
<td>CTFPHC</td>
<td>Canadian Task Force on Preventive Health Care</td>
</tr>
<tr>
<td>DTA</td>
<td>Diagnostic test accuracy</td>
</tr>
<tr>
<td>Dx</td>
<td>Diagnosis</td>
</tr>
<tr>
<td>EFNS</td>
<td>European Federation of Neurological Societies Tool</td>
</tr>
<tr>
<td>EGAPP</td>
<td>Evaluation of Genomic Applications in Practice and Prevention</td>
</tr>
<tr>
<td>ESC</td>
<td>European Society of Cardiology</td>
</tr>
<tr>
<td>EULAR</td>
<td>European League Against Rheumatism</td>
</tr>
<tr>
<td>EvP</td>
<td>Evidence Profile</td>
</tr>
<tr>
<td>EWG</td>
<td>EGAPP Working Group</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>GIN</td>
<td>Guideline International Network</td>
</tr>
<tr>
<td>GISIG</td>
<td>Italian Study Group on Severe Infections</td>
</tr>
<tr>
<td>GoR</td>
<td>Grade of Recommendation</td>
</tr>
<tr>
<td>GRADE</td>
<td>Grading of Recommendations Assessment, Development and Evaluation</td>
</tr>
<tr>
<td>HCTDS</td>
<td>Healthcare related tests and diagnostic strategies</td>
</tr>
</tbody>
</table>
IKNL  Comprehensive Cancer Centre the Netherlands (Integraal Kankercentrum Nederland)
IQWiG  German Institute for Quality and Efficiency in Healthcare
JRGCSG  Japanese Research Group for the Development of Cancer Screening Guidelines
LEGEND  Let Evidence Guide Every New Decision
LiST  Lives Saved Tool
MERGE  Method for Evaluating Research and Guideline Evidence
MeSH  Medical Subject Headings
NACB  National Academy of Clinical Biochemistry
NASS  North American Spine Society
NGC  National Guideline Clearinghouse
NHMRC  Australian National Health and Medical Research Council
NICE  National Institute for Health and Clinical Excellence
NOQAS  Newcastle–Ottawa Quality Assessment Scale
PNLG  Programma nazionale per Le linne guida
QoE  Quality of Evidence
QUADAS  Quality Assessment of Diagnostic Accuracy Studies
RCOG  Royal College of Obstetricians and Gynaecologists
RCR  Royal College of Radiologists
RCT  Randomized Controlled Trial
SIGN  Scottish Intercollegiate Guidelines Network
SoF  Summary of Findings
SoR  Strength of Recommendation
SORT  Strength of Recommendation Taxonomy
STARD  Standards for Reporting of Diagnostic Accuracy
TA  Test Accuracy
TN  True Negative
TP  True Positive
USPSTF  U.S. Preventive Services Task Force
WHO  World Health Organization
DECLARATION OF ACADEMIC ACHIEVEMENT

I was the main contributor and first author for all studies.
CHAPTER 1: INTRODUCTION
INTRODUCTION

While therapeutic interventions are increasingly subject to critical evaluation of a) the evidence supporting such interventions and b) decision-making processes to recommend interventions, such approaches and requirements either do not exist for healthcare related tests and diagnostic strategies (HCTDS) or there is a lack of clear regulation on how tests and strategies should be used. In this thesis I provide a systematic evaluation of the available published and grey literature as well as a review of methods used by various organizations. Additionally, based on our findings I conclude by suggesting a framework to support decision makers including guideline panels when making decisions about HCTDS.

This thesis is based on a series of articles submitted for publications, addressing issues related to decision-making about healthcare related tests and diagnostic strategies. This thesis consists of the following articles and corresponding chapters:

**Article 1 (Chapter 2). Decision-making about healthcare related tests and diagnostic strategies: A review of methodological and practical challenges and introduction to a new series**

In this article we provide frameworks for the roles and applications of a HCTDS. Also, we summarise the challenges unique to the field of diagnosis. Then, we discuss different study designs including diagnostic randomized controlled trials (RCT) and retrospective validation. We specifically make the utility of currently suggested designs clear (which questions can they answer and which they cannot).
Article 2 (Chapter 3). Decision-making about healthcare related tests and diagnostic strategies: User testing of GRADE evidence tables

In this article we summarise the user testing and use of evidence tables for HCTDS over the last 12 years. We have conducted user testing and interviews in a variety of settings (workshops, meetings) with a variety of end users (authors of systematic reviews, guideline developers and many others). The current suggested format of diagnostic evidence tables provides a solution that, with some training, is an easy to use presentation of complex test accuracy data and quality of evidence.

Article 3 (Chapter 4). Decision-making about healthcare related tests and diagnostic strategies: A systematic review of available instruments to assess the quality of evidence and strength of recommendations

In this article we present our findings from a comprehensive systematic review to identify and summarise all available instruments, checklists, critical appraisal tools, and indices published in the literature that are designed for assessing the quality of evidence (QoE) and strength of recommendations (SoR) dealing with HCTDS. We identified 43 tools and modifications of existing tools to assess the QoE and SoR. Most of the tools reviewed acknowledge the importance of assessing the QoE and SoR separately. When moving from evidence to recommendations only few tools require the consideration of patient values and preferences and resources. There is confusion about the terminology that describes the various factors that influence the QoE and the criteria for moving from evidence to recommendations are incomplete for most guideline development frameworks that we
evaluated.

**Article 4 (Chapter 5). Decision-making about healthcare related tests and diagnostic strategies: Assessing international guidelines**

In this article we summarise our findings from reviewing a sample of 37 international clinical practice and public health guidelines about HCTDS. We followed a comprehensive search strategy and consulted international experts to identify these guidelines. We provide an assessment of their processes, methods and adherence to credibility criteria.

**Article 5 (Chapter 6). Decision-making about healthcare related tests and diagnostic strategies: what do experts say?**

In this article we describe and summarise our findings from interviewing experts in the field of diagnosis and guideline development. We conducted in-depth semi structured interviews with 24 international experts and key informants in this area. The findings are remarkable. Diagnostic test accuracy was the factor most commonly considered by organizations when formulating recommendations. However, most experts quickly pointed out that accuracy alone is rarely if ever sufficient and that recommendations and decisions based on accuracy alone may be misleading. Experts identified additional factors as essential in making decisions about diagnostic tests, which we outline in this paper.

**Chapter 7. Thesis conclusion, recommendations and future direction**

In this section I summarise the appropriate steps when moving from evidence to recommendations about HCTDS. I describe the process by which treatment and diagnostic
recommendations and coverage decisions are developed, which are relatively similar. I provide detailed suggestions about how to determine if a test is new. Then I summarise our suggested strategy into 7 major recommendations and I present an algorithm that can be used to systematically support coverage decision, regulatory approval or health care recommendations addressing HCTDS.

I also highlight additional products based on our work and future direction and projects of this research program.
CHAPTER 2: A REVIEW OF METHODOLOGICAL AND PRACTICAL CHALLENGES
Decision-making about healthcare related tests and diagnostic strategies: A review of methodological and practical challenges and introduction to a new series

Reem A. Mustafa, MD, MPH 1,2, Wojtek Wiercioch, B.H.Sc1, Adrienne Cheung, B.H.Sc3, Barbara Prediger, B.Sc.4, Jan Brozek, MD1,5, Patrick Bossuyt, MSc, PhD6, Amit X. Garg, MD, PhD1,7, Monika Lelgemann, MD, MSc8, Diedrich Büehler, MD9, Holger Schünemann, MD, MSc, PhD1,5.

(1) Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada

(2) Department of Internal Medicine, University of Missouri-Kansas City, Kansas City, USA

(3) Faculty of Medicine, University of British Colombia, Vancouver, Canada

(4) Center for Medical Biometry and Medical Informatics, University of Freiburg, Germany

(5) Department of Medicine, McMaster University, Hamilton, Canada

(6) Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherland

(7) Department of Medicine, Western University, London, Canada

(8) Medizinischer Dienst des Spitzenverbandes Bund der Kranken-kassen e.V. (MDS) Theodor Althoff-Str. 47 45133 Essen, Germany

(9) Abteilung Medizin. GKV—Spitzenverband Mittelstraße 51 10117 Berlin, Germany
Objective

To provide an overview of the literature about the application of healthcare related tests and diagnostic strategies (HCTDS) and the decision making process that guides the use of these tests with our interpretation and suggested frameworks.

Design and setting

We conducted a review by searching MEDLINE, the references of identified articles, related chapters in relevant textbooks and forward search for articles citing classic papers on this topic.

Results

In this paper we provide frameworks for the potential roles and applications of tests with suggested definitions and practical examples. We also discuss study designs that are commonly used to assess tests’ performance and the effects of tests on patients’ health. These designs include diagnostic randomized controlled trials and retrospective validation. We clarify the utility of currently suggested designs - which questions they can answer and which ones they cannot. In addition, we summarize the challenges unique to diagnostic decision-making.

Conclusion

This article lays the background to a series of eight papers on diagnostic decision-making in health care. In this series we attempt to shed more light on some of the challenges related to decision-making about tests and propose possible solutions.
WHAT IS NEW?

- We identify potential roles and applications of healthcare related tests and diagnostic strategies based on extensive literature search.

- We discuss the advantages and disadvantages of different study designs that can be used when assessing HCTDS.

- We summarize a variety of diagnostic randomized trial designs that can be employed to assess HCTDS. We clarify which questions can each of these designs answer and which questions they cannot address.

- We summarize challenges unique to diagnostic decision-making.
BACKGROUND

An accurate diagnosis is one of the first critical steps for successful management of a health care problem. In most instances, decision makers perform healthcare related tests and diagnostic strategies (HCTDS), which often are used for reasons beyond establishing an accurate diagnosis (i.e. establishing response to treatment, prognostication). Tests are used at different times through the care pathway, for example, to select the most appropriate next steps in an often complex decision making process\(^2\).

A larger number of HCTDS evaluations focus on their diagnostic accuracy. A test with high diagnostic accuracy (i.e. ability to correctly classify a patient as having or not having a disease/condition) may not necessarily be the test of choice in clinical practice or may not lead to improved health outcomes\(^3\). Currently available guidance about how best to consider evidence from diagnostic test accuracy (DTA) studies in decision-making is insufficient. This, at least in part, is related to the practical realities and challenges involved in using diagnostic evidence in decision-making.

In this first of a series of eight articles in the *Journal of Clinical Epidemiology* on diagnostic decision-making in health care, we review the roles and applications of tests in the literature. We also discuss different study designs, including diagnostic randomized controlled trials (RCT) and retrospective validation designs. Following this review, we describe the challenges encountered in conducting diagnostic research and in applying diagnostic research to health care decision-making. We conclude with an outline of the forthcoming articles that describe the research we conducted to suggest solutions to these challenges.
METHODS

Literature search strategy

To identify tests’ roles and applications, we conducted a comprehensive review. We searched MEDLINE using the PubMed Clinical Queries for a random sample of DTA studies published between 1998 and 2012 using the keyword, ‘test’ and limited the search in the category field to ‘Diagnosis’ and in the Scope field to ‘Broad’. We also limited our search to human studies that were published in English or French.

To identify challenges in diagnostic research, we searched the references of identified articles and publications for relevant research, including the Methods Guide for Medical Test Reviews prepared by the Agency for Healthcare Research and Quality, six systematic reviews of diagnostic test accuracy prepared by the Diagnostic Test Accuracy Working Group at the Cochrane Collaboration, as well as the Handbook for DTA Reviews prepared by the DTA Working Group at the Cochrane Collaboration. We conducted a forward search using Web of Science indexing for articles citing the classic papers on this topic. We also searched for papers authored or coauthored by known experts in this field. Additionally, we consulted relevant chapters in reference books in the areas of clinical epidemiology, evidence-based medicine and systematic reviews of DTA.

Study selection and data extraction

For identifying tests’ roles and applications, we randomly selected 50 studies from the list of references. From each study we extracted the test under investigation, the suspected
condition, and the reference test used. We then categorized the tests by their diagnostic application and role. We went through two iterations of this process until we reached theme saturation. For diagnostic research design and challenges in diagnostic research, we reviewed all the articles identified by our search strategy.
APPLICATIONS AND ROLES OF TESTS IN HEALTH CARE

Table 1 summarises our findings of all applications of tests. Tests are used for many different applications: screening or surveillance, risk assessment and classification, diagnosis (ruling in), ruling out diagnosis, treatment triage, treatment monitoring, staging and determining prognosis. In Table 1 we also provide definitions for each of the applications.

We identified studies that investigated whether or not a “new test” could be used as an appropriate replacement for the reference standard itself. We have labeled this role “replacement of reference standard”. However, despite the common use of the terminology “new test” in the scientific literature, our review indicated that no consistent definition on when to consider a test as “new” existed. We therefore developed a definition that encompasses the intended use by various authors; a “new test” or strategy is a test that has not previously been used in the intended role or for the intended purpose. Interest in this “new test” may exist because it is perceived to be faster in determining test results, require less expertise, be less invasive, improve accuracy or efficacy, require less resources, or be more feasible for various reasons.

Table 2 summarizes our findings of the roles of tests with their definitions. If a new test is conceived to perform better than the existing reference standard, it might be considered the new reference standard. However, a new reference standard can be developed only by convention because better accuracy than a reference standard usually cannot be scientifically demonstrated. The other scenarios relate to the comparison of an index test against an alternative test or test strategy. To make judgments about the comparative
accuracy a reference test is required against which the index and alternative test are compared.

Table 2 describes instances when the test of interest would be placed before an existing test (triage), after an existing test (add-on) or would replace the existing test (replacement)\(^2\). We identified uses of tests not covered by the above three situations. Parallel testing may occur when two perceived equally useful tests are combined and both would be done regardless of the results of the other to enhance the ability of each other (table 2). This may occur less frequently, because the addition of a concurrent test would have to justify the use of additional resources and may only be useful if the results of the new test are complementary to the results of the existing test and thus require their simultaneous application or interpretation if they have to be applied sequentially.
DIAGNOSTIC RESEARCH DESIGNS

1. Diagnostic randomized controlled trial (RCT)

The diagnostic RCT is considered the optimal study design to minimize bias in HCTDS evaluation. Box 1 summarises advantages of RCT over other methods of comparing tests.

<table>
<thead>
<tr>
<th>Box 1: Advantages and disadvantages of RCTs over other methods of comparing healthcare related tests and diagnostic strategies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Advantages include:</strong></td>
</tr>
<tr>
<td>1. Proper randomization should prevent bias in the allocation of patients to the index and reference test</td>
</tr>
<tr>
<td>2. Possibility of applying experimental statistical designs such as testing for significance and calculating confidence intervals directly for patient important outcomes (i.e. without modeling)</td>
</tr>
<tr>
<td>3. Trial design can closely mimic existing diagnostic and management questions</td>
</tr>
<tr>
<td><strong>Disadvantages include:</strong></td>
</tr>
<tr>
<td>1. Trials can be logistically more complex and should be used when there is true equipoise for a test or test-treatment strategy (^1)</td>
</tr>
<tr>
<td>2. Trials are likely to require more resources compared to other designs.</td>
</tr>
</tbody>
</table>

To maximize the confidence in the results of diagnostic RCT designs, the following considerations are essential:
• A pre-specified link between test results and management/treatment decisions should be provided.

• If high quality evidence about management/treatment efficacy exists then patients who are classified as having the disease and those who do not have the disease will undergo different management/treatment strategies.

• Detailed information about the management/treatment options is necessary to help other groups replicate the findings and implement the suggested strategy.

• Investigators should make a careful decision about at what point in the management scheme should randomization be done. Usually the point of randomization coincides with the clinical decision whether or not to perform the test or which test to use.

• Basic methodological considerations to reduce risk of bias like blinding and allocation concealment are essential in all RCTs including diagnostic trials.

• Investigators should pay extra attention to methodological issues that are specific to tests. For example, physician’s knowledge about which test has been performed may influence clinical management and patient outcomes which may affect the results of the trial.

• Different designs may be more efficient to answer specific questions. Investigators should aim for the most efficient design.

In this section we summarize diagnostic RCT designs that evaluate a single test and compare HCTDS in case of a replacement, add-on, or triage tests. We also discuss trials for predictive test validation. Then, we discuss other designs including before and after design and case-control and two gate designs.
1.1. **Diagnostic RCTs evaluating a single test**\(^{1,231,221,221,221,221,222}\)

Table 3 summarises our findings of different diagnostic RCT designs, their prerequisites, and what questions can and cannot be answered by each of these designs for single test and compared HCTDS in case of replacement. In summary we identified the following designs: a. Focus: Prognostic value of a test, b. Testing at baseline in RCTs, c. A test result as an RCT inclusion criterion, d. Random disclosure design, e. RCT of testing (A test based strategy design).\(^{21,22,24}\)

1.2. **Diagnostic RCTs comparing tests and test strategies in the case of a Replacement test**

In summary we identified the following RCT designs in this category: f. Comparing the prognostic value of two tests, g. Comparing predictive value, h. Discordant test results RCT, i. Random disclosure RCT, j. RCT comparing different tests. For details about these designs refer to table 3.

1.3. **Diagnostic RCTs comparing tests and test strategies in the case of an add-on or a triage test:**

Many new tests are intended to be added to the management strategies rather than replace an older test. These new tests can be added after or before an older test is done (triage or add-on). If the test is being used as an add-on test it may be considered as the final stage in the diagnostic pathway and hence the designs that apply to single tests may apply here. If the test is being used as a triage test randomized designs can be utilized to evaluate the effectiveness of this new test in the care pathway.
1.4. **Clinical trials for predictive test validation (predictive marker validation)**

A predictive test is a test that separates a population with respect to the outcome of interest in response to a particular target therapy.\textsuperscript{24,25} These tests are being established to predict patient important outcomes to specific management. They are being used increasingly in the field of biomarkers, in particular testing related to oncological disease. With validation we refer to confirming how well the test can predict the outcome. Validation is not a yes or no answer, it rather requires a variety of approaches and processes to increased one’s confidence in this prediction. The validation of these predictive tests can be done by employing one of the RCT designs that we discussed earlier specifically design b and c. Additionally, a retrospective validation of these tests may be possible. However, retrospective validation analysis is conceptually similar to subgroup analysis in treatment efficacy studies and should be dealt with cautiously. It should only be done if all the requirements for a valid retrospective assessment of a predictive test are met. These requirements include: 1. Data from well-conducted RCT are available; 2. Availability of samples on a majority of patients to avoid selection bias; 3. A priori stated hypothesis and analysis plan; 4. A priori defined assay and scoring system; 5. A priori justification of sample size requirement.

A classic example of this type of validation is the use of previously collected RCT data to retrospectively validate KRAS gene mutations as a predictor of efficacy of monoclonal antibodies panitumumab and cetuximab in advanced colorectal cancer\textsuperscript{26-28}. It is unclear why one would choose a retrospective validation rather than a prospective randomized design given that the investigators have stated the hypothesis and defined the assay a priori.
These test validation designs have also been used to validate specific tests or scoring systems based on genetic profiling like Oncotype DX. Complex scoring systems on the basis of genetic profiling present even more challenging problems because they require validation and methodological evaluation of the scoring system (typically a selection of a genetic expression pattern of several candidate genes) itself.

2. Before-after study design

The before-after study design examines changes to clinicians’ patient management plans that occur in response to a test result. A drawback of this design is that it is only appropriate for measuring the added accuracy or efficacy of an add-on test. The main criticism of this design is that clinicians’ reported plans during a research study may differ from their decision-making in an actual scenario, and so implementation of the test may not produce the same results achieved in trials.

3. Case-control and two-gate design

Some investigators use case-control study designs in an attempt to determine DTA of a test. While it is generally accepted that this design cannot be used to determine effects on patient important outcomes, the design has been criticized specifically for producing inflated estimates of test accuracy, sometimes two to three times higher than estimates derived from evaluations with a single series of patients. The difference is in part attributed to spectrum bias since cases and controls are selected with different sampling schemes (“gates”). The design can, however, be less expensive and easier to conduct. Some experts in this field maintain that the design can be used without bias if appropriate
sampling is applied \(^{33}\), but others, and we tend to agree, consider this design entirely inappropriate for the reasons mentioned above.

4. **Classic cross sectional diagnostic test accuracy (DTA) design**

The classic DTA study design determines the accuracy of a HCTDS or compares the accuracy of more than one test by establishing the test(s) accuracy compared to an acceptable reference standard that should be applied to the same population. This design is frequently used in the literature. Although it could be perceived as an easy design, it requires clear understanding of methodological issues in DTA studies to avoid fatal flaws that may affect the validity of the study and its findings. An established standard on how to evaluate these studies and how to report them is available\(^{34,35}\).
CHALLENGES IN CONDUCTING AND APPLYING DIAGNOSTIC RESEARCH

1. Inconsistent use of terminology

This issue is a major limitation that hinders facilitating communication or implementing the findings of diagnostic research done by different disciplines. There is wide range of terms that are used in this area without clear agreement on their meanings. Additionally, the same term may mean different things in different disciplines. There is a clear need for standardization of terminology and definitions that can be referred to by different discipline.

2. Poor communication among disciplines involved

Conducting diagnostic research, assessing the evidence and implementing the evidence are the necessary steps in improving patient outcomes through diagnostic techniques. Each step involves different decision-making parties and communication between these parties is also necessary to ensure that decisions are evidence-based. Problems in communication may result in research findings remaining unknown outside of the scientific community and therefore not impacting policy and societal decisions, or economic assessments being conducted in inappropriate areas since cost-analysis experts are not as aware of areas of needs as the scientific or medical community may be.

3. Lack of regulatory standards

Regulatory standards for HCTDS are often lacking in comparison to those that are established for pharmacological interventions. Although the lack of standards is
particularly evident for tests targeting diseases that primarily affect the poor and disadvantaged the issue of inflated or false claims of test performance likely exists in high income countries due to lack of regulatory oversight on the design and conduct of diagnostic studies 37.

4. **Fast technology development**

Diagnostic technologies, particularly in the field of radiology and biomarkers, has developed dramatically and increasingly quickly, outpacing the rate at which thorough evaluations of efficacy take place 15,38. Although some investigators argue that research cannot keep up with the distribution of technologies and introduction of new applications for their use, this should not be used as an excuse to justify a lack of regulatory standards. The lack of understanding of how to timely and efficiently test the central principles results in justified suspicion among policymakers and payers of health care that expensive technologies are not being used cost-effectively. This presents a further challenge to researchers and policymakers since the latest technology should not be approved for use without thorough evaluation of its safety and efficacy.

5. **Problems with reference standards causing high risk of bias**

Evaluating diagnostic accuracy requires a comparison to an “independently established standard diagnosis” 15. In some instances, an accurate reference standard may not exist or the standard diagnostic procedure may be too invasive for routine use in a study. Reitsma et al. 39 suggest solutions to handling issues related to imperfect or unavailable reference standard. These solutions include constructing data for a reference standard from clinical assessment data using panel judgment, using composite reference
standard, imputing or adjusting for the incomplete verification and using latent class analysis.

6. **Tests are usually used within diagnostic strategies**

The use of tests is more complex than the simplistic set-up of a single test used to assess the presence or absence of a target condition. This simplistic set-up is frequently applied in DTA research and education because of the challenges encountered in study design or education about diagnostic tests. In practice, however, tests are seldom used in isolation but rather as a component of a protocol of management strategies. To evaluate testing strategies, studies should aim to account for other factors in the diagnostic pathway such as timing of tests and their use in combination with results of other tests.

7. **Measuring effects of HCTDS on patient important outcomes**

Diagnostic research should be conducted with the aim of providing net benefit for patients or the population at large (including improving patient or population important outcomes, resource utilization, enhancing accessibility or reducing testing related complications). The relation between the use of HCTDS and important outcomes is often indirect because an intervention usually follows the use of a test and only rarely can immediate benefits be considered in isolation. For example, a patients’ reassurance about absence of a disease, a psychological immediate benefit, is unlikely to be the last step in a patient’s management as the reasons for testing do not disappear after reassurance and further testing and management is likely required. Thus, using a test based on diagnostic accuracy evidence alone may be harmful to patients because a direct link from the test result to improved health has not been made.
Lee et al. suggest a framework for evaluating the potential medical and nonmedical value of testing that considers three dimensions of value: medical, planning and psychological \(^{43}\). Medical value considers a test’s impact on treatment decisions; planning value reflects how test results can better allow patients to make informed decisions on work, finance, long-term plans and other life decisions; and psychological value refers to a patient’s sense of self-satisfaction and can have positive or negative effects (e.g. ruling out a health concern or delivering bad news). Consideration of planning and psychological value can change perceptions of test benefit and have implications in health policy and insurance design by influencing the conclusions of cost-effectiveness analyses. Because of confusion about the terminology of “values”, i.e. how this term is used in decision and cost-utility analyses, we describe outcomes as desirable and undesirable consequences as they relate to patient important outcomes and the use of the term as an expression of cost. Desirable consequences include direct health outcomes, such as a mortality or morbidity reduction, psychological well-being, reduced burden and lower cost. Undesirable consequences represent the opposite. Given the indirect relation between performing a test and changes in important health outcomes, a large sample size in diagnostic studies is usually required in order to detect - with precision - effects on downstream patient important outcomes \(^{40}\). Studies with large sample sizes may be resource-intensive and logistically challenging.

8. **Defining key outcomes**

Defining key outcomes amongst a multitude of possible endpoints and determining when evidence is sufficient is challenging to researchers and health care decision makers (including policy makers, guideline developers and clinicians). Gazelle et al.
propose that the size of the at-risk population, the anticipated clinical benefits, and the potential economic impact should be weighed to determine the level of outcomes data required, based on Fryback and Thornbury's hierarchy of diagnostic imaging efficacy \(^3,4^4\). This hierarchical model consists of six levels of classification for diagnostic assessment studies. The continuum begins with technical efficacy of the technology at level 1, then considers measures of diagnostic accuracy at level 2, level 3 reflects the effect of test results on clinician thinking, level 4 evaluates effect of test results on clinicians' choice of therapy, level 5 looks at patient outcomes and level 6 has the broadest scope and considers effects on society. Different frameworks exist to help guide the level of diagnostic studies and outcomes required for healthcare decision making. According to the framework developed by Gazelle et al., technologies that affect a small population, have large anticipated clinical benefit, and have low cost to payers and society or high value for money should require less extensive outcomes data (e.g. level 1 or 2 data). Conversely, technologies that affect a large population, have a small anticipated clinical benefit, and have high cost to payers and society or have low value for money should be held to evidence on more extensive outcomes data (e.g. level 5 or 6 data). Although this framework is proposed as useful for determining the level of outcomes data necessary to make decisions regarding resource allocation, insurance coverage and research study design, it has limitations which may hinder its applicability. A potential concern with this framework is that the relative importance of the three factors will be based on the judgment of those assessing the benefits and harms of the test. Additionally, it requires a high degree of certainty, which is difficult to achieve, about the anticipated consequences. Finally, weighing the importance of these
three factors can also become complex, and those involved may benefit from structured information aids regarding costs, clinical benefit and the population affected to inform their decisions.

9. Availability of complex diagnostic research designs

As discussed earlier, efficient diagnostic research designs have emerged and implemented that allow for limiting the potential for risk of bias. However, some of these designs are sophisticated and if not implemented correctly may lead to inaccurate conclusions about the potential benefits and harms from using a specific test. In this paper we attempted to describe the key features of these diagnostic research designs and demystified their complexity.

10. Methodological issues in DTA evidence synthesis

Systematic reviews and meta-analyses contribute to more accurate performance indications of tests. However, poor reporting in DTA studies impedes the quality of the evidence. Initiatives like the Standards for Reporting of Diagnostic Accuracy (STARD) project have led to significant progress in encouraging accurate and complete reporting in studies of DTA as well as evaluating potential for bias and generalizability of results. Additionally, there is a great deal of literature addressing the statistical complexity of systematic reviews of DTA and the lack of user-friendly programs for pooling DTA effect estimates.
CONCLUSION

In this first article in a series of eight articles on diagnostic decision-making in health care, we provide an overview of known tests’ roles and applications, different diagnostic study designs and the main challenges encountered in conducting diagnostic research and in applying it to health care decision-making. We did not address issues regarding the different applications of HCTDS in depth in this paper but we will discuss some of these issues in more details in following papers. In this series we use a combination of literature search of the evidence and its applications, formal user testing, survey and comprehensive interviews with experts using a mixed methods approach as we attempt to shed more light on some of the challenges and propose possible solutions.

Following this paper that provides an overview of the methodological and practical challenges encountered in the field of diagnosis, we will address the following:

Article 2. Decision-making about healthcare related tests and diagnostic strategies: User testing and feedback about grading the quality of evidence and preparing evidence tables

Article 3. Decision-making about healthcare related tests and diagnostic strategies: A systematic review of available instruments to assess the quality of evidence and strength of recommendations

Article 4. Decision-making about healthcare related tests and diagnostic strategies: Assessing state-of-the-art international guidelines and health technology

Article 5. Decision-making about healthcare related tests and diagnostic strategies: what do experts say?

Article 6. Decision-making about healthcare related tests and diagnostic strategies: A
framework for when is diagnostic research sufficient and when is it not

Article 7. Decision-making about healthcare related tests and diagnostic strategies: A validation study of the factors and for when diagnostic research is sufficient and when it is not

Article 8. Decision-making about healthcare related tests and diagnostic strategies: presenting benefits and harms information about HCTDS to decision maker.

Acknowledgment

This work was partially funded by the German Insurance Fund agency as part of a larger project about decision-making for HCTDS. The views presented here are those of the authors and should not be attributed to the funding agency or its staff.
References:

16. Leeflang M. *Systematic Reviews of Diagnostic Test Accuracy.* Amsterdam 2008.


### TABLE 2.1. SUMMARY OF HCTDS APPLICATIONS AND THEIR DEFINITIONS

<table>
<thead>
<tr>
<th>Applications of HCTDS</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Screening or surveillance</td>
<td>Monitor general population or a high risk group for an early detection of a disease/condition</td>
</tr>
<tr>
<td>Risk assessment and classification</td>
<td>Determine pre-test probability, existence of specific risk groups (e.g. high and low risk) and the need for close monitoring for a disease/condition</td>
</tr>
<tr>
<td>Diagnosis (ruling in)</td>
<td>Confirm presence of a disease/condition</td>
</tr>
<tr>
<td>Ruling out disease/condition</td>
<td>Exclude presence of a disease/condition</td>
</tr>
<tr>
<td>Treatment triage</td>
<td>Determine appropriateness of starting a treatment or type of treatment</td>
</tr>
<tr>
<td>Treatment monitoring</td>
<td>Follow-up for regression of a disease, possible recurrence, or appropriateness of continuing treatment during and/or post-treatment</td>
</tr>
<tr>
<td>Grading and staging</td>
<td>Determine severity of disease or phase of disease progression</td>
</tr>
<tr>
<td>Determining prognosis</td>
<td>A prediction of the probable course and outcome of a disease/condition</td>
</tr>
</tbody>
</table>
### TABLE 2.2. ROLES OF HCTDS, THEIR DEFINITIONS, AND EXAMPLES FOR EACH

<table>
<thead>
<tr>
<th>Role</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Replacement of a reference standard</td>
<td>Accuracy of a “new test”* is determined against the reference standard. If the new test performs at least as well as the reference standard, it can be considered for use in settings/populations where it is more appropriate, it can replace the reference standard or become a new reference standard.</td>
<td>Using polymerase chain reaction methods to identify difficult-to-culture bacteria instead of regular cultures.</td>
</tr>
<tr>
<td>Triage</td>
<td>A “new test” positioned before the existing test or testing pathway – only patients with a particular result continue the testing pathway. The results of the test strategy are compared against the reference standard or an alternative strategy (which may be the existing test alone). Usually done because the result of the triage test allows avoiding further testing on everyone suspected of the condition because the existing test has resource implications or side effects. A diagnosis on the basis of the triage test alone would not be made.</td>
<td>Screening all women for cervical intraepithelial neoplasia (CIN), which is a pre cervical cancer lesion, with a human papilloma virus (HPV) test, only those who have a positive HPV test will receive visual inspection of the cervix with possibly biopsies to confirm the diagnosis of CIN. Women with a negative HPV test will be saved from the visual inspection and the invasive biopsies.</td>
</tr>
<tr>
<td>Comparison against an existing test or test strategy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replacement</td>
<td>A “new test”, intended to replace an existing test, is compared with the existing test and results are verified against the same reference standard. If the results of a single test are compared against the reference standard, they would be considered under “single comparison against reference standard”.</td>
<td>For patients with suspected spinal cord injury, magnetic resonance imaging (MRI) has replaced myelography in most centres. This is because MRI is simpler, safer, and does not require exposure to radiation. The reference standard in this case is clinical findings on follow-up.</td>
</tr>
</tbody>
</table>

\* New test refers to a test that is not currently recognized as a reference standard.
| Add-on | A “new test” positioned after the existing or in addition to the existing test or test strategy. The results need to be compared against the reference standard or an alternative strategy. | To stage patients with cancer; only those with negative computed tomography (CT) scan for metastasis receive a positron emission tomography (PET). This will save resources, as PET scan is expensive and not widely available in all centers. |
| Parallel or combined | A “new test” intended to be used concurrently with an existing test. Both the existing and the new test are combined to make a diagnosis. | For patients with chest pain suggestive of myocardial infarction, They get an electrocardiogram (ECG) and troponin levels checked regardless of the results of each test. The results of both tests (ECG and troponin) are considered to determine next steps in management. |

* “New test” or strategy is a test that has not previously been used in the intended role or for the intended purpose and is not necessarily a new test in the market.
### TABLE 2.3. DIFFERENT DIAGNOSTIC RCT DESIGNS, THEIR PREREQUISITES, AND WHAT QUESTIONS CAN AND CANNOT BE ANSWERED BY EACH OF THESE DESIGNS.

<table>
<thead>
<tr>
<th>Study Design</th>
<th>Prerequisites and Conditions</th>
<th>Questions which can or can not be answered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Evaluation of a single test</strong></td>
<td>• Uncertainty about available treatment</td>
<td>Questions that can be answered</td>
</tr>
<tr>
<td><strong>a. Testing at baseline in RCTs (unselected or all-comers design, sequential testing strategy design, test based design, tests by treatment interaction design, test-based strategy design)</strong></td>
<td>• Uncertainty about the DTA of the test (all components)</td>
<td>• What is the prognostic value (accuracy) of a test within the context of subsequent decision-making and management?</td>
</tr>
<tr>
<td></td>
<td>• Interaction will help with defining subgroups to determine when will index test improve the prognostic value (accuracy)</td>
<td>• What is the overall prognostic value (accuracy) of a test?</td>
</tr>
<tr>
<td></td>
<td>• Decisions about the important subgroups should be made a priori</td>
<td>• What is the prognostic value (accuracy) of a test in each treatment arm separately?</td>
</tr>
<tr>
<td></td>
<td>• Results of test need to be concealed (unknown) at the time of randomization to treatment</td>
<td>• Which treatment is the most effective for all patients included in the trial?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Which treatment is more effective for groups with identical test results?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Questions that cannot be answered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• If use of index test provides net benefit</td>
</tr>
</tbody>
</table>

| **b. A test result as an RCT inclusion criterion (alternative names include targeted design, enriched design)** | • Certainty about DTA of a test (High quality DTA data)                                          | • What is the natural progression history of the people without the disease?                              |
|                                                                                                                                  | • Certainty about benign natural history in people without the disease                            | • Does treatment improve the outcomes of patients with the disease?                                       |
|                                                                                                                                  |                                                                                               | • What is the prognostic value (accuracy) of test positive within the context of                         |

*Figure: Study Design Diagram*
(Test-)
- Certainty about lack of treatment benefit in people without the disease
- Uncertainty about treatment effect in people with the disease (Test+)

subsequent decision-making?
- What is the prognostic value (accuracy) of the test in predicting the outcome of interest in the absence of treatment? (by comparing outcomes of patients in the placebo arm with those that tested negative only if the other management in both groups was similar. To calculate the prognostic accuracy in this design one has to adjust for the randomization ratio, keeping in mind that probably only about half of the test positive group would have been randomized to receiving placebo)

c. Random disclosure design

- Uncertainty about the added value of stratifying based on test results
- Uncertainty about treatment efficacy in test negative group

- What is the prognostic value (accuracy) of the test in predicting outcome of interest? (by comparing outcomes of patients in the test positive arm and test negative in the none disclosed arm)
- Does disclosure of the results affect patient important outcomes? (by comparing outcomes of patients in the test negative group in the disclosed arm with the test negative group in the non-disclosed arm)

d. RCT of testing (test-based strategy design)

- Standard of care involves not relying on test results for treatment.
- This design may be helpful if the routine management involves treating everyone without a test being used. Having a negative test may avoid receiving unnecessary treatments.
Evaluation of a replacement test

e. Comparing predictive value

- Uncertainty about available treatment
- Uncertainty about DTA of test A and test B in all results categories
- Interaction will help with defining subgroups to determine when each of the tests or their combination improves the prognostic value (accuracy).
- Decisions about subgroups should be made a priori.
- Results of both tests must remain undisclosed
- Very large sample size required to answer all of the posted questions

- What is the DTA of test A and test B within the context of subsequent decision-making?
- How does the DTA of test A and test B compare within the context of subsequent decision-making?
- What is the DTA of the combination of test A and test B within the context of subsequent decision-making?
- What is the overall DTA of test A, test B and their combination?
- What is the DTA of test A, test B and their combination in each treatment arm separately?
- Which treatment is the most effective for all patients included in the trial?
- Which treatment is more effective for groups in respective test results categories?

f. Discordant test results RCT

- Uncertainty about DTA of discordant test results (that is one test

- What is the natural progression history of the people without the disease?
- Does treatment improve the prognosis of
indicating disease is present and the other that disease is not present)
• Certainty about diagnosis when test results are concordant (i.e. DTA)
• Certainty about benign natural history in people without disease (Test A- and Test B-)
• Certainty about lack of treatment benefit in people without the disease
• Certainty about treatment effect in people with the disease (Test A+ and Test B+)

patients with the disease?
• What is the DTA of test A within the context of subsequent decision-making?
• What is the DTA of test B within the context of subsequent decision-making?
• What is the effect of the treatment (clinical pathway) based on test A only for patients with discordant test results?
• What is the effect of the treatment (clinical pathway) based on test B only for patients with discordant test results?
• What is the relative risk associated with each management strategy? (to calculate this one needs information about even rates in each of the concordant groups)
• What is the risk difference between the clinical pathways based on test A and the clinical pathway based on test B? (This is calculated based on the difference in poor outcomes between the groups)

**Random disclosure RCT**

• Certainty about the DTA of test A and test B when used in isolation
• Uncertainty about the added value from performing both tests together both for the concordant and the discordant test results
• Certainty about treatment efficacy in people with the disease
• What is the DTA of test A in predicting outcome of interest?
• What is the DTA of test B in predicting outcome of interest?
• What is the prognostic value of test A and test B together?
• Does disclosure of the results affect patient important outcomes? (by comparing outcomes of patients in the test negative group in the disclosed arm with the test negative group in the non-disclosed arm)
h. RCT comparing different tests

- Ideally also order of tests should be randomized
- Condition does not change during administration of tests

- Certainty about treatment effect in people with the disease
- Relative certainty about the DTA of test A and test B
- Uncertainty about the best-combined test-management strategy. This may relate to factors other than the accuracy of the test, e.g. how fast the test results can be obtained or how feasible a test is?

Questions that can be answered
- Which test-management strategy is more effective?

Questions that cannot be answered
- One is unable to distinguish the treatment effect from the prognostic or predictive value of the test. Additionally, one is unable to compare the outcome in the subgroups with discordant test results.
CHAPTER 3: USER TESTING OF GRADE EVIDENCE TABLES
Decision-making about healthcare related tests and diagnostic strategies: User testing of GRADE evidence tables

Reem A. Mustafa, MD, MPH, Wojtek Wiercioch, B.H.Sc, Jan Brozek, MD, Nancy Santesso, RD, MLIS, Adrienne Cheung, B.H.Sc, Barbara Prediger, B.Sc, Tejan Buleh, Alonso Carrasco-Labra, DDS, MSc, Romina Brignardello-Petersen, DDS, MSc, Ignacio Neumann, MD, MSc, Patrick Bossuyt, MSc, PhD, Amit X. Garg, MD, PhD, Monika Lelgemann, MD, MSc, Diedrich Bühler, MD, Holger Schünemann, MD, MSc, PhD (10)

Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada

Department of Internal Medicine, University of Missouri-Kansas City, Kansas City, USA

Department of Medicine, McMaster University, Hamilton, Ontario, Canada

Faculty of Medicine, University of British Colombia, Vancouver, British Columbia, Canada

Center for Medical Biometry and Medical Informatics, University of Freiburg, Germany

Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Chile

Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada

Department of Internal Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands

Department of Medicine, Western University, London, Ontario, Canada
(20) Medizinischer Dienst des Spitzenverbandes Bund der Kranken-kassen e.V. (MDS)
Theodor Althoff-Str. 47 45133 Essen, Germany

(21) Abteilung Medizin. GKV—Reinhardtstraße 28 10117 Berlin, Germany
ABSTRACT

Objective
To develop guidance on what information to include in and how to present it in tables summarizing the evidence from systematic reviews of test accuracy following the GRADE approach.

Methods
To design and refine the evidence tables, we used an iterative process based on the analysis of data from four rounds of discussions, feedback and user testing. During the final round, we conducted one-on-one user testing with target end users. We presented a number of alternative formats of evidence tables to participants and obtained information about users’ understanding and preferences.

Results
More than 150 users participated in initial discussions and provided their formal and informal feedback. 21 users completed one-on-one user testing interviews. Almost all participants preferred summarizing the results of systematic reviews of test accuracy in tabular format rather than plain text. Users generally preferred less complex tables but found presenting sensitivity and specificity estimates only as too simplistic. Users found the presentation of test accuracy for several values of prevalence initially confusing but modifying table layout and adding sample clinical scenarios for each prevalence reduced this confusion. Providing information about clinical consequences of testing result was viewed as not feasible for authors of systematic reviews.
Conclusion

We present the current formats for tables presenting test accuracy following the GRADE approach. These tables are being further developed into electronic interactive tables that will suit the needs of different end users. The formatting of these tables, and how they influence result interpretation and decision-making will be further evaluated in a randomized trial.
WHAT IS NEW?

• The results of systematic reviews of test accuracy can be conceptually complicated. Presenting the data in a clear, comprehensive, comprehensible way that is tailored to different users is critical.

• We present the current GRADE diagnostic summary tables that were developed through extensive user testing.

• The current formats are being further developed into electronic interactive tables that may solve some of the challenges identified by different users.
BACKGROUND

In the previous article in our series addressing diagnostic decision-making in health care, we provided an overview of known tests’ roles and applications, different diagnostic study designs and the main challenges encountered in conducting diagnostic research and applying it to health care decision-making. In this second of a series of eight articles we describe research supporting the development and modifications of tables summarizing the evidence about test accuracy based on a systematic review of the literature. We specifically aim to present the key findings about users’ perspectives and feedback on various formats of diagnostic evidence tables. We also aim to provide an overview of the remaining challenges in presenting TA systematic review results for users in preparation for follow-up work in this area.

Test accuracy (TA) (i.e. performance characteristics of a diagnostic test) is rarely if ever the sole determining factor for the selection of that test in a specific clinical and broader health care situation\(^1\). However, TA results remain an important factor that should be considered and is often the factor that has the most available evidence to synthesize especially when direct evidence, from prospective studies, about effects of a test on patient important outcome is lacking.

Systematic reviews and meta-analysis of TA summarize the available evidence and assess its quality (certainty or confidence in the effect estimates). TA systematic reviews typically focus on tests to establish the presence or absence of a disease, condition or syndrome, and on tests that ultimately categorize results as positive or negative.

Despite significant developments in the methodology of TA systematic reviews, authors still face many challenges, including how best to present their results to different users. The
Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group\(^2\) has previously laid out its approach to making recommendations about diagnostic tests including existing challenges and limitations\(^3\). Appreciating these limitations among users including healthcare providers, guideline developers, and policymakers and their need for TA information in decision making led to the development of diagnostic evidence tables.

In this article “tests” for simplicity of communication refers to all healthcare related tests and diagnostic strategies that are used for different application and roles and not necessarily to make a diagnosis sensu stricto. Tests are used for many different applications: screening or surveillance, risk assessment and classification, diagnosis (ruling in), ruling out diagnosis, treatment triage, treatment monitoring, staging and determining prognosis. Tests are used in different roles in the care pathways: triage, add-on, replacement, parallel testing and replacement of a reference test (reference paper 1 of the series).
METHODS

Setting and design

A variety of formats of the diagnostic evidence tables have been evaluated. We used an iterative process to modify the diagnostic evidence tables based on analysis of data from each round of feedback and user testing. Figure 1 summarizes the different rounds that led to the current suggested formats.

The first round involved discussions among members of the GRADE working group, fueled by specific examples of systematic reviews and addressing how the results can be presented. The second round included collecting feedback from various stakeholders attending several diagnostic GRADE workshops. The feedback was obtained informally during the large and small group discussions as well as formally using specifically designed questionnaires. Appendix 1 presents an example of a work package used in one of such workshops. The third round involved several large and small group discussions in the GRADE working group meeting in Barcelona in January 2013. The fourth round consisted of formal one-on-one user testing either in person or via teleconferencing. Appendix 2 presents two examples of material used in the one-on-one user testing.

Initially we developed tables summarizing the information about the use of a single test; later on we explored ways of summarizing information from comparative TA reviews in which ≥2 tests are compared against each other and a common reference standard.

In this article we provide a summary of our final results after considering the feedback from all rounds. We summarise the finding of the background work and we focus on the results of the final round of user testing.

Background work (rounds 1 and 2)
In the year 2000, the GRADE working group was established to address the shortcomings of existing systems for grading quality of evidence and strength of recommendations in health care. The working group has met regularly 2-3 times a year to develop the GRADE approach and to address methodological challenges in assessing quality of the evidence and strength of recommendations for health care questions including preventive, therapeutic and diagnostic interventions. Since 2002, members of the GRADE Working Group have been involved in preparing and facilitating multiple workshops aimed at training and informing various users of TA systematic reviews representing many organizations and professional groups. Each of 12 GRADE diagnostic workshops consisted of a brief presentation about TA results and applying the GRADE approach to evaluate the quality of evidence, a small group exercise to work through practical examples of TA systematic reviews, and the presentation and evaluation of results in diagnostic evidence tables. Workshops duration ranged between 1.5 hours to two days. Workshops were conducted in different countries around the world, including Canada, Germany, Italy, Singapore, Spain, Switzerland and the United States.

**Diagnostic evidence tables**

Diagnostic evidence tables include three main parts: a heading, the body of the table and the footnotes. The heading contains a description of the health question in a PICO format (population, intervention, comparison and outcomes) and specifications of the index test(s) and the reference test of interest. Similar to the tables used to present results of therapeutic interventions the body of the table summarizes the results of a systematic review of test accuracy and the judgments about the quality of the evidence. Footnotes provide explanations for specific issues in the body of the table, including justification for any judgments made, e.g. about quality of evidence.
There are two main types of GRADE evidence tables: the evidence profile (EvP) and summary of findings (SoF) tables. The EvP reports the estimates for test accuracy measures and the detailed judgments about the five domains for rating the quality of evidence - considerations of risk of bias, indirectness, inconsistency, imprecision, publication bias, and others -. The SoF table follows a similar format. However, it is less complex as detailed judgments about each domain of quality of evidence being explained only in footnotes. Appendix 3 presents an example of GRADE EvP and SoF tables summarizing test accuracy data.

**Participants**

Participants represented an international group who had broad range of experience with DTA systematic reviews, health research methods and GRADE. We surveyed and collected feedback from: authors of TA systematic reviews, methodologists, guideline developers, policy makers, and health care professionals that addressed, developed or used systematic reviews or recommendations about tests.

**Data collection**

The formal one-on-one user testing specifically was intended to compare various formats of evidence tables and to collect user perspectives about the most useful and best possible presentation of information in tables. The results were summarized for TA systematic reviews of single tests and multiple tests that were compared either directly in the same studies or indirectly in different studies against the same reference standard. We used the domains summarized in Figure 2 for our data analysis. We used the different table components as our guide to compile feedback. We also analyzed the comments that addressed the single test versus those that addressed comparative tests separately.

**Data analysis**
Two investigators (RAM and WW) separately reviewed the notes and transcripts of participants’ comments and results of user testing and then discussed their findings. We gathered users’ views on the presentation and formatting, content, comprehensiveness, usefulness and accessibility of results in the evidence tables. We also analyzed reasons for confusion and misunderstanding related to the evidence tables’ content. Before we carried out the next round we modified the tables based on the findings of the previous round.
RESULTS

A range between 11 and 72 members attended each of 25 GRADE working group meetings between 2002 and 2012. More than 150 stakeholders participated in large and small group discussions during workshops and 52 of them completed formal feedback questionnaires about GRADE diagnostic evidence tables. 62 members participated in large and small group discussions and feedback in GRADE working group meetings in 2013 and 21 participants completed one on one user testing interviews (10 for 90 minutes and 11 for 30-60 minutes).

Presenting TA results using different format

Almost all participants preferred summarizing the results of TA systematic reviews in table format. They considered evidence tables as useful and easy to follow.

During the different rounds of feedback collection and user testing we assessed four main formats of tables presenting: 1. sensitivity and specificity estimates only, 2. individual TA numbers (true positives (TP), false positive (FP), true negative (TN) and false negative (FN)) organised based on test results (test positive and negative), 3. individual TA numbers (TP, FN, FP and TN) organised based on disease status (disease present or absent), and 4. likelihood ratios with pre- and post-test probabilities.

Sensitivity and specificity alone (format 1). In early discussions some experts noted that a simple format including only sensitivity and specificity would be sufficient. However, once we tested this simplest format, participants unanimously noted that they did not prefer it. Participants noted that sensitivity and specificity are parameters of the test that
are familiar to most users, but they are often misinterpreted and may not reflect well the effects expected in the population of interest. Participants also noted that this simple table is missing critical information including estimates of prevalence and other measures of test accuracy such as likelihood ratios, predictive values, and absolute numbers of TP, FP, TN and FN that may be more useful for decision making. Hence, later rounds focused on the other three formats of the tables.

*Individual TA values – TP, TN, FP and FN (formats 2 and 3).* Participants generally liked this format but did not have a clear preference for arranging TP, TN, FP and FN in any specific order. Some noted that arranging the rows by test positive (TP and FP) and test negative (TN, FN) makes it more difficult to make a link between the individual test result values and the sensitivity and specificity which are the direct results of a systematic review as they are usually combined across studies in meta-analysis. However, views about what is the more useful order for users varied. Some noted that clinicians are accustomed to thinking about positive and negative test results and that this order may be useful to them. Others noted that arranging the outcomes according to disease status (disease positive (TP and FN) and disease negative (TN and FP)) and making the link to sensitivity and specificity in the table helps better highlight the results, which is most typically used by clinicians and decision makers when applying the test on a population level.

*Likelihood ratios with pre- and post-test probability (format 4).* Some respondents noted that this format represents clinicians’ implicit thinking in terms of changes in post-test probability based on a test result. Others noted that this format with likelihood ratios and probabilities is a more difficult format to use and takes more time to interpret, especially for those who are not familiar with accuracy measures. Some respondents suggested that it is easier to think of patients and test results in absolute numbers, rather than in changes in
probability. It was noted that in this table format, users would most likely use the post-test probability to make decisions about the test while likelihood ratios were considered difficult to understand.

Additionally, multiple participants pointed out the need for flexible tables that allow for qualitative representation of the TA reviews. They explained that this is frequently needed when pooling is not possible either due to methodological challenges or differences among the index test(s) and reference standard in the studies included.

**Evidence tables heading**

The aim of the header section was to give a brief description of the population, condition and the index and reference tests. The intent was to provide enough information about how the tests were applied in the studies, to allow users to judge to what extent the results are applicable in their own setting. This included the tests’ role, application, cut-off values, setting and the population included in the studies. Participants reported that presenting background information about the index and reference tests in the header of the table is helpful to contextualize the information before looking at the TA results. The majority preferred to place the number of participants and studies in the header as a method of avoiding repetition and saving space in the table if the information was the same for each row. It was noted that these evidence tables might be more difficult to use if there are different reference standard tests and multiple index tests or cut-off values.

**Evidence tables summarizing single TA systematic reviews**

Prevalence/pre-test probability/baseline risk estimates
Presenting multiple prevalence estimates allows for interpretation of the test results in different populations as well as comparisons of the test performance in various populations and clinical settings. It may also help users decide which estimate is more applicable to their setting.

Presentation of this information went through multiple changes based on users' feedback. Initially we removed presentation of three prevalence estimates to one/two prevalence estimates in columns (Appendix 4 show earlier tables with three prevalence estimates in rows). We subsequently added clinical scenarios to describe a typical patient in each prevalence group. Also, we changed the label of prevalence from percentage to natural numbers (per 1000) to be consistent with how results were presented in the reminder of the table.

When TA results were presented for 3 prevalence estimates in rows, respondents reported it was confusing and the information was overwhelming. At that time many preferred to have results of the test presented with only a single prevalence estimate. After changing to presenting different prevalence values in columns rather than in rows, more respondents preferred the table with two prevalence estimates compared with one prevalence estimate. Some noted it gave more information and demonstrated test performance in different settings but others felt it remained unclear what the different prevalence estimates represented. Hence, we added a clinical scenario to the label of the column to describe an average patient in each risk group (prevalence). Users viewed this positively and more respondents preferred showing the TA results based on two prevalence values in the table.

Users noted that the number of prevalence estimates that should be presented depends on the test and condition and the available evidence in the studies in the review, and may vary. However, it was noted that the prevalence estimates should be obtained from the highest
quality evidence available that is applicable to the population of interest. For most tests presenting the results for two prevalence settings based on data from observational studies or from the included TA studies seems appropriate. All participants viewed clinical scenarios explaining an average patient in a pre-test probability group as helpful and thus this information should be provided whenever possible. The vast majority preferred clinical scenario to be placed inside the table as opposed to in the footnotes.

**Comments and explanation column**

The comments section of the table is intended to provide generic implications of the test results, e.g. “false positives may lead to unnecessary treatment or additional testing.” or specific information about the downstream consequences, whenever available. Comments were not seen as necessary and helpful to the interpretation of the values in the table for most users. Participants noted that, when authors are explaining the consequences of test results, comments should preferably be based on systematic review(s) of the literature to summarize the highest quality evidence available, and not based on assumptions. It was also suggested that the discussion of probable downstream implications be included in the body of the systematic review rather than in the summary table.

**Labeling of the effect estimate column**

Respondents noted that adding definitions to columns’ labels might be helpful for users not familiar with the definitions, or as a refresher. However, the labels increase the complexity of the table and some labels could be avoided such as definitions of sensitivity and specificity. The majority of respondents preferred ‘Number of results per 1000’ compared to ‘illustrative comparative numbers’ label as they explained it is easier to understand.
Participants noted, although ‘per 1000’ label is repetitive and makes the table busier, it may be helpful to retain as a reminder for the reader and to avoid confusion with percentages.

**Rationale for quality of evidence rating**

We tested the best location to indicate the rational of quality of evidence (QoE) in the table versus in footnotes. Some respondents noted that providing the rationale for QoE rating in the footnotes is sufficient and may not need to be repeated in the table (with brief reason). While some respondents noted that a possible advantage of having the brief reason inside the table is drawing attention to the footnotes, others noted it could have an opposite effect and distract attention from the footnotes. Respondents generally agreed that footnotes explaining QoE rating should be concise and briefly provide the reasons for the rating beyond just stating to which domain concerns applied (e.g. explain why there was concern about the risk of bias instead of just stating “risk of bias”). It was also noted that further elaboration about the QoE rating can and should be provided in the text of the systematic review.

**Evidence tables summarizing comparative TA systematic reviews of ≥2 tests**

These tables summarise the TA results of a comparison of two or more tests. Tests can be compared to each other and to the same reference standard directly in the same studies (direct comparison). Alternatively, one test can be compared to the reference standard in one set of studies and another test compared to the same reference standard in another set of studies to assess the same condition (indirect comparisons).

Most participants indicated that having data about two tests in one table is useful for comparison. However, this increased complexity of the tables. A single table that combines
all information together with the comparison of the tests was viewed by majority of respondents as sufficient and more practical compared to presenting two tables for each test separately. Because there is a comparison between tests, users viewed the absolute differences between the values of the outcomes of each test as helpful information, but this made the table more complex. They also preferred to keep the number of studies, participants and the sensitivity and specificity values in the header. However, when tests were not directly compared in the same studies, participants did not always realize that the lower quality of evidence was due to indirect comparison.

**Evidence tables footnotes**

The aim of the footnotes section was to give further details that are needed to explain judgments and the information provided in table cells. Many respondents did not read the footnotes. They noted that critical information that is needed to understand the results should be included directly in table cells and not be “hidden” in the footnotes, such as information about reference test or prevalence estimates, which were viewed as critical. Respondents preferred short, informative and easy to read statements. They noted that footnotes explaining the QoE ratings are helpful. It was also suggested that it is best to fit the footnotes on one page with the tables. There was general agreement that the labels “explanations” or “clarifications” would be preferred to the label “footnotes”.
DISCUSSION

In this article we present the research supporting the development of GRADE diagnostic evidence tables that display the results of TA systematic reviews. Overall, users viewed presenting the results of TA reviews in table layout as useful. However, they preferred simple tables that present the results in a format that is easy to apply to patients and populations. As a result of our sequential rounds of user testing and revisions, current versions of the TA systematic reviews evidence tables were agreed upon (Tables 1, 2, and 3: diagnostic evidence tables).

Our study has multiple strengths. We conducted extensive usability testing with different user groups with a variety of experience with GRADE, TA systematic reviews, and decision making about tests. This helped us to develop table formats that will be useful for a diverse group of users as well as different applications and settings. We interviewed an international group of users whose first languages were often other than, but not limited to English, adding transferability of our findings and the usability of these tables around the world. We also contacted participants to clarify any confusion about their responses. Additionally, we asked users to suggest solutions to make the tables more useful and minimize any misunderstanding about the results. At least two investigators reviewed the notes and transcripts of the user testing interviews to minimize bias in describing our findings.

Our study has a few limitations. Regardless of table layout, there is a learning curve associated with general understanding the information about test accuracy and systematic reviews of studies measuring it. Understanding and using GRADE diagnostic evidence
tables also requires getting accustomed to this presentation of information. The more often they are used, the more familiar users become with the tables. To avoid “learning effect” bias that may affect users feedback, we randomized the order by which participants saw the different formats. Additionally, we observed variability in comprehension and preference for the different formats. In some instances, preference was not in line with the correct understanding of the results, which created a challenge for the developers. When discordance between understanding and preference was observed, we relied on understanding as our primary outcome of interest to guide the development of different formats and to investigate the need to modify any parts of the table to make them more intuitive and transparent. Also, respondents were volunteers that agreed to participate in our study, which may be viewed as a limitation due to self selection. To avoid that, we contacted a variety of potential respondents and included users that have a wide range of experience and background.

When designing evidence tables for systematic reviews of test accuracy we considered all experience from developing GRADE evidence tables for interventions and plain language summaries\(^5\)\(^-\)\(^8\). However, this is the first study assessing users views about diagnostic evidence tables, which entail unique challenges. While determining patient important outcomes is relatively straightforward in intervention reviews, determining which outcomes to present in TA systematic reviews is more complicated. Since only TA results are available to authors of systematic reviews, we collected feedback about which form of TA results to present. We identified three formats that presented TA results in a layout that was acceptable to different users. Participants agreed that while sensitivity and specificity are the most commonly reported in TA systematic reviews, presenting pre- and post-test probability based on likelihood ratios and absolute TA results of TP, FP, TN and FN may be
needed to comprehend the implications of the results by users. Probabilities may be a more useful measure for decision making about the use of a test, albeit likely more difficult for some users to understand. Table formats presenting individual TA values (i.e. TP, TN, FP, FN) may be more useful at the population level, while the format presenting likelihood ratios with pre- and post-test probabilities is more useful for decision making about individual patients in the clinical setting.

Presentation of multiple estimates of prevalence was identified as important for the application of the test and deciding whether to use it in a given setting. The decision to use the test may differ for a population with low prevalence versus higher prevalence as one considers the test results (e.g. TP, FN, TN, FP) and proportions for the specific population. However, unlike presentation of several baseline risk groups when estimating absolute effects of therapeutic interventions, the importance of presenting test accuracy results based on several assumed values of prevalence (pre-test probability) was not clear to participants until we added clinical scenarios describing an average patient in each prevalence (risk) group. How best to develop the clinical scenarios remains an open question. They could be obtained from the literature or suggested by an expert panel based on the observations from clinical practice. Participants also suggested that TA systematic review authors would need to ensure it is representative of an average patient in that risk category, a challenge that requires clinical expertise. For specific conditions authors of systematic reviews may be able to use validated prediction models of baseline risk or pre-test probability to support their scenarios (e.g. Framingham coronary heart disease prediction score). However, it is likely that studies with an applicable example clinical scenario may be challenging for many topics.
Systematic reviews of interventions typically address questions comparing the outcomes from two management options, even if one of them is no intervention or placebo. This, however, changes with the more widespread use of network meta-analytical approaches that enable multiple direct and indirect comparisons. Systematic reviews of TA assess the accuracy of a single test or multiple tests. This creates a unique challenge, conceptually similar to the challenges of simple presentation of the results of multiple comparisons of treatments, as one has to develop tables that can present results of the accuracy estimates either for single test or multiple tests while attempting to keep the structure similar to minimize confusion among users. There was general consensus among our study participants that tables should be simple both in the information they provide and in their design. Achieving this was naturally more challenging when developing evidence tables for comparative diagnostic accuracy of more than one test.

Majority of the users agreed that evidence tables should be included in TA systematic reviews, as it improves accessibility of the results. However, similar to intervention reviews, we also learned that authors may have difficulty preparing the tables since not all tests or conditions will fit a standard layout. There may also be other methodological issues such as ability to pool the data, especially from different types of studies (e.g. differences in reference test, index tests, indirect comparisons, heterogeneity, etc.). Participants reported that evidence tables would be more acceptable to authors of TA reviews if there were some options for content and layout to accommodate different clinical topics (e.g. interactive table). This led to the development of an electronic version of a summary table that integrates narrative interpretation of the results as well as definitions without crowding the table and making it overwhelming. This interactive format is currently under development and testing.
In this article we presented the results of the extensive user testing that led to the development of the current formats of the evidence tables for test accuracy as suggested by the GRADE approach. These formats are being further developed into electronic interactive tables that will suit the needs of different users. These tables and their impact on understanding the results and making decisions in regards to the test(s), will be evaluated through a randomized trial.
Acknowledgement

This work was partially funded by the German Insurance Fund agency as part of a larger project about decision-making for healthcare related tests and diagnostic strategies. It was also partially funded by a Methods Innovation Fund from the Cochrane Collaboration. The views presented here are those of the authors and should not be attributed to the funding agency or its staff.

We are grateful to all who participated in user testing.
References

### TABLE 3.1: DTA SOF TABLE BASED ON TEST RESULTS (TP, FP AND TN, FN)

**Abdominal ultrasound for the diagnosis of pancreatic cancer**

| Patients or population: | symptomatic patients in primary care with suspicion of pancreatic cancer |
| Setting: | mainly outpatients |
| New Test: | abdominal ultrasound¹ |
| Reference Test: | endoscopic ultrasound with biopsy² |
| Threshold: | Proven or probable pancreatic cancer |

**Pooled Sensitivity (95% CI): 0.64 (0.50 to 0.77) | Pooled Specificity (95% CI): 0.95 (0.91 to 0.97)**

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested³ (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prevalence 20 per 1000⁴:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in otherwise healthy adults presenting with symptoms of jaundice, fatigue, pain of the abdomen, and dark urine.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prevalence 400 per 1000⁴:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in older adults presenting with symptoms of jaundice, fatigue and pain, with a family history of pancreatic cancer, history of chronic pancreatitis, who have diabetes, and are current or past smokers.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Test-Positive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td>282 per 1000 (220 to 339 per 1000)</td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td>49 per 1000 (30 to 89 per 1000)</td>
<td>28 per 1000 (17 to 50 per 1000)</td>
<td>2777 (18 studies) ⊕⊕⊕⊕ High⁵</td>
</tr>
<tr>
<td><strong>Test-Negative</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td>532 per 1000 (510 to 543 per 1000)</td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td>7 per 1000 (4 to 10 per 1000)</td>
<td>158 per 1000 (101 to 220 per 1000)</td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval

**Footnotes:**

¹ A diagnostic test for pancreatic cancer needs to be less invasive than the current reference standard and lessen the burden to patients.

² Endoscopic ultrasound with biopsy as a reference test is a more invasive test, placing higher burden on the patient, with risk of complications such as infection.

³ Assumed numbers with abdominal ultrasound compared to reference test.

⁴ Estimates of prevalence of pancreatic cancer were based on the median and range of values in included studies.

⁵ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### TABLE 3.2: DTA SOF TABLE BASED ON DISEASE STATUS (TP, FN AND TN, FP)

**Abdominal ultrasound for the diagnosis of pancreatic cancer**

**Patients or population:** symptomatic patients in primary care with suspicion of pancreatic cancer  
**Setting:** mainly outpatients  
**New Test:** abdominal ultrasound 1  
**Reference Test:** endoscopic ultrasound with biopsy 2  
**Threshold:** Proven or probable pancreatic cancer

**Pooled Sensitivity (95% CI):** 0.64 (0.50 to 0.77) | **Pooled Specificity (95% CI):** 0.95 (0.91 to 0.97)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test-Positive</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td>282 per 1000 (220 to 339 per 1000)</td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td>49 per 1000 (30 to 89 per 1000)</td>
<td>28 per 1000 (17 to 50 per 1000)</td>
<td>2777 (18 studies) ⊕⊕⊕⊕ High</td>
</tr>
<tr>
<td>Test-Negative</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td>532 per 1000 (510 to 543 per 1000)</td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td>7 per 1000 (4 to 10 per 1000)</td>
<td>158 per 1000 (101 to 220 per 1000)</td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval

**Footnotes:**
1 A diagnostic test for pancreatic cancer needs to be less invasive than the current reference standard and lessen the burden to patients.  
2 Endoscopic ultrasound with biopsy as a reference test is a more invasive test, placing higher burden on the patient, with risk of complications such as infection.  
3 Assumed numbers with abdominal ultrasound compared to reference test.  
4 Estimates of prevalence of pancreatic cancer were based on the median and range of values in included studies.  
5 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
# TABLE 3.3: DTA SOF TABLE BASED ON PRE- AND POST-TEST PROBABILITY

Abdominal ultrasound for the diagnosis of pancreatic cancer

**Patients or population:** symptomatic patients in primary care with suspicion of pancreatic cancer  
**Setting:** mainly outpatients  
**New Test:** abdominal ultrasoundⁱ  
**Reference Test:** endoscopic ultrasound with biopsy²  
**Threshold:** Proven or probable pancreatic cancer

<table>
<thead>
<tr>
<th>Pooled Sensitivity (95% CI): 0.64 (0.50 to 0.77)</th>
<th>Pooled Specificity (95% CI): 0.95 (0.91 to 0.97)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Pre-test probability³</th>
<th>Test result</th>
<th>Post-test probability⁴ (95% CI)</th>
<th>Likelihood ratio (LR) (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Probability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td>+</td>
<td>21% (17% to 26%)</td>
<td>+ LR = 13 (10 to 17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in otherwise healthy adults presenting with symptoms of jaundice, fatigue, pain of the abdomen, and dark urine.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>−</td>
<td>1% (1% to 1%)</td>
<td>− LR = 0.38 (0.27 to 0.54)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| **High Probability**  |             |                                 |                               |                               |                               |
| 44%                   | +           | 91% (89% to 93%)                | + LR = 13 (10 to 17)          | 2777 (18 studies) | High⁵ |
| Which is typically seen in older adults presenting with symptoms of jaundice, fatigue and pain, with a family history of pancreatic cancer, history of chronic pancreatitis, who have diabetes, and are current or past smokers. | | | | | |
| −                     | −           | 23% (22% to 24%)               | − LR = 0.38 (0.27 to 0.54)   |                               |                               |

CI: Confidence interval

**Footnotes:**

¹ A diagnostic test for pancreatic cancer needs to be less invasive than the current reference standard and lessen the burden to patients.  
² Endoscopic ultrasound with biopsy as a reference test is a more invasive test, placing higher burden on the patient, with risk of complications such as infection.  
³ Pre-test probability of pancreatic cancer was selected based on the median and range of prevalence values in included studies.  
⁴ Percentages are rounded to the nearest integer.  
⁵ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
Figure 3.1. Outline of the rounds of feedback and user testing to develop GRADE diagnostic summary tables

<table>
<thead>
<tr>
<th>Event/Dates</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRADE working group members' discussion 2002-2012</td>
<td>Between 11 and 72 members attended each of 25 meetings</td>
</tr>
<tr>
<td>GRADE diagnostic workshops 2008-2012</td>
<td>&gt;150 attendees participated in large and small groups discussions (52 completed questionnaires)</td>
</tr>
<tr>
<td>GRADE working group meeting 2013</td>
<td>62 members participated in large and small group discussion</td>
</tr>
<tr>
<td>One-on-one user testing 2013</td>
<td>21 users (10 users --&gt; 90 minutes, 11 users --&gt; 30-60 minutes)</td>
</tr>
</tbody>
</table>

Figure 3.2. Summary of the domains used for data analysis of user testing and feedback

<table>
<thead>
<tr>
<th>Presenting TA results using different format</th>
<th>Heading (summarizing question of interest, index and reference tests’ characteristics)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Single TA systematic review</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Comparative TA systematic review</strong></td>
</tr>
<tr>
<td>Prevalence/pre-test probability/baseline risk</td>
<td>Comments and explanations</td>
</tr>
<tr>
<td></td>
<td>Labeling of the effect estimate column</td>
</tr>
<tr>
<td></td>
<td>Quality of Evidence</td>
</tr>
<tr>
<td>Same studies directly assessing the accuracy of ≥2 tests</td>
<td>Different studies indirectly assessing the accuracy of ≥2 tests</td>
</tr>
</tbody>
</table>
APPENDIX 3.1: WORKSHOP PACKAGE GRADING THE QUALITY OF EVIDENCE AND PREPARING SUMMARY OF FINDINGS TABLES ABOUT DIAGNOSTIC TESTS

Aim of this exercise
Summary of Findings (SoF) tables provide a concise, easy to understand summary of results of a systematic review. The GRADE Working Group and the Cochrane Collaboration Applicability and Recommendations Methods Group have suggested an approach to grading the quality of evidence for questions about diagnostic accuracy. This workshop will introduce the approach based on examples. We will discuss the content of the currently suggested format of SoF tables and seek feedback on improving them.

1. Read the abstract of the review
2. Identify the clinical question asked and its components


Question: _______________________________________________________________

Setting: __________________________________________________________________

Population (prevalence): ____________________________________________________

New test: __________________________________________________________________

Existing test: __________________________________________________________________

Cut-off criterion: __________________________________________________________________

Exercise
Please review attached examples of proposed Summary of Findings table and an Evidence Profile. The aim of the following questions is to assess whether the currently proposed table format is easy to understand – we do not evaluate your knowledge.

Do you have experience with diagnostic test accuracy studies? □ Yes □ No
Do you have experience with systematic reviews of diagnostic test accuracy studies? □ Yes □ No

Overall appearance of the tables
To what extent do you agree with the following statements?

1. The information in the Evidence Profile is clearly presented.
   - [ ] I strongly disagree
   - [ ] I disagree
   - [ ] I somewhat disagree
   - [ ] I am not sure
   - [ ] I somewhat agree
   - [ ] I agree
   - [ ] I strongly agree

2. The information in the Summary of Findings table is clearly presented.
   - [ ] I strongly disagree
   - [ ] I disagree
   - [ ] I somewhat disagree
   - [ ] I am not sure
   - [ ] I somewhat agree
   - [ ] I agree
   - [ ] I strongly agree

3. Generally, the Evidence Profile is easy to understand.
   - [ ] I strongly disagree
   - [ ] I disagree
   - [ ] I somewhat disagree
   - [ ] I am not sure
   - [ ] I somewhat agree
   - [ ] I agree
   - [ ] I strongly agree

4. Generally, the Summary of Findings table is easy to understand.
   - [ ] I strongly disagree
   - [ ] I disagree
   - [ ] I somewhat disagree
   - [ ] I am not sure
   - [ ] I somewhat agree
   - [ ] I agree
   - [ ] I strongly agree

Please provide suggestions for improvement, in particular if you disagree with the above statements:

Comprehensibility
5. In the row “true positives” under the heading “prevalence 1%” the number “6 (5 to 8)” means that:
   - 6 out of 100 patients who receive the new test (95% confidence interval: 5 to 8) are correctly identified as having the disease.
   - 6 out of 1000 patients who receive the new test (95% confidence interval: 5 to 8) are correctly identified as having the disease.
   - 6 out of 2777 patients who receive the new test (95% confidence interval: 5 to 8) are correctly identified as having the disease.
   - only 6 out of 1000 patients who suffer the disease (95% confidence interval: 5 to 8) will test positive.
   - if the prevalence decreases from 12% to 1%, 6 additional people with the disease will be identified.
   - I cannot find the answer
   - I do not understand the question

6. The risk of false positive results in patients who receive a new test in a setting with a prevalence of 1% is:
   - 6 per 1000
   - 28 per 1000
   - 30 per 1000
   - 50 per 1000
   - 89 per 1000
   - I cannot find the answer
   - I do not understand the question

7. “High quality of evidence” for the outcome “true positives” refers to:
   (Check the most complete answer)
   - the extent to which we can be confident that the estimate of the rate of “true positives” is correct
   - the extent to which we can be confident about the consequences important to patients correctly identified as having the disease
   - an overall estimate of the quality of all reviewed diagnostic and therapeutic studies
   - the overall quality of evidence of the reviewed diagnostic studies and the confidence about the consequences of being classified as “true positive” for patients in whom the test will be used
   - the risk of bias of reviewed diagnostic studies.
   - I cannot find the answer
   - I do not understand the question
8. The number 7 in the “importance” column for “true positive” results in the Evidence Profile indicates: (Check the most complete answer)

☐ low importance of “true positive results, implying they are not critical to making a decision

☐ that the correct labeling of patients as having the disease in question is sufficient for recommending the test in clinical practice.

☐ that the patient-important consequences related with being correctly diagnosed as having the disease are regarded as very important

☐ that “true positives” are the only outcome that should be considered for decision-making

☐ that the clinical decision about using the test in question in immunocompromised patients should not be based on the true positive results

☐ I cannot find the answer

☐ I do not understand the question

9. In the framework or grading the quality of evidence, directness refers to: (Choose ALL answers that you think are correct)

☐ the degree to which we are confident how the outcomes important to patients will be affected by the correct or incorrect diagnosis and subsequent management

☐ the degree to which the tests studied are comparable to those being used in clinical practice

☐ the degree to which the diagnostic expertise of people applying the tests in the studies is comparable to the expertise of those who will interpret the test in clinical practice

☐ the degree to which patients in the studies resemble those in whom the test will be used in clinical practice

☐ whether the tests being assessed and the reference standard were all assessed in the same study populations or whether each of the tests was compared to the reference standard in a different study population

☐ I cannot find the answer

☐ I do not understand the question

10. Please provide any comments that you may have about the comprehensiveness of the tables, including suggestions about how it could be improved, e.g. should there be a graphical presentation of the results?
**Accessibility of the results**

**To what extent do you agree with the following statements?**

**11. The authors of a systematic review have indicated what they considered to be the most important outcomes for someone considering the diagnostic test.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree

**12. What outcomes were the most important according to the authors?**

[ ] True positive  [ ] True negative  [ ] False positive  [ ] False negative  [ ] Inconclusive  [ ] Resource use  [ ] I cannot find the answer  [ ] I do not understand the question

**13. It was easy to find the information about what outcomes are the most important.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree

**14. It was easy to find the results for each of the outcomes.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree

**15. It was easy to understand the results for each of the outcomes.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree

**16. The main findings of the review are presented in a way that would help me to make a decision.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree

**17. It was easy to find information about the quality of the evidence for each outcome.**

[ ] I strongly disagree  [ ] I disagree  [ ] I somewhat disagree  [ ] I am not sure  [ ] I somewhat agree  [ ] I agree  [ ] I strongly agree
18. **Overall, how would you rate the accessibility of the main findings of this review?**
   
   By “accessibility” we mean the extent to which the main findings are easy to find, understand, and use by someone making a decision.

   - [ ] Very inaccessible
   - [ ] Inaccessible
   - [ ] Somewhat inaccessible
   - [ ] I am not sure
   - [ ] Somewhat accessible
   - [ ] Accessible
   - [ ] Very accessible

19. **Please provide comments you may have about the accessibility of the main findings of the review in the space below, including suggestions about how this could be improved:**

20. **Is there any other information you would like to see in the tables summarizing the results of a review?**

   - [ ] Yes
   - [ ] No
   - [ ] I do not know

   If yes, please describe the information you would like to see presented additionally.

<table>
<thead>
<tr>
<th>Evaluation of this QUESTIONNAIRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>21. <strong>Do you have any suggestions how to improve the comprehension and clarity of the above questions?</strong></td>
</tr>
</tbody>
</table>
| - [ ] Yes
| - [ ] No
| - [ ] I do not know

If yes, please list the questions that were challenging and provide suggestions for improvement.
Workshop date:
Venue:
Facilitators:

Exercise
- Work in small groups
- Select someone from your small group to report back to the whole group (take notes!)
- Watch the time
- Follow the subsequent instructions

Instructions (outline – you will find detailed instructions on the respective pages as labeled)
1. Read the abstract of the systematic review (attachment)
2. Identify the clinical question asked in the review and its components (page 2)
3. Identify the main comparison from the review that you want to work on (e.g. BNP vs no BNP). Read the relevant sections of the review focusing on the text marked-up on the margin.
4. Complete the assessment of the quality of evidence in an evidence profile (page 3, instructions on page 4 and 5)
   - Go to the marked-up text in the systematic review to evaluate the quality of evidence
   - Make judgment about the quality of the evidence for each outcome
   - Make judgments about the overall quality of evidence
5. Provide estimates of the effect of using a test in a population of patients with an assumed prevalence. Obtain the sensitivity and the specificity of the test in order to calculate the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) results (page 6)
   - Use a hypothetical pre-test probability (prevalence) of the disease of 60%.
   - Calculate true positives, false positives, true negatives, and false negatives (page 6)
6. Provide information about the magnitude of desirable and undesirable outcomes (page 7 and 8)
   - Choose the most important outcome for decision making
   - Rate the relative importance of the effect
   - Fill out the table to provide information about the outcomes
7. Develop a recommendation for the use of this test (page 11 and 12)

During the exercise please contemplate how a Summary of Findings table for systematic reviews of diagnostic accuracy studies and an Evidence Profile for decision making (choosing among available tests) should look like.
1. Specify details of the review and the clinical question asked

*Title of the review*: Diagnostic accuracy of pleural fluid NT-pro-BNP for pleural effusions of cardiac origin: a systematic review and meta-analysis

**Identify the following components:**

- **Patients** (P: Population) and **Prevalence**:

- **Purpose** of the test (triage, replacement or add-on, screening, making diagnosis, monitoring)

- **New/index test** (I: Intervention):

- **Existing/comparator/reference test** (C: Comparator):

- **Outcomes of interest** (O): [consider all outcomes that patients could experience with or without treatment – use worksheet 1 on page 3]

- **Cut-off threshold** (if relevant):
2. Assess **quality of evidence** across studies for **each aspect** in a review
(see example at the end of this work package)

**Quality assessment for diagnostic accuracy studies**

<table>
<thead>
<tr>
<th>Test result</th>
<th>No. of studies</th>
<th>Study design</th>
<th>Factors that may decrease quality of evidence</th>
<th>Final Quality (symbols)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Risk of bias</td>
<td>Publication bias</td>
</tr>
<tr>
<td>sensitivity (TP + FN)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>specificity (FP + TN)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inconclusive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Footnotes:
### GRADE quality assessment criteria for diagnostic accuracy studies

#### Underlying study design

Valid diagnostic accuracy studies (cross-sectional or cohort) in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are initially rated as **high quality** evidence. These studies are rare, however.

#### Factors that may decrease the quality of evidence

- Limitations in design or execution of the study (risk of bias)
- Indirectness (comparison or the population, new test, comparison test, and outcomes)
- Inconsistency in study results
- Imprecise results
- High probability of publication bias

If any of the factors warranting downgrading is present, consider if the limitations are **serious** (downgrade by **one** level) or **very serious** (downgrade by **two** levels).

#### Factors that may increase the quality of evidence

#### Factors that determine or decrease the quality of evidence for studies of diagnostic accuracy

<table>
<thead>
<tr>
<th>Factors that determine quality</th>
<th>Explanations how they differ from evidence about treatment options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study design</strong></td>
<td><strong>Different criteria</strong> for accuracy studies than for management trials</td>
</tr>
<tr>
<td></td>
<td>Cross-sectional or cohort studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard are considered high quality and can move to moderate, low or very low depending on the following factors.</td>
</tr>
<tr>
<td><strong>Limitations in design or execution</strong> (risk of bias)</td>
<td><strong>Different criteria</strong> for accuracy studies than for management trials</td>
</tr>
<tr>
<td></td>
<td>Consecutive patients should be recruited as a single cohort and not classified by disease state and the selection as well as the referral process should be clearly described. Tests should be performed in all patients in the same patient population for the new test and a well described reference standard – the evaluators should be blind to the results of the alternative test and reference standard.</td>
</tr>
<tr>
<td><strong>Indirectness</strong> of evidence</td>
<td><strong>Similar criteria</strong> for accuracy studies and for management trials</td>
</tr>
<tr>
<td><strong>Outcomes</strong></td>
<td>Diagnostic accuracy studies do not provide direct evidence about patient-important outcomes. One must make deductions about the balance between the presumed influences of any differences in true and false positives and true and false negatives on patient-important outcomes in relationship to test complications and costs (diagnostic accuracy studies typically provide low quality evidence for making recommendations due to indirectness of outcomes, similar to surrogate outcomes for treatments).</td>
</tr>
<tr>
<td></td>
<td><strong>Similar criteria</strong> for accuracy studies and for management trials</td>
</tr>
<tr>
<td></td>
<td>The quality of evidence can be lowered if there are important differences:</td>
</tr>
<tr>
<td></td>
<td>1) between the populations studied and those for whom the recommendation is intended (e.g. the spectrum of disease or comorbidity)</td>
</tr>
<tr>
<td></td>
<td>2) in tests studied and the expertise of those applying them in the studies compared to the settings for which the recommendations are intended</td>
</tr>
<tr>
<td><strong>Patient populations, diagnostic test/intervention, comparison test/intervention, and indirect comparisons</strong></td>
<td>The quality of evidence can be also lowered if the tests being compared are each compared to a reference (gold) standard in different studies but not directly in the same studies.</td>
</tr>
</tbody>
</table>
Inconsistency of the results

**Similar criteria** for accuracy studies and for management trials
For accuracy studies, unexplained inconsistency in sensitivity, specificity or likelihood ratios (rather than relative risk or mean differences) can lower the quality of evidence.

Imprecision of the results

**Similar criteria** for accuracy studies and for management trials
For accuracy studies, wide confidence intervals for estimates of test accuracy, or true and false positive and negative rates can lower the quality of evidence.

Publication bias

**Similar criteria** for accuracy studies and for management trials
A high risk of publication bias (e.g. evidence from small studies for a new intervention or test, or asymmetry in a funnel plot) can lower the quality of evidence.

References:

**Quality criteria of diagnostic accuracy studies**


<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Yes</th>
<th>Unclear</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Was the spectrum of patients representative of the patients who will receive the test in practice? (representative spectrum)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Is the reference standard likely to classify the target condition correctly? (acceptable reference standard)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests? (acceptable delay between tests)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Did the whole sample or a random selection of the sample, receive verification using the intended reference standard? (partial verification avoided)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Did patients receive the same reference standard irrespective of the index test result? (differential verification avoided)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)? (incorporation avoided)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Were the reference standard results interpreted without knowledge of the results of the index test? (index test results blinded)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Were the index test results interpreted without knowledge of the results of the reference standard? (reference standard results blinded)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Were the same clinical data available when test results were interpreted as would be available when the test is used in practice? (relevant clinical information)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Provide estimates of the effect of using a test in a population of patients with an assumed prevalence. Assumed typical pre-test probability (“prevalence”): 60 %

<table>
<thead>
<tr>
<th>Test findings</th>
<th>Quality of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled sensitivity</td>
<td></td>
</tr>
<tr>
<td>___ (95% CI: ___ to ___)</td>
<td></td>
</tr>
<tr>
<td>Pooled specificity</td>
<td></td>
</tr>
<tr>
<td>___ (95% CI: ___ to ___)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consequences</th>
<th>Number per 1000 tested*</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>Inconclusive results**</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
</tr>
</tbody>
</table>

* all results are given per 1000 patients tested based on the prevalence of ___ % and pooled sensitivity and specificity.
** inconclusive results are either uninterpretable, indeterminate or intermediate test results
Calculate an absolute number of patients with a given test result based on the combined sensitivity and specificity from meta-analysis and an assumed prevalence of a target condition using a 2x2 table.

<table>
<thead>
<tr>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease present</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
</tbody>
</table>

Prevalence: ___ %

Example calculation for an **assumed prevalence of 20%**.

<table>
<thead>
<tr>
<th>Reference standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease present</td>
</tr>
<tr>
<td>Positive</td>
</tr>
<tr>
<td>Negative</td>
</tr>
</tbody>
</table>

Prevalence: **20 %**

4. Choose the most important outcomes for decision making

Choose the most important outcomes for decision making that are a consequence of the condition, disease or a problem (e.g. mortality, stroke, disability, bleeding complications, etc.). Consider all consequences regardless of whether or not the correct diagnosis is made (e.g. false positives and whether the consequences happen with or without subsequent therapy (e.g. false negatives will not receive potentially beneficial treatment and false positives may receive unnecessary but potentially harmful therapy). If you are not familiar with the topic talk to your neighbor or make assumptions and the best guesses.

Consider the outcomes that:
- are part of the natural history of a disease
- may have been reported in systematic reviews and/or individual studies
Might be important to someone making a decision to use or not to use a test, including complications of performing tests being compared (include both benefits and adverse effects, and costs if relevant)

Consider efficacy of a treatment for all important outcomes and the rate of adverse effects.

Complete a list of the assumptions about patient outcomes/consequences (use worksheet below):

- E.g. Relative Risk Reduction for Mortality: 50%
- E.g. Relative Risk Increase for severe adverse outcomes: 20%

You may draw a flow diagram of what may happen to patients to support your thinking.

List all outcomes and specify how frequently they would occur in each of the four categories (overwrite the grey example):

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Assumptions about the effect of appropriate treatment (RRR)</th>
<th>Include in TP, TN, FP, FN (at what rate would the outcome occur with or without correct therapy?)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. mortality at 30 days</td>
<td>50%</td>
<td>TP: 35%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TN: 5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FP: 7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FN: 70%</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Footnotes:
1 Those correctly diagnosed with the disease and treated successfully (RRR: 50%)
2 This proportion will die regardless within 30 days
3 Adverse effects of therapy cause an additional 2% mortality
4 Proportion that will die because they will have not received therapy or therapy will have been delayed

5. Rate the relative importance of the effect
Consider patient-important consequences of being correctly or incorrectly classified as having or not having a disease. Assess the relative importance of each outcome and decide which is critical for the decision to use the test. Use the percentages that
you have assumed in worksheet 1. Begin by listing the consequences followed by the percentage and then rate the outcomes according to the following scale.

Rate the importance on a 9-point scale (please note that you can assign the same rating several times):

| 1 – 3  | not important (not included in the evidence profile) |
| 4 – 6  | important, but not critical for making a decision (included in the evidence profile) |
| 7 – 9  | critical for making a decision (included in the evidence profile) |

<table>
<thead>
<tr>
<th>Patient-important consequences</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP: (e.g. mortality 35%; severe stroke 5%; etc)</td>
<td></td>
</tr>
<tr>
<td>TN:</td>
<td></td>
</tr>
<tr>
<td>FP:</td>
<td></td>
</tr>
<tr>
<td>FN:</td>
<td></td>
</tr>
<tr>
<td>Inconclusive results*:</td>
<td></td>
</tr>
<tr>
<td>Cost:</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

*uninterpretable, indeterminate or intermediate test results*
6. Consider the patient important outcomes and how directly they relate to diagnostic accuracy

Summary of Findings Table based on patient outcomes

<table>
<thead>
<tr>
<th>Population/ Setting</th>
<th>New Test/ cut-off value</th>
<th>Comparison Test/ cut-off value</th>
<th>Reference Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity ___ (CI - - )</td>
<td>Specificity ___ (CI - -)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Patient outcomes (importance)</th>
<th>Assumed patient consequences</th>
<th>Basis of assumptions</th>
<th>Results per 1000 patients/year for a given pre-test probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mortality (9)</td>
<td>True positive will have improved mortality due to early diagnosis and false negative will have worse mortality due to late diagnosis</td>
<td>We assumed 10% increased mortality risk every year from CHF to be uniformed across different stages. We also assumed 10% RRR in mortality for early diagnosis of CHF.</td>
<td>0 more patients will live</td>
</tr>
</tbody>
</table>

Footnotes: **
(1)
7. Move from evidence to recommendation
6.1. Determine overall quality of evidence

Overall quality of evidence across all critical outcomes is:

<table>
<thead>
<tr>
<th>Choose one</th>
<th>Symbol</th>
<th>Quality</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>⊗⊗⊗⊗</td>
<td>High</td>
<td>Further research is very unlikely to change our confidence in the estimate of effect or accuracy.</td>
</tr>
<tr>
<td></td>
<td>⊗⊗⊗</td>
<td>Moderate</td>
<td>Further research is likely to have an important impact on our confidence in the estimate of effect or accuracy and may change the estimate.</td>
</tr>
<tr>
<td></td>
<td>⊗⊗</td>
<td>Low</td>
<td>Further research is very likely to have an important impact on our confidence in the estimate of effect or accuracy and is likely to change the estimate.</td>
</tr>
<tr>
<td></td>
<td>⊗</td>
<td>Very low</td>
<td>Any estimate of effect or accuracy is very uncertain.</td>
</tr>
</tbody>
</table>

6.2. Values and preferences (assume a set of values for each outcome considered)

Example: A high value may be placed on the true positives, false negatives, and false positives because the associated treatments both have important benefits and harms, but relatively low values on true negatives and resources use.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Values and preferences</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td></td>
</tr>
<tr>
<td>TN</td>
<td></td>
</tr>
<tr>
<td>FP</td>
<td></td>
</tr>
<tr>
<td>FN</td>
<td></td>
</tr>
<tr>
<td>Inconclusive results</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
</tr>
</tbody>
</table>

6.3. Draft recommendation
8. Decide about the strength of a recommendation (strong or conditional/weak)

Use the table below to make a judgment. The four factors in this table will determine whether the recommendation is likely to be “strong” or “conditional” (aka “weak”). Frequent positive answers increase the likelihood of a strong recommendation. Make sure to add an explanation for your judgment.

<table>
<thead>
<tr>
<th>Factors that can weaken the strength of a recommendation</th>
<th>Decision</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High or moderate quality evidence</strong> (is there high or moderate quality evidence?)</td>
<td>□ Yes □ No</td>
<td>The higher the quality of evidence, the more likely is a strong recommendation. The lower the quality of evidence the more likely is a conditional/weak recommendation.</td>
</tr>
<tr>
<td><strong>Certainty about the balance of benefits versus harms and burdens</strong> (is there certainty?)</td>
<td>□ Yes □ No</td>
<td>The larger the difference between the desirable and undesirable consequences and the certainty around that difference, the more likely a strong recommendation. The smaller the net benefit and the lower the certainty for that benefit, the more likely is a conditional/weak recommendation.</td>
</tr>
<tr>
<td><strong>Certainty or similarity in values</strong> (is there certainty?)</td>
<td>□ Yes □ No</td>
<td>The smaller the variability or the greater the certainty around values and preferences, the more likely is a strong recommendation.</td>
</tr>
<tr>
<td><strong>Resource implications</strong> (are the resources consumed worth the expected benefit)</td>
<td>□ Yes □ No</td>
<td>The higher the costs of an intervention compared to the alternative that is considered and other cost related to the decision – that is, the more resources consumed – the more likely is a conditional/weak recommendation.</td>
</tr>
</tbody>
</table>

If consensus has not been reached by discussion, the panel can use the following table to record their views (votes) about the strength of the recommendation related to a specific management option, based on their analysis of the available evidence and its quality, the benefits and downsides, values and preferences, and resource use (cost). This assessment is then mapped to the strength of recommendation for the use, or non-use, of each intervention.

If you need to vote: Insert the number of votes for the recommendation in each category

<table>
<thead>
<tr>
<th>Assessors’ view of the balance between desirable and undesirable consequences of the intervention</th>
<th>Desirable consequences clearly outweigh undesirable consequences</th>
<th>Desirable consequences probably outweigh undesirable consequences</th>
<th>Undesirable consequences probably outweigh desirable consequences</th>
<th>Undesirable consequences clearly outweigh desirable consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Strength of recommendation</strong></td>
<td>Strong for an intervention</td>
<td>Conditional (weak) for an intervention</td>
<td>Conditional (weak) against an intervention</td>
<td>Strong against an intervention</td>
</tr>
<tr>
<td><strong>Wording of a recommendation</strong></td>
<td>We recommend to “do something”</td>
<td>We suggest (conditionally recommend) to “do something”</td>
<td>We suggest (conditionally recommend) not to “do something”</td>
<td>We recommend not to “do something”</td>
</tr>
</tbody>
</table>

| Number of votes | | | | |
| Strength of the recommendation: | □ Strong  
□ Conditional (weak) |

<table>
<thead>
<tr>
<th>Final recommendation</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Strength:</th>
<th>Quality of evidence:</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Assumptions about underlying values and preferences</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Remarks</th>
</tr>
</thead>
</table>


**Alternative Example**


**Question:** Should multislice spiral computed tomography rather than conventional coronary angiography be used to diagnose coronary artery disease (CAD)?

**Patient or population:** Adults suspected of coronary artery disease

**Settings:** The included trials were conducted in Europe and North America

## Quality assessment for diagnostic accuracy studies – example

<table>
<thead>
<tr>
<th>Outcome</th>
<th>No. of studies</th>
<th>Study design</th>
<th>Factors that may decrease quality of evidence</th>
<th>Final quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives (patients with CAD)</td>
<td>21 studies (1570 pts)</td>
<td>cross-sectional¹</td>
<td>No serious limitations</td>
<td>⊗⊕⊕O moderate</td>
</tr>
<tr>
<td>True negatives (patients without CAD)</td>
<td>21 studies (1570 pts)</td>
<td>cross-sectional¹</td>
<td>No serious limitations</td>
<td>⊗⊕⊕O moderate</td>
</tr>
<tr>
<td>False positives (patients incorrectly classified as having CAD)</td>
<td>21 studies (1570 pts)</td>
<td>cross-sectional¹</td>
<td>No serious limitations</td>
<td>⊗⊕⊕O moderate</td>
</tr>
<tr>
<td>False negatives (patients incorrectly classified as not having CAD)</td>
<td>21 studies (1570 pts)</td>
<td>cross-sectional¹</td>
<td>No serious limitations</td>
<td>⊗⊕⊕O low</td>
</tr>
<tr>
<td>Inconclusive</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

¹ All patients were selected to have conventional coronary angiography and were, therefore, generally presenting with high probability of coronary artery disease (median prevalence in included studies 63.5%, range 6.6-100%)

Summary of findings – example. Assumed pre-test probability (prevalence) was 20%.

### Test findings

<table>
<thead>
<tr>
<th>Pooled sensitivity</th>
<th>0.96 (95% CI: 0.94 to 0.98)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled specificity</td>
<td>0.74 (95% CI: 0.65 to 0.84)</td>
</tr>
</tbody>
</table>

### Consequences

<table>
<thead>
<tr>
<th>Number per 1000¹</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP²</td>
<td>192</td>
</tr>
<tr>
<td>TN²</td>
<td>592</td>
</tr>
<tr>
<td>FP²</td>
<td>208</td>
</tr>
<tr>
<td>FN²</td>
<td>8</td>
</tr>
<tr>
<td>Inconclusive results⁶ ⁷</td>
<td>–</td>
</tr>
<tr>
<td>Cost³</td>
<td>–</td>
</tr>
</tbody>
</table>

¹ all results are given per 1000 patients tested based on the prevalence of 20% and pooled sensitivity and specificity.

² Inconclusive results are either uninterpretable, indeterminate or intermediate test results

³ Important because mandates drugs, angioplasty and stents, bypass surgery.

⁴ Important because spares patients unnecessary interventions associated with adverse effects.

⁵ Important because patients are exposed to unnecessary potential adverse effects from drugs and invasive procedures.

⁶ Important because increase risk of coronary events as a result of patients not receiving efficacious treatment.

⁷ Uninterpretable, indeterminate, or intermediate test results; important because generate anxiety, uncertainty as to how to proceed, further testing, and possible negative consequences of either treating or not treating.

Although the results for these consequences are not reported because they are not exactly known on the basis of the available data, they are important.
APPENDIX 3.2.1A: INTERVIEW PACKAGE 6A – FULL 30-MIN INTERVIEW

Summary of Findings tables for diagnostic test accuracy reviews – Interview Guide

A. Interviewer Checklist
For in person interviews:
- Printed copy of all SoF table formats to be evaluated
- Interview Guide form. Take notes (point-form preferred) in the spaces provided.
- Additional paper to take notes if needed.
- Audio recorder. Test the recorder before each interview (Press REC to record a test message, then press STOP. Press PLAY to make sure the message was recorded clearly and is audible and press ERASE to delete the test message)

<table>
<thead>
<tr>
<th>Test participant No.:</th>
<th>Name:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant name and contact information:</td>
<td>Email/Telephone:</td>
</tr>
<tr>
<td>Location:</td>
<td></td>
</tr>
<tr>
<td>Date:</td>
<td></td>
</tr>
<tr>
<td>Interviewer/Note taker:</td>
<td></td>
</tr>
<tr>
<td>Recorder interview No. and Total Time (e.g. 01, 1h12min):</td>
<td></td>
</tr>
</tbody>
</table>

For interviews by telephone or Skype:
- Make sure you have a land line phone number or a Skype username
- For interviews by Skype, search for and add in the username of the interviewee in advance
- Send in advance a pdf copy of all the SoF table formats to be evaluated
- Confirm that the participant has received the SoF tables by email
- Confirm the date and time of the call, and send an email explaining that he/she needs to have the tables available
- Send a reminder to the participant two days before the meeting
- Interview Guide form
- Additional paper to take notes if needed.
- Audiorecorder for telephone/Recording Software for Skype. Test audio recorder before each interview.

B. Introduction – 1 minute
What we are testing and why:

**Say:** In this user testing interview, we will be evaluating the usability and understanding of Summary of Findings tables, referred to as “SoF” tables for short. Summary of Findings tables are used in the GRADE approach to summarize and present data from systematic reviews. They are intended to facilitate access to the key information of a systematic review. The format that SoF tables should have in diagnostic test accuracy (DTA) reviews has not been fully explored.

We will use your feedback to help determine the ideal presentation of information and the preferred content in SoF tables for diagnostic tests. We really appreciate you giving us a bit of your time.

Participant Consent Statement:

**Say:** The research study has been reviewed by the Hamilton Health Sciences Research Ethics Board. With your permission, the session will be recorded on tape for transcription and erased afterwards. Do you agree to have the interview recorded and for the data collected in the study to be used anonymously in publication?

☐ Yes ☐ No  Notes:

Is it ok to contact you in the future if we have any questions?

☐ Yes ☐ No  Notes:

*As soon as you have finished the introduction, TURN ON AUDIORECORDER.*
C. Background questions – 3 minutes

**Say:** I will first ask a few questions about your background.

1. **Ask:** What is your current position?

2. **Ask:** What is your formal education?: *(e.g. MSc, PhD Edpidemiology, Health Economics, etc.)*

3. **Ask:** We would also like to know about your professional background, and you may fit into more than one of the following categories, so select all that apply.

   **Note to Interviewer:** For interviews with DTA review authors, skip C, for interviews with Clinicians skip B and ask directly about specialty:

   a. Would you identify yourself as a **Researcher:** ☐ Yes ☐ No

   b. Would you identify yourself as a **Health Professional:** ☐ Yes ☐ No

   Please specify *(e.g. profession or clinical specialty):*

   c. Are you an **author of DTA systematic review(s):** ☐ Yes ☐ No

   d. Are there any other roles that you have had that you would identify as relevant to this interview *(e.g. guideline developer) (Specify):*

   e. How long have you been in these roles, in other words how many years of experience do you have overall?: *Years *

**Note to Interviewer:** For authors of DTA reviews, skip to Q6.

**Say:** For the next set of questions, I will ask about your experience with systematic reviews. We have provided an Answer Options Sheet as a guide. Please refer to the **BLUE section** of the sheet to help you choose the appropriate answers.

**Ask:** To what extent do you agree or disagree with the following statement?

4. I am familiar with systematic reviews.
   
   Note that in the introduction I mentioned the topic of diagnostic testing, but here I am referring to systematic reviews on any topic, not DTA reviews specifically.

   ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

   Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly
5. Approximately how many systematic reviews do you read per week, or per month, or per year? You can answer this question using the time frame you prefer:  

Circle time frame used: [per week] [per month] [per year]  

6. Approximately how often do you access the Cochrane Library per week, or per month, or per year? You can also answer this question using the time frame you prefer:  

Circle time frame used: [per week] [per month] [per year]  

7. For what purpose do you read/use systematic reviews?:

8. I am familiar with diagnostic test accuracy (DTA) systematic reviews.  

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat</th>
<th>Neither Agree</th>
<th>Somewhat</th>
<th>Agree</th>
<th>Strongly Agree</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

9. Would you identify yourself as a user of DTA systematic reviews: □ Yes □ No  

If Yes, in what capacity have you used DTA systematic reviews (e.g. to make clinical decisions for patients, to inform policies, to inform guideline developers)? [Alternative question for authors of DTA reviews if have difficulty answering: In what capacity are the DTA reviews that you have produced intended to be used by others?]

10. I am familiar with the GRADE approach.
D. Instructions for Evaluation of SoF Formats – 30 Seconds

Say:
I will present to you alternative formats of the SoF tables with different ways of presenting information. For the questions I will ask you, think out loud, we want to capture your opinion. If you think something is easy or difficult, clear or confusing, please point it out. There are no right or wrong answers, we are not testing your knowledge, we are testing our material.

E. Review SoF Format 1 – Layout 1 – 7 minutes

Say: Please refer to the printout of SoF Format 1 – Layout 1 on Page 1. Review this format of the SoF table for about 2 minutes to grasp the information being presented. Let me know when you are ready. (Note: Time the 2 minutes and notify participant when time is up.)

1. Ask:
   a. What is your overall impression of this SoF table?
   b. Which features of the table helped you to understand the systematic review data displayed?
      Which features were difficult to understand?

Note to Interviewer: If interviewee is unsure provide cue: What do you think about the table overall, do you understand everything the table is presenting? Does it present all the information you think is essential for reporting and making conclusions about the test?

2. For Clinicians: Would you recommend or use this diagnostic test?
   For DTA review authors: What would your conclusion about the test be given the results presented in the SoF table?

   Yes □   No □   Need More Information □

   Why or why not? What information did you use to come to your conclusion?:
**Note to Interviewer:** This question is to gauge the overall understanding of the table. We would like to know what components of the table the user looked at and considered in providing their answer. We want to see if some users for example find only sensitivity and specificity sufficient and do not consider individual DTA results or impact on patient-important outcomes. ‘Need more information’ could refer to information on purpose of the test, place of the test in the test-treatment strategy, or downstream consequences of test results or treatment.

3. Did you initially read the content in the footnotes when reviewing the table?:

Yes □  No □  Why or Why Not?:

**Note to Interviewer:** If answer above is yes, and the answer addresses the follow-up to Q4 below, then ask only the question on the 7-point scale for Q4, if not addressed then ask the follow-up questions.

**Say:** If you did not initially read the footnotes, have a look through them now. For the next question, referring to the RED section of the answer sheet, to what extent do you agree or disagree with this statement.

4. The content in the footnotes helped me to better understand the data.

☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disagree</td>
<td>Disagree</td>
<td>or Disagree</td>
<td>Agree</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Why? What information were you looking for in the footnotes to help with understanding and interpretation of the data in the table? Did the footnotes provide the information you expected to see?
F. Review SoF Format 1 – Layout 2 – 7 minutes
Say: We have also prepared a second layout for this SoF format. Please refer to the printout of SoF Format 1 – Layout 2 on Page 2. Review this layout for about 1 minute. This layout uses the same systematic review example as the previous one.

Say: Referring to the GREEN section of the answer sheet, to what extent do you agree or disagree with these statements.

1. The purpose of the two different prevalence values presented was clear to me.

   ![Strongly Disagree] ![Disagree] ![Somewhat or Disagree] ![Neither Agree] ![Somewhat Agree] ![Agree] ![Strongly Agree]

   Why? For what reason did you think two prevalence estimates are presented in this SoF table layout?

2. The presentation of various prevalence estimates is important for readers of the systematic review to help make appropriate conclusions and demonstrate diagnostic test accuracy results in different clinical settings.

   ![Strongly Disagree] ![Disagree] ![Somewhat or Disagree] ![Neither Agree] ![Somewhat Agree] ![Agree] ![Strongly Agree]

   For Clinicians: How do you think the presentation of different prevalence values affects making conclusions about the test? How might your conclusion about the test change when presented with the different prevalence estimates?

   For DTA authors: How do you think the presentation of different prevalence values affects making appropriate conclusions about the test by the users of the systematic review?

3. Do you think that an example clinical scenario should be presented in a SoF table to accompany the prevalence value estimate?

   Yes, present clinical scenario □  No, clinical scenario is unnecessary □

   If NO, why do you think the scenario is not necessary?
4. Could the clinical scenario information be placed in the footnotes instead of inside the table?
   
   Yes, place in footnotes □  No, leave in table □  Other answer □
   
   What is the reason for your preference?

Say: I will now ask about your overall preference between the two layouts we have reviewed.

5. Do you prefer Layout 1 (Pg 1) with one prevalence estimate and no clinical scenario, or Layout 2 (Pg 2) with two prevalence estimates and the clinical scenarios?:

   Layout 1 □  Layout 2 □

   a. What is the reason for your preference?
   b. Would you change anything about the overall presentation or look of either layout?

G. Review SoF Format 2 – 2 minutes

Say: We will now review an alternative format of the SoF table. Please refer to the printout of SoF Format 2 on Page 3. Review this format for about 30 seconds to familiarize yourself with it. This SoF format uses the same systematic review example as the first format.

1. Ask:
   a. What is your overall impression of this format?
   b. How does the arrangement of the rows in the table by test-positive and test-negative (versus according to sensitivity and specificity as in the first format) affect your interpretation of the test results?
   c. Does this arrangement help or hinder using the SoF table for making conclusions about the test?
H. Review SoF Format 3 – Layouts 1-3 – 5 minutes

Say: We will now review a third alternative format of the SoF table. This SoF table is a bit different from the first two formats you have reviewed, but uses the same systematic review example. Please refer to the printout of SoF Table Format 3 – Layout 1 on Page 4 and review it for about 1 minute. Let me know when you are ready.

1. Ask:
   a. What is your overall impression of this format?
   b. How do the features of this table affect your understanding and interpretation of the systematic review results when compared to the features in the previous two SoF formats?
   c. Which features of this table are more and less helpful? Is there any information that is missing?

2. Ask:
   a. How do you think the presentation of post-test probability and likelihood ratios affects making appropriate conclusions about the test by the users of the systematic review?
   b. Do the likelihood ratios provide important information, or is post-test probability the more important piece of information in this table?

Say: We have also prepared two other layouts for this SoF format. Please refer to the printouts of SoF Format 3 – Layout 2 on Page 5 and SoF Format 3 – Layout 3 on Page 6. I will now ask about your overall preference between the three layouts.

3. Do you prefer Layout 1 (Pg 4) with no clinical scenarios, Layout 2 (Pg 5) with the clinical scenarios, or Layout 3 (Pg 6) with the clinical scenarios and the likelihood ratios placed in the header of the table instead of as a column inside the table?

   Layout 1 ☐    Layout 2 ☐    Layout 3 ☐

   a. What is the reason for your preference?
   b. Would you change anything about the overall presentation or look of these layouts?
   c. If you prefer the layouts with the clinical scenario, where should the clinical scenario be placed in this SoF format?
I. Review SoF Format 4 – 2 minutes
Say: We will now review the last alternative format. Please refer to the printout of SoF Table Format 4 on Page 7 and review it for about 30 seconds. Let me know when you are ready.

Say: Referring to the Yellow section of the answer sheet, to what extent do you agree with the following statement?

1. The table provides sufficient information for understanding the systematic review results and making conclusions about the diagnostic test.

   ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐ ☐
   Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
   Disagree Disagree or Disagree Agree Agree

2. Ask:
   a. What is your overall impression of this format?
   b. Is this table sufficient to present DTA systematic review results? Why or why not?
   c. What information is missing that is essential for understanding the results and making an appropriate conclusion about the diagnostic test?

J. Overall Preferences – 2 minutes
Say: Having viewed all the formats, I will now ask you about your overall preferences.

1. Ask:
   a. Overall, do you prefer prefer SoF Format 1 (Pg 1 & 2, with rows arranged by sensitivity and specificity), Format 2 (Pg 3, with rows arranged by test-positive and test-negative), Format 3 (Pgs 4-6, with presentation of post-test probability and likelihood ratios), or Format 4 (Pg 7, with presentation of sensitivity and specificity).
   b. Rank the other formats in order of preference. You can choose not to rank a certain format if you feel that it is not useful at all and should not be used.

Format 1 (by sens/spec): ☐ ☐ Format 2 (by test +/-): ☐ ☐
What is the reason for your overall preference and the rankings you assigned? Which features of the tables were you taking into consideration in deciding your preferences?

Say: Considering the SoF tables overall and referring to the Yellow section of the answer sheet, to what extent do you agree with the following statement?

2. Presenting the SoF table as part of a DTA systematic review is a helpful way of reporting the outcomes.

   \[\begin{array}{cccccccc}
   \text{Strongly Disagree} & \text{Disagree} & \text{Somewhat Disagree} & \text{Neither Agree} & \text{Somewhat Agree} & \text{Agree} & \text{Strongly Agree} \\
   \end{array}\]

Why or why not? Does this table represent what you would expect from a SoF to help with interpretation of the results and making appropriate conclusions about a diagnostic test? Do you think a SoF table should be included in a DTA review?

K. Conclusion – 1 minute

1. Ask: Do you have any final comments that you would like to make about the content and presentation of information in SoF tables, keeping in mind that the goal of this interview was to identify problematic areas for users and to determine the usefulness and usability of the tables for presentation of DTA systematic review data?

Say: That marks the end of our interview. Thank you very much for your participation, we really appreciate you dedicating your time to this.
### APPENDIX 3.2.1B: GALACTOMANNAN ELISA FOR THE DIAGNOSIS OF INVASIVE ASPERGILLOSIS

**Patients or population:** immunocompromized patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum | **Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria | **Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence</td>
<td>20 per 1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Sensitivity (95% CI):** 0.64 (0.50 to 0.77)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td>7 per 1000 (4 to 10 per 1000)</td>
<td>2777 (18 studies)</td>
<td>⭐⭐⭐⭐⭐ High</td>
</tr>
</tbody>
</table>

**Specificity (95% CI):** 0.95 (0.91 to 0.97)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True negatives</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td>49 per 1000 (30 to 89 per 1000)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

4. Estimate of prevalence of IA was based on the median and range of values in included studies.

5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
# Galactomannan ELISA for the diagnosis of invasive aspergillosis

Patients or population: immunocompromized patients, mostly hematology patients¹
Settings: mainly hematology or cancer departments, mainly inpatients
New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum² | Cut-off value: 1.5 ODI
Reference Test: composite of EORTC/MSG clinical and histological criteria³ | Threshold: Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested⁴ (95% CI)</th>
<th>Prevalence 400 per 1000⁵: Which is typically seen in adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity (95% CI):</td>
<td>0.64 (0.50 to 0.77)</td>
<td>Prevalence 20 per 1000⁵: Which is typically seen in adults undergoing transplant, no neutropenic patients.</td>
<td>2777 (18 studies)</td>
<td>⊕⊕⊕⊕ High⁶</td>
</tr>
<tr>
<td>True positives</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td>282 per 1000 (220 to 339 per 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td>7 per 1000 (4 to 10 per 1000)</td>
<td>158 per 1000 (101 to 220 per 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity (95% CI):</td>
<td>0.95 (0.91 to 0.97)</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td>532 per 1000 (510 to 543 per 1000)</td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>931 per 1000 (901 to 960 per 1000)</td>
<td>532 per 1000 (510 to 543 per 1000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td>49 per 1000 (30 to 89 per 1000)</td>
<td>28 per 1000 (17 to 50 per 1000)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:
¹ One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
² A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
³ A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
⁴ Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).
⁵ Estimates of prevalence of IA were based on the median and range of values in included studies.
⁶ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### Galactomannan ELISA for the diagnosis of invasive aspergillosis

**Patients or population:** immunocompromized patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum

**Reference Test:** composite of EORTC/MSG clinical and histological criteria

**Quality:** not downgraded based on the GRADE criteria.

**Pooled Sensitivity (95% CI):** 0.64 (0.50 to 0.77) | **Pooled Specificity (95% CI):** 0.95 (0.91 to 0.97)

<table>
<thead>
<tr>
<th>Test result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test-Positive</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>13 per 1000 (9 to 15 per 1000)</td>
<td>282 per 1000 (220 to 339 per 1000)</td>
<td></td>
</tr>
</tbody>
</table>
| False positives | 49 per 1000 (30 to 89 per 1000) | 28 per 1000 (17 to 50 per 1000) | 2777 (18 studies) | ⊕⊕⊕⊕ High

| **Test-Negative** | | | |
| True negatives | 931 per 1000 (901 to 960 per 1000) | 532 per 1000 (510 to 543 per 1000) | |
| False negatives | 7 per 1000 (4 to 10 per 1000) | 158 per 1000 (101 to 220 per 1000) | |

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

4. Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).

5. Estimates of prevalence of IA were based on the median and range of values in included studies.

6. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
Galactomannan ELISA for the diagnosis of invasive aspergillosis

Patients or population: immunocompromized patients, mostly hematology patients.
Settings: mainly hematology or cancer departments, mainly inpatients
New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum. Cut-off value: 1.5 ODI
Reference Test: composite of EORTC/MSG clinical and histological criteria. Threshold: Proven or probable invasive aspergillosis

Pooled Sensitivity (95% CI): 0.64 (0.50 to 0.77) | Pooled Specificity (95% CI): 0.95 (0.91 to 0.97)

<table>
<thead>
<tr>
<th>Pre-test probability</th>
<th>Test result</th>
<th>Post-test probability (95% CI)</th>
<th>Likelihood ratio (LR) (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Probability</td>
<td>+</td>
<td>21% (17% to 26%)</td>
<td>+ LR = 13 (10 to 17)</td>
<td>2777 (18 studies)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>1% (1% to 1%)</td>
<td>- LR = 0.38 (0.27 to 0.54)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Probability</td>
<td>+</td>
<td>91% (89% to 93%)</td>
<td>+ LR = 13 (10 to 16)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>23% (22% to 24%)</td>
<td>- LR = 0.38 (0.35 to 0.41)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:
1 One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4 Pre-test probability of IA was selected based on the median and range of prevalence values in included studies.
5 Percentages are rounded to the nearest integer.
6 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### Galactomannan ELISA for the diagnosis of invasive aspergillosis

Patients or population: immunocompromized patients, mostly hematology patients

Settings: mainly hematology or cancer departments, mainly inpatients

New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum | Cut-off value: 1.5 ODI

Reference Test: composite of EORTC/MSG clinical and histological criteria | Threshold: Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Pooled Sensitivity (95% CI): 0.64 (0.50 to 0.77)</th>
<th>Pooled Specificity (95% CI): 0.95 (0.91 to 0.97)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pre-test probability</strong></td>
<td>Test result</td>
</tr>
<tr>
<td>Low Probability</td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in adults undergoing transplant, no neutropenic patients</td>
<td>+</td>
</tr>
<tr>
<td>–</td>
<td>1% (1% to 1%)</td>
</tr>
<tr>
<td>High Probability</td>
<td></td>
</tr>
<tr>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</td>
<td>+</td>
</tr>
<tr>
<td>–</td>
<td>23% (22% to 24%)</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:

1 One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

2 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

3 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

4 Pre-test probability of IA was selected based on the median and range of prevalence values in included studies.

5 Percentages are rounded to the nearest integer.

6 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
Galactomannan ELISA for the diagnosis of invasive aspergillosis

Patients or population: immunocompromized patients, mostly hematology patients¹
Settings: mainly hematology or cancer departments, mainly inpatients
New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum² | Cut-off value: 1.5 ODI
Reference Test: composite of EORTC/MSG clinical and histological criteria³ | Threshold: Proven or probable invasive aspergillosis

Pooled Sensitivity (95% CI): 0.64 (0.50 to 0.77) | Pooled Specificity (95% CI): 0.95 (0.91 to 0.97)

<table>
<thead>
<tr>
<th>Pre-test probability⁴</th>
<th>Test result</th>
<th>Post-test probability⁵ (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Low Probability</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td>+</td>
<td>21% (17% to 26%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in adults undergoing transplant, no neutropenic patients.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>1%</td>
<td>1% (1% to 1%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>High Probability</strong></td>
<td></td>
<td></td>
<td>2777 (18 studies)</td>
<td>High⁶</td>
</tr>
<tr>
<td>44%</td>
<td>+</td>
<td>91% (89% to 93%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which is typically seen in adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>−</td>
<td>23%</td>
<td>23% (22% to 24%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:
¹ One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
² A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
³ A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
⁴ Pre-test probability of IA was selected based on the median and range of prevalence values in included studies.
⁵ Percentages are rounded to the nearest integer.
⁶ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
Galactomannan ELISA for the diagnosis of invasive aspergillosis

Patients or population: immunocompromized patients, mostly hematology patients¹
Settings: mainly hematology or cancer departments, mainly inpatients
New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum² | Cut-off value: 1.5 ODI
Reference Test: composite of EORTC/MSG clinical and histological criteria³ | Threshold: Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test property</th>
<th>Summary estimate (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.64 (0.50 to 0.77)</td>
<td>2777 (18 studies)</td>
<td>⬛️⬜⬜⬜ High⁴</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.95 (0.91 to 0.97)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:
¹ One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
² A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
³ A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
⁴ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
APPENDIX 3.2.2A: INTERVIEW PACKAGE 4A – FULL INTERVIEW

Summary of Findings tables for diagnostic test accuracy reviews – Interview Guide

A. Interviewer Checklist

For in person interviews:
- Printed copy of all SoF table formats to be evaluated
- Interview Guide form. Take notes (point-form preferred) in the spaces provided.
- Additional paper to take notes if needed.
- Audiorecorder. Test the recorder before each interview (Press REC to record a test message, then press STOP. Press PLAY to make sure the message was recorded clearly and is audible and press ERASE to delete the test message)

<table>
<thead>
<tr>
<th>Test participant No.:</th>
<th>Name: __________________________</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>__________________________</td>
</tr>
<tr>
<td>Location:</td>
<td></td>
</tr>
<tr>
<td>Date:</td>
<td></td>
</tr>
<tr>
<td>Interviewer/Notetaker:</td>
<td></td>
</tr>
<tr>
<td>Recorder interview No.</td>
<td></td>
</tr>
<tr>
<td>and Total Time</td>
<td></td>
</tr>
<tr>
<td>(e.g. 01, 1h12min):</td>
<td></td>
</tr>
</tbody>
</table>

For interviews by telephone or Skype:
- Make sure you have a land line phone number or a Skype username
- For interviews by Skype, search for and add in the username of the interviewee in advance
- Send in advance a pdf copy of all the SoF table formatss to be evaluated
- Confirm that the participant has received the SoF tables by email
- Confirm the date and time of the call, and send an email explaining that he/she needs to have the tables available
- Send a reminder to the participant two days before the meeting
- Interview Guide form
- Additional paper to take notes if needed.
- Audiorecorder for telephone/Recording Software for Skype. Test audio recorder before each interview.
**Note to interviewer:** All the introductory material and background questions (Sections B-D) should take a total of 10 minutes.

**B. Introduction – 4 minutes**

**Say:** You have been asked to participate in a user testing interview to help provide feedback on material that is intended to help with presenting information of diagnostic test accuracy (DTA) reviews. We really appreciate you giving us a bit of your time.

**What we are testing and why:**

**Say:** We will be evaluating the usability and usefulness of Summary of Findings tables. I will refer to these as “SoF” tables for short throughout the interview. Summary of Findings tables are used in the GRADE approach to summarize and present outcome data from systematic reviews and are now a key element of Cochrane Intervention Systematic Reviews. They are intended to accompany a systematic review and be a tool to support transparent and facilitate access to the key information of a systematic review. The format that SoF Tables should have in DTA reviews is incompletely explored.

You are one of about 20 people from a variety of backgrounds that represent users of systematic reviews, including clinicians, researchers, guideline developers and decision makers, that we are collecting feedback from. We will use your feedback to help determine the preferable content and presentation of information in Summary of Findings tables for diagnostic tests with the aim to maximize their usability and usefulness for authors and users like you.

**What the interview session will consist of:**

**Say:** For this interview, I will first ask some questions about your background and experience with systematic reviews. Then, I will present to you alternative formats of the SoF table with different ways of presenting information in the tables. I will ask you questions about these different formats. Questions will be in the form of statements, with which you will be asked to agree or disagree, open ended questions seeking your comments, or questions asking about your preferences. We have provided you with an Answer Options sheet which will help with choosing the appropriate answer for some questions.

**Participant Consent:**

**Say:** The research study has been reviewed by the Hamilton Health Sciences Research Ethics Board. I have provided you with an introduction to the study and I have informed you about the study purpose and what is being asked of you as a participant. With your permission, the session will be recorded on tape. The recording will only be used for transcribing the interview and will be erased afterwards. Do you agree to have the interview recorded and for the data collected to be used for the research study and to be summarized anonymously in publication?

☐ Yes  ☐ No  Notes:____________________________

May we also contact you in the future in case we need to follow up with you about the interview, for example if we need to clarify any of your comments or if any of your comments did not get recorded properly?
□ Yes □ No Notes:___________________________

Ask: Do you have any questions before we proceed?

> As soon as you have finished the introduction, TURN ON AUDIORECORDER.

C. Background questions – 5 minutes

Say: I will first ask a few questions about your background.

12. Ask: What is your current position?: __________________________

13. Ask: What is your formal education?: __________________________
   (e.g. MSc, PhD Epidemiology, Health Economics, etc.)

14. Ask: We would also like to know about your professional background, and you may fit into more than one of the following categories, so select all that apply. Of the following 3 options:

   f. Would you identify yourself as a Researcher: □ Yes □ No
   g. Would you identify yourself as a Health Professional: □ Yes □ No
      Please specify (e.g. profession or clinical specialty): __________________________

   h. Are you an author of DTA systematic review(s): □ Yes □ No
   i. Are you a user of DTA systematic review(s): □ Yes □ No
      If Yes, in what capacity have you used DTA systematic reviews (e.g. to make clinical decisions for patients, to inform policies, to inform guideline developers):? ______

   How many times have you participated in this role: ________________

   j. Are there any other roles that you have had that you would identify as relevant to this interview (Specify): __________________________

   How long have you been in these roles, in other words how many years of experience do you have overall?: ______ Years
**Say:** For the next set of questions, I will ask about your experience with systematic reviews. These questions will mainly be in the form of statements for which I will ask whether you agree or disagree on a 7-point scale. Please refer to the Answer Options Sheet we have provided you, and refer to the BLUE section to help you choose the appropriate answers.

**Ask:** To what extent do you agree or disagree with the following statements.

15. **Say:** I am familiar with systematic reviews. Based on the information in the BLUE section of the answer sheet, the two endpoints for this question would be as follows. Someone who strongly disagrees is someone who has never used or read systematic reviews, or does not have any methods training in systematic reviews. Someone who strongly agrees is someone who frequently uses systematic reviews, or produces systematic review, or has methods training in systematic reviews. Note that in the introduction I mentioned the topic of diagnostic testing, but here I am referring to systematic reviews on any topic, not on diagnosis specifically. **Note to Interviewer:** If needed, read out the 7 answer options.

![Answer Options]

16. Approximately how many systematic reviews do you read per week, or per month, or per year? You can answer this question using the time frame you prefer:__________

   *Circle time frame used:* [per week] [per month] [per year]

17. Approximately how often do you access the Cochrane Library per week, or per month, or per year? You can also answer this question using the time frame you prefer:_____

   *Circle time frame used:* [per week] [per month] [per year]

18. For what purpose do you read the systematic reviews?: ____________________________

   __________________________________________________________________________

   **Note to interviewer:** If the participant has difficulty answering, provide the cue: Someone might read a systematic review for their own learning, or to make policy decisions, or to make recommendations for clinical practice guidelines or coverage decisions, or to make clinical decisions for a patient.

**Say:** For the next questions I will ask again to what extent do you agree with the statements. For these, refer also to the BLUE section of the answer sheet to help you choose the appropriate answers.

19. I am familiar with diagnostic test accuracy (DTA) systematic reviews. Please note that from now on I will refer to diagnostic test accuracy as “DTA” to simplify things. For this question, someone who strongly disagrees is someone who has never used or read DTA systematic reviews, or does not have
any methods training in DTA systematic reviews. Someone who strongly agrees is someone who frequently uses DTA systematic reviews, or produces DTA systematic reviews, or has methods training in DTA systematic reviews.

Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
Disagree Disagree or Disagree Agree Agree

20. I am familiar with the GRADE approach. Referring again to the BLUE section of the answer sheet, someone who strongly disagrees is someone who has never heard about, read, or used the GRADE approach. Someone who strongly agrees is someone who has received training in applying GRADE and has used the GRADE approach in practice.

Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
Disagree Disagree or Disagree Agree Agree

21. I am familiar with using Summary of Findings (SoF) tables. For this question, someone who strongly disagrees is someone who has never read or used SoF tables. Someone who strongly agrees is someone who frequently uses SoF tables in decision-making or has produced SoF tables.

Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
Disagree Disagree or Disagree Agree Agree

D. Instructions for Evaluation of SoF Formats – 30 seconds

Say:

For the following part of the interview we aim to find what works well and what doesn’t work well, both regarding content, use of language or terminology, as well as presentation and formatting of the tables. We want to know what are any major problems for all users regardless of experience and are there any striking things that don’t make sense or are problematic. Think out loud, we want to capture your opinion. There are no right or wrong answers, we are not testing your knowledge, we are testing our material.

From our experience, we are fairly certain that things you find difficult to understand, other people will also find difficult. When we ask for your comments describe your interpretation of the tables, if you are unsure or surprised by anything, if there are things you don’t understand, just say “I don’t know
what this means...”, if you think something is easy or difficult, clear or confusing, please point it out. This is the information we can use to improve the summary of findings tables.

My role is to ask the interview questions. But, since it is your opinion we are interested in, I will be otherwise saying as little as possible during this part of the interview.

**[Layer 1: (mandatory) assessing different formats of SoF tables to summarise DTA SR results]**

**E. Review Single Test SoF Format 1 – 7 minutes**

_Say:_ First, please refer to the printout of SoF Format 1 – Layout 1 on Page 1. Review this format of the SoF table for about 2 minutes to grasp the information being presented. Let me know when you are ready. You will give us your thoughts by answering the following question, and we will go more in depth later in the interview regarding these topics:

**Note to interviewer: Time the 2 minutes and notify participant when time is up.**

5. **Ask:** What is your overall impression of this SoF table format?
   **Note to interviewer: If interviewee is unsure of how to answer provide cue:** What do you think about the table overall, do you understand everything the table is presenting, what do you think about it in terms of content and presentation/aesthetics?

6. **Ask:** Which features of the table helped you to better understand the systematic review data displayed? Are there any parts of the table that are not necessary?

7. **Ask:** Would you recommend or use this diagnostic test? Why or why not:
**Note to Interviewer:** This question is to gauge the overall understanding of the table. We would like to know what components of the table the user looked at and considered in making the decision. First see if the interviewee offers up the option of ‘insufficient or need more information’ as this is what we think should be the response, but we want to see if some users for example find only sensitivity and specificity sufficient and do not identify needing to know impact on downstream patient-important outcomes.

Yes □  No □  Need More Information □  Why?

---

**Say:** For the following set of questions I will ask to what extent do you agree with these statements on a 7-point scale. The answer options for each question are as before; Strongly Disagree, Disagree, Somewhat Disagree, I am not sure, Somewhat Agree, Agree, Strongly Agree. Please refer to the GREEN section of the answer sheet now to help you choose the appropriate answers.

**Note to interviewer:** Repeat the answer options for each question (“The answer options again are Strongly Disagree, Disagree, Somewhat Disagree...”) until not necessary.

8. The purpose of the different prevalence values presented was clear to me:

<table>
<thead>
<tr>
<th></th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes □</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Why?: ________________________________

_________________________________________________________________

9. The reason for the assigned quality of evidence scores assigned in the SoF table was clear to me:

**Note to interviewer:** If interviewee has difficulty answering this question provide cue: “Did you understand why the quality of evidence was rated as high?”

<table>
<thead>
<tr>
<th></th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes □</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Why?: ________________________________

_________________________________________________________________
10. The purpose of the footnotes was clear to me:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: ________________________________________________

______________________________________________________________________________________________

11. **Say:** For the following question the answer options are Yes or No and I will ask you the reason for the answer.

Did you read the content in the footnotes?:

Yes □       No □       Why or Why Not?:

Say: For the next 2 questions, refer to the RED section of the answer sheet now and I will ask to what extent do you agree with these statements.

12. The content in the footnotes helped me to better understand the data:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: ________________________________________________

______________________________________________________________________________________________

13. The content in the comments column helped me to better understand the data:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: ________________________________________________
Say: For the following 3 questions, refer to the YELLOW section of the answer sheet. For these questions we are simply asking about the extent of your agreement with the statement.

14. Presenting the SoF table as part of a DTA systematic review is a helpful way of reporting the outcomes:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: __________________________________________________________

15. The density of information presented in this SoF table format is appropriate:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: __________________________________________________________

16. The information presented in this SoF table format is easy to read:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: __________________________________________________________

Say: We have also proposed an alternative look for this SoF format which I will ask you some questions about. Please refer to the printout for SoF Format 1 – Layout 2 on Page 2.

17. Ask: Of the two layouts which do you prefer? Layout 1 where the sensitivity and specificity information is presented similarly as other information in the header in the same font, or Layout 2 where the sensitivity and specificity information is boxed from the rest of the information in the header and the number of participants and studies is also placed in the header instead of in a column?:

Layout 1 □    Layout 2 □    Any comment?:________________________
F. Instructions for Evaluation of Components – 20 seconds

Say: In addition to evaluating the SoF table formats overall, we are also interested in gathering detailed feedback on specific parts of the table. For these next questions, we will be focusing on one component of the table.

G. Quality of Evidence Rating Component – 4 minutes

Say: The component we will evaluate is the Quality of Evidence Rating. Please refer to the printout of SoF Table Component: Quality of Evidence Rating on Page 3. This SoF table uses a different systematic review example with different results, so review the table for about a minute, paying attention specifically to the Quality of the Evidence Column.

Say: For the first question I will ask to what extent do you agree with these statements on the 7-point scale. Please refer to the RED section of the answer sheet to help you choose the appropriate answer.

3. The brief reason provided in the table for the quality of evidence rating helps me to better understand the assigned grade.

□ □ □ □ □ □ □ □
Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
Disagree or Disagree Agree

Why?: ______________________________________________________
________________________________________________________________
________________________________________________________________

Say: For the next question, refer to the YELLOW section of the answer sheet.

4. SoF table users will be more likely to read the footnotes if a brief reason for the quality of evidence rating is provided in the Quality of Evidence Column.

□ □ □ □ □ □ □ □
Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Agree
Disagree or Disagree Agree

Why?: ______________________________________________________
________________________________________________________________
________________________________________________________________
**Say:** I will now ask you a few open ended questions about the quality of evidence rating.

5. **Ask:** How detailed should the explanation of the quality of evidence rating in the footnotes be? What information would you like to have presented in the explanation?

6. **Say:** The two footnotes in this table referred to the same reasons for downgrading quality of evidence, but one footnote was more detailed than the other. Given what you mentioned above, we would like to confirm your preference. Do you prefer a more detailed explanation about the quality of evidence rating, with specific information about the studies, for example a statement explaining that 65% of the studies reported blinding or how wide the confidence intervals were, as is demonstrated in Footnote #1. Or, do you prefer a more minimal explanation for the quality of evidence rating, as is demonstrated in Footnote #2?

   Detailed (Footnote 1) □    Minimal (Footnote 2) □    Any comment?:

   __________________________________________________________

   __________________________________________________________

7. **Ask:** Would you change anything about the presentation of the brief reason provided in the Quality of Evidence column inside the table? Was this a helpful feature of the table?

8. **Ask:** The rationale for the GRADE quality of evidence rating may not be found anywhere else in a systematic review except in the SoF table footnotes. Should the information be repeated elsewhere, and if so, where?
H. Review Single Test SoF Format 2 – 7 minutes

Say: We will now return to reviewing alternative formats of the SoF table. Please refer to the printout of SoF Format 2 – Layout 1 on Page 4. Review this format for about 1 minute to familiarize yourself with it. This SoF format uses the same systematic review example as the first format you reviewed. Let me know when you are ready.

Say: For the following questions I will ask to what extent do you agree with these statements on the 7-point scale. Please refer to the RED section of the answer sheet to help choose the appropriate answer for the first question.

2. Presenting the sensitivity and specificity test properties as rows in the table helped me to better understand the data:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Why?: __________________________________________

___________________________________________________

Say: For the following 3 questions, refer to the YELLOW section of the answer sheet to help choose the appropriate answer.

3. Presenting the sensitivity and specificity as rows in the table as opposed to inside the header of the table improves accessibility of the systematic review results:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Why?: __________________________________________

___________________________________________________

4. The density of information presented in this SoF table format is appropriate:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree or Disagree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Why?: __________________________________________
5. The information in this SoF table format is easy to read:

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat Disagree</th>
<th>Neither Agree</th>
<th>Somewhat Agree</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: ____________________________

______________________________

Say: I will now ask a few open ended questions seeking your thoughts about SoF format 2. Remember to think aloud so we can capture your opinion:

6. Ask: What is your overall impression of this SoF table format? Are sensitivity and specificity properties necessary in a SoF table at all? Why or why not? Does having this information inside the table help you make a link between the sensitivity and specificity and the test results?

Sens/Spec are Necessary in a SoF □ Sens/Spec are Not Necessary in a SoF □

Say: We have also proposed an alternative layout for this SoF format. Please refer to the printout for SoF Format 2 – Layout 2 on Page 5.

7. Ask: Consider the presentation order of the columns in SoF Format 2. Do you prefer for the TP, FN, TN, FP test results to be presented on the left side of the table, adjacent to the sensitivity and specificity column, as in Layout 2 (Pg. 5), or do you prefer Layout 1 (Pg. 4) where the test result columns are on the right side of the table adjacent to the comments column? Please compare the two layouts on Pages 4 and 5 to help visualize the difference.

Test Result Columns on Left (Layout 2) □ Columns on Right (Layout 1) □

Any comment?____________________________________________________

_______________________________________________________________
Say: I will now ask you about your overall preference between SoF Format 1 that we first reviewed on Page 1 and 2, and SoF Format 2 that we just reviewed now.

8. **Ask:** Do you prefer SoF Format 1 (on Pg. 1 and 2) with sensitivity and specificity information in the header or SoF Format 2 (on Pg. 4 and 5) with sensitivity and specificity as rows inside the table?

   SoF Format 1 ☐  SoF Format 2 ☐  Any comment?:_______________________

   ______________________________________________________________________

   ______________________________________________________________________

I. Review Single Test SoF Format 3 (SoR Table) – 3 minutes

**Say:** We will now review a third alternative format of the SoF table. This SoF table is a bit different from the first two formats you have reviewed. Please refer to the printout of SoF Table Format 3 on Page 6. Review this format for about 1 minute. Let me know when you are ready to answer the questions about this format.

**Say:** I will now ask a few open ended questions seeking your thoughts about SoF format 3.

4. **Ask:** What is your overall impression of this SoF table format?

**Box:**

5. **Ask:** Does this SoF table format have specific features or does it present information that you find more helpful to *understanding the systematic review results* when compared to the features in the previous two SoF formats? Which features were more helpful?

**Box:**
Is there something missing?

Say: I will now ask you about your overall preference between this SoF format, format 3, and the previous two SoF formats, 1 and 2, we reviewed.

6. Ask: Do you prefer SoF format 3 (Pg. 6) over SoF formats 1 (Pg. 1) and 2 (Pg. 4)?

Prefer SoF Format 3 □ Prefer SoF Format 1 or 2 □ Any comment?:

__________________________

J. Prevalence Value Component – 5 minutes

Say: We would like to focus on a specific component of the SoF table. The component we will evaluate now is the prevalence values presented in the SoF tables. Please refer to the printout of SoF Table Component: Prevalence – Layout 1 on Page 7. Review the SoF table, noting the example clinical scenario provided, and let me know when you are ready.

Say: For the following questions I will ask to what extent you agree with these statements on the 7-point scale. For the first question, please refer to the RED section of the answer sheet to help choose the appropriate answer.

1. The presentation of a clinical scenario is valuable in helping with the interpretation of the different prevalence values:

   □ □ □ □ □ □ □

   Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly

   Disagree or Disagree Agree Agree

Why?: ________________________________

__________________________

Say: For the next question, please refer to the YELLOW section of the answer sheet to help choose the appropriate answer.
2. The presentation of various prevalence estimates is important for decision making as it helps to demonstrate the diagnostic tests accuracy outcomes in different clinical settings

<table>
<thead>
<tr>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Somewhat</th>
<th>Neither Agree</th>
<th>Somewhat</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
</table>

Why?: _______________________________

______________________________

Say: I will also ask you a few open ended questions about the prevalence component.

3. **Ask:** Do you think that different prevalence estimates should be presented in a SoF table? Why or why not?

4. **Ask:** Ideally, how many estimates should be presented?

5. **Ask:** What do you think should be the source for obtaining the prevalence value estimates, and what should be considered when choosing the prevalence values to present?

6. **Ask:** Do you think that an example clinical scenario should be presented in a SoF table to accompany the prevalence value estimate?
   - Yes, present clinical scenario □
   - No, clinical scenario is unnecessary □

If you answered “yes” what should be the source for the clinical scenario that is presented?
Say: We have also proposed alternative layouts for presenting the prevalence value estimates and would like to ask you about your preferences. Please refer to the printouts of SoF Table Component: Prevalence – Layout 2 Table 1 on Page 8 and SoF Table Component: Prevalence – Layout 2 Table on Page 9.

7. Ask: Do you prefer for the different prevalence value estimates to be presented in a single table, as in Layout 1, or do you prefer the presentation of separate tables for different prevalence values, as in Layout 2?

Layout 1 (Single Table) □       Layout 2 (Separate Tables) □

Ask: What is the reason for your preference?:______________________________

____________________________________________________________________

____________________________________________________________________

8. Ask: Finally, what is your preference for the labelling of the prevalence value estimate, should it be called:

   Prevalence □   Pre-test Probability □   Initial Probability □   Baseline Risk □

   Or, Other □ (specify): ________________________________

   Any comment?:________________________________________

   ___________________________________________________________________

[This can be the end of the interview if time only allows for part of the interview]

K. Review Comparative Test SoF Format 1 – 3 minutes

Say: We will now move on to reviewing a SoF table format for comparative diagnostic tests. These SoF table formats present data from DTA systematic reviews in which two tests are compared against a reference standard. Please refer to the printout of Comparative Test SoF Table Format 1 on Page 10. Review this format for about 1 minute. Let me know when you are ready to give us your thoughts by answering the following questions.
1. **Ask:** What is your overall impression of this SoF table format?

2. **Say:** For the following questions I will ask to what extent do you agree with these statements on the 7-point scale. For the first 3 questions, refer to the RED section of the answer sheet to help choose the appropriate answer.

   2. The presentation of the absolute difference helped me to better understand the data:

      □  □  □  □  □  □  □  □
      Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Disagree Somewhat or Disagree Agree Agree

      Why?: _____________________________________________

      _____________________________________________

   3. The confidence intervals presented in the table are easy to understand:

      □  □  □  □  □  □  □  □  □
      Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Disagree Somewhat or Disagree Agree Agree

      Why?: _____________________________________________

      _____________________________________________

   4. Presenting confidence intervals in the SoF table is helpful for decision making about the diagnostic tests:

      □  □  □  □  □  □  □  □  □
      Strongly Disagree Somewhat Neither Agree Somewhat Agree Strongly Disagree Somewhat or Disagree Agree Agree
Why?: ______________________________________________________
______________________________________________________________

Say: For the next 3 questions, refer to the YELLOW section of the answer sheet to help choose the appropriate answer.

5. The SoF table is clear in showing how the diagnostic test accuracy results compare between the two tests:

   □ □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Agree

   Why?: ______________________________________________________
   ______________________________________________________________

6. The density of information presented in this SoF table format is appropriate:

   □ □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Agree

   Why?: ______________________________________________________
   ______________________________________________________________

7. The information in this SoF table format is easy to read:

   □ □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Agree

   Why?: ______________________________________________________
   ______________________________________________________________

Say: I will now ask one more open ended question relating to Comparative Test SoF table Format 1.

8. Ask: What is your interpretation of the confidence intervals presented in the table? How does this data affect your decision making about the diagnostic tests?
L. Review Comparative Test SoF Format 2 – 3 Minutes

Say: We will now briefly review an alternative format of the Comparative Test SoF table. Please refer to the printout of Comparative Test SoF Table Format 2 on Page 11. Review the SoF table and let me know when you are ready for the questions about this format.

Say: I will now ask you about your preferences between the two Comparative Test SoF formats we have reviewed.

1. Ask: Do you prefer Comparative Test SoF Format 1 (on Pg. 10) with sensitivity and specificity information in the header, or Comparative Test SoF Format 2 (on Pg.11) with sensitivity and specificity as rows inside the table?

Comparative Test SoF Format 1 □ Comparative Test SoF Format 2 □ Any comment?:___

________________________________________________________________________

________________________________________________________________________

2. Ask: Should confidence intervals also be included for the absolute difference as in Comparative Test SoF Format 2, or should they be left out as in Comparative Test SoF Format 1? Why?

Include (Format 2) □ Exclude (Format 1) □ Why?:_____________________

________________________________________________________________________

________________________________________________________________________

3. Ask: Is there any additional information you would like to see presented in a SoF table comparing two tests? Would you change anything about the presentation of Comparative Test SoF Format 1 or 2?
M. Review Comparative Test SoF Format 3 – 2 Minutes

**Say:** We will now review a third alternative format of the Comparative Test SoF table. Please refer to the printout of Comparative Test SoF Table Format 3 on Page 12. Review this format for about 1 minute. Let me know when you are ready to answer the questions about this format.

**Say:** I will now ask a few open ended questions seeking your thoughts about SoF format 3.

1. **Ask:** What is your overall impression of this SoF table format?

2. **Ask:** Does this SoF table format have specific features or does it present information that you find more helpful to *understanding the systematic review results* and *decision making about the two diagnostic tests* when compared to the features in the previous two SoF formats? Which features were more helpful?

**Say:** I will now ask you about your overall preference between this Comparative Test SoF format, format 3, and the previous two Comparative Test SoF formats, 1 and 2, we reviewed.

3. **Ask:** Do you prefer SoF format 3 (Pg. 12) over SoF formats 1 (Pg. 10) and 2 (Pg. 11)?

   Prefer SoF Format 3 □  
   Prefer SoF Format 1 or 2 □  
   Any comment?:___________

   _________________________________________________________________
   _________________________________________________________________
N. Review Indirect Comparative Test SoF Formats – 5 minutes

Say: There are also instances when we would like to compare two diagnostic tests, but they may not have been evaluated in the same diagnostic test accuracy study. For this, we must draw an indirect comparison using data on the tests from separate studies in different populations. We would like to ask a few questions about your thoughts on the best approach for the indirect comparison of diagnostic tests.

Say: Please refer to the printouts of Indirect Comparative Tests: SoF Table 1 – ParaCheck Test and Indirect Comparative Tests: SoF Table 2 – ParaSight Test on Pages 13 and 14. Review the two SoF tables for about 1 minute to grasp the information presented.

Say: Now take a look at Indirect Comparative Tests: SoF Table 3 - Combined Tests on Page 15. Review this table for about another minute and let me know when you are ready to answer the questions about this approach.

Say: For the following 3 questions I will ask to what extent do you agree with these statements on the 7-point scale. Please refer to the YELLOW section of the answer sheet to help choose the appropriate answer.

1. Using the two separate SoF tables for the individual tests, it is clear which test has better diagnostic test accuracy results.

   □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Agree

   Why?: ____________________________________________
   ____________________________________________

2. Presenting two separate SoF tables for the individual diagnostic tests is sufficient for comparing and decision making about the tests.

   □ □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Agree

   Why?: ____________________________________________
   ____________________________________________

3. Presenting a combined SoF table with data for both tests and absolute difference in test results is necessary for decision making about the tests.

   □ □ □ □ □ □ □ □
   Strongly Disagree Somewhat or Disagree Neither Agree Somewhat Agree Strongly Ag
Say: I will now ask a few open ended questions seeking your thoughts about the Indirect Comparative Test SoF formats.

4. **Ask:** Do you think that presenting one combined SoF table for the indirect comparison is sufficient for decision making about the two tests? Why or why not? In your opinion, what does the combined SoF table add over the two individual SoF tables for each test, or what is the advantage to having the two individual SoF tables for decision making?

5. **Ask:** In the GRADE approach, in the case of an indirect comparison it is recommended to downgrade the quality of evidence by one level for indirectness of comparison. This is explained in the first footnote in Indirect Comparative Tests: SoF Table 3 - Combined Tests on Page 1. Did you notice this difference between the individual test SoF tables and the Combined Table? How does this affect your opinion on the presentation of individual and a combined SoF table for indirect comparative tests?

   Yes, Noticed the Difference □   Do, Did Not Notice Difference □

Say: In summary of these questions, I will now confirm your preference for which SoF tables to present when making an indirect comparison between two diagnostic tests.
6. **Ask:** Do you prefer having the individual SoF tables for each test (on Pg. 13 and 14), or only the combined SoF table (on Pg. 15), or all 3 SoF tables available in order to make the indirect comparison between two diagnostic tests?

Individual SoF Formats □   Combined SoF Format Only □   All 3 SoF Formats □

Any comment?:__________________________________________________
_______________________________________________________________

**[Layer 2: (Optional) to assess usefulness of SoF tables for decision making]**

---

0. **Comments Column Component – 4 minutes**

**Say:** The next component we would like some insight about is the Comments Column. Please refer to the printouts of Comments Column – Example Table 1 on Page 16 and Comments Column – Example Table 2 on Page 17. Review the SoF tables, noting the information provided in the comments column of each table.

**Say:** I will now ask you a few open-ended questions about the Comments Column.

1. **Ask:** As the two tables demonstrate, the contents of the comments column can be very simple, or more detailed, for instance providing some considerations about the perceived downstream consequences of the diagnostic test results. What content do you think should be presented in this column?

   ![Comment column](image)

2. **Ask:** Diagnostic accuracy studies do not provide direct evidence about patient-important outcomes and one must make deductions about the downstream impact on patient-important outcomes. What do you think should be considered in making the link between diagnostic test accuracy results and downstream patient-important outcomes? How do you think this link between the diagnostic test accuracy results and patient outcomes should be presented in a SoF table?

   ![Comment column](image)
3. **Ask:** Is there any content from the footnotes that you would like to see presented in the comments column instead? Or, vise versa, is there any content from the comments column that should be placed in the footnotes?

---

**P. Component: Additional Outcomes – 2 minutes**

1. **Ask:** Is there any additional information you would like to see presented in a SoF table to better evaluate diagnostic tests and make decisions? What information would be more helpful?

2. **Say:** Take a brief look at the SoF Table on Page 18, SoF Table Component: Presentation of Outcomes for an example. This table aims to present additional outcomes about a diagnostic test, including inconclusive results, complications of the test, and resource use, to help with decision making. Is additional information like this necessary for decision making, or is DTA sufficient? Why or why not?

   Yes, Necessary □   No, DTA is Sufficient □
3. **Ask:** Currently, these types of outcomes are often not reported in diagnostic test accuracy studies. Do you think more effort should be made to evaluate these outcomes in DTA studies? Why or why not?

Yes □  No □

---

Q. **Conclusion – 2 minutes**

2. **Ask:** Do you have any final comments that you would like to make about the content and presentation of information in SoF tables, keeping in mind that the goal of this interview was to identify problematic areas for users and to determine the usefulness and usability of the tables for presentation of DTA systematic review data?

3. **Say:** We would also like your feedback on how we might have organised this session better. Do you have any suggestions for improving the user testing and this interview?
Say: That marks the end of our interview. Thank you very much for your participation, we really appreciate you dedicating your time to this.
### APPENDIX 3.2.2B: ENHANCING THE USABILITY AND USEFULNESS OF SUMMARY OF FINDINGS TABLES FOR DECISION MAKING ABOUT DIAGNOSTIC TESTS – TABLES FOR EVALUATION

#### Summary of Findings (SoF) Table Format 1 – Layout 1

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</td>
<td>Prevalence 2%(^4) Prevalence 12%(^4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
<td>2777 (18) High(^5)</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality.</td>
</tr>
<tr>
<td>True negatives</td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
<td>2777 (18) High(^5)</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td>False positives</td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

**Notes:**

- CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### SoF Table Format 1 – Layout 2

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

**Patients or population:** immunocompromized patients, mostly hematology patients¹

**Settings:** mainly hematlogy or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum² | **Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria³ | **Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Number of Results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 2%⁴</td>
<td>Prevalence 12%⁴</td>
<td></td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
<td>🟢🟢🟢🟢 High⁵</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
<td>🟟🟢🟢🟢 High⁵</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
<td>🟢🟢🟢🟢 High⁵</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
<td>🟟🟢🟢🟢 High⁵</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


**Footnotes:**

¹ One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

² A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

³ A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

⁴ Estimates of prevalence of IA were based on the median and range of values in included studies.

⁵ Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### SoF Table Component: Quality of Evidence Rating

**ParaSight-F compared to microscopy in diagnosis of malaria in outpatients with symptoms suggestive of malaria in P. falciparum endemic areas**

**Patients or population:** patients with symptoms suggestive of malaria  
**Settings:** ambulatory health facilities in P. falciparum endemic areas  
**New Test:** ParaSight-F  
**Cut-off value:** -  
**Reference Test:** microscopy or PCR  
**Threshold:** Proven or probable malaria

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Pooled Sensitivity</th>
<th>Pooled Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,591 (17)</td>
<td>0.94 (0.90 to 0.97)</td>
<td>0.95 (0.90 to 0.97)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives</strong></td>
<td><strong>94 (90 to 97)</strong></td>
<td><a href="#">☆☆☆☆☆</a> Moderate Due to Risk of Bias¹ ²</td>
<td>True positives are important, because patients will be treated with improved survival.</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td><strong>6 (3 to 10)</strong></td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on morbidity and mortality.</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td><strong>851 (813 to 872)</strong></td>
<td><a href="#">☆☆☆☆☆</a> Moderate Due to Risk of Bias¹ ²</td>
<td>True negatives are important, because patients will not be treated unnecessarily and will be reassured.</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td><strong>50 (28 to 87)</strong></td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

CI: Confidence interval; PCR: Polymerase chain reaction


### Footnotes:

¹ Downgraded for risk of bias. Downgraded based on the assessment of all Type 1 studies included in the review together (only 25% reported an adequate reference standard, only half of the included studies were explicit about patient recruitment involving a consecutive or random series of patients, blinding of the index and reference tests was reported in 65% and 70%, respectively).

² Downgraded for risk of bias due to lack of adequate reference standard, blinding, and lack of information about patient recruitment in studies.
### SoF Table Format 2 – Layout 1

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Patients or population:</th>
<th>immunocompromized patients, mostly hematology patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings:</td>
<td>mainly hematology or cancer departments, mainly inpatients</td>
</tr>
<tr>
<td><strong>New Test:</strong></td>
<td>commercial Platelia® sandwich ELISA detecting galactomannan in serum</td>
</tr>
<tr>
<td><strong>Cut-off value:</strong></td>
<td>1.5 ODI</td>
</tr>
<tr>
<td><strong>Reference Test:</strong></td>
<td>composite of EORTC/MSG clinical and histological criteria</td>
</tr>
<tr>
<td><strong>Threshold:</strong></td>
<td>Proven or probable invasive aspergillosis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Property (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td></td>
<td></td>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td>0.64 (0.50 to 0.77)</td>
<td>2777 (18)</td>
<td>High</td>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality.</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td></td>
<td>True negatives</td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td>0.95 (0.91 to 0.97)</td>
<td>2777 (18)</td>
<td>High</td>
<td>False positives</td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
<td>False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

**Notes:**

- CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group

**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
Galactomannan ELISA for the diagnosis of invasive aspergillosis

**Patients or population:** immunocompromised patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum.

**Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria.

**Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test Property (95% CI)</th>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity</strong></td>
<td>True positives</td>
<td>6 (5 to 8) 77 (60 to 92)</td>
<td>2777 (18)</td>
<td>⫡idfidf</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td></td>
<td>False negatives</td>
<td>4 (2 to 5) 43 (28 to 60)</td>
<td></td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality.</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>True negatives</td>
<td>941 (901 to 960) 836 (801 to 854)</td>
<td>2777 (18)</td>
<td>⫡idfidf</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td></td>
<td>False positives</td>
<td>50 (30 to 89) 44 (26 to 79)</td>
<td></td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

**CI:** Confidence interval; **ODI:** Optical density index; **EORTC:** European Organization for Research and Treatment of Cancer; **MSG:** Mycoses Study Group


**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

4. Estimates of prevalence of IA were based on the median and range of values in included studies.

5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
## SoF Table Format 3

### What is the diagnostic accuracy of the galactomannan ELISA for invasive aspergillosis for different cut-off values?

**Patients/population:** immunocompromized patients, mostly hematology patients.

**Prior testing:** varied, mostly underlying disease or symptoms (fever, neutropenia)

**Settings:** mainly hematology or cancer departments, mainly inpatients.

**Index test:** a sandwich ELISA for galactomannan, an *Aspergillus* antigen.

**Importance:** depends on the time-gain the test may give.

**Reference standard:** gold standard is autopsy, but that is nearly never done; so in most studies the reference standard is composed of clinical and microbiological criteria

**Studies:** patient series or case-control studies, not using an in-house test and not excluding possibly infected patients. Studies (n = 29) had to report cut-off values that were used; 0.5 ODI, 1.0 ODI, or 1.5 ODI. Each study can be present in more than one subgroup. The results presented in this table are for the 1.5 ODI subgroup only.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Effect (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Prevalence (median, range)</th>
<th>What do these results mean?</th>
</tr>
</thead>
</table>
| **Cut-off 1.5 ODI** | Sensitivity 0.64 (0.50 to 0.77) | 2777 (18) | Median 12.4% (0.8% to 44%) | With a prevalence of 12%\(^1\), 12 out of 100 patients will develop IA. Of these, 4 will be missed by the Platelia test (36% of 12), but will be tested again. Of the 88 patients without IA, only 4 will be unnecessarily referred for CT scanning.  

In children (1 study, 17 participants), the sensitivity was higher (100%) and the specificity was lower (50%). |
| Specificity 0.95 (0.91 to 0.97) | | | | |

---

\(^1\) Prevalence over all 28 studies (children-studies excluded): 4501 participants; median 12% (range 0.8% to 44%).

---

SoF Table Component: Prevalence – Layout 1

Galactomannan ELISA for the diagnosis of invasive aspergillosis

| Patients or population: immunocompromized patients, mostly hematology patients. | Settings: mainly hematology or cancer departments, mainly inpatients |
| New Test: commercial Platelia® sandwich ELISA detecting galactomannan in serum¹. | Cut-off value: 1.5 ODI |
| Reference Test: composite of EORTC/MSG clinical and histological criteria². | Threshold: Proven or probable invasive aspergillosis |

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Pooled Sensitivity</th>
<th>Pooled Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2777 (18)</td>
<td>0.64 (95% CI 0.50 to 0.77)</td>
<td>0.95 (95% CI 0.91 to 0.97)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)³</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
<td>⬤⬤⬤⬤ High⁵</td>
</tr>
<tr>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
<td>⬤⬤⬤⬤ High⁵</td>
</tr>
<tr>
<td>False positives</td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
<td>⬤⬤⬤⬤ High⁵</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


Footnotes:
¹ A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
**Table Component: Prevalence – Layout 2 Table 1**

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Patients or population:</th>
<th>Immunocompromized patients, mostly hematology patients</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Settings:</strong></td>
<td>Mainly hematology or cancer departments, mainly inpatients</td>
</tr>
<tr>
<td><strong>New Test:</strong></td>
<td>Commercial Platelia® sandwich ELISA detecting galactomannan in serum</td>
</tr>
<tr>
<td><strong>Cut-off value:</strong></td>
<td>1.5 ODI</td>
</tr>
<tr>
<td><strong>Reference Test:</strong></td>
<td>Composite of EORTC/MSG clinical and histological criteria</td>
</tr>
<tr>
<td><strong>Threshold:</strong></td>
<td>Proven or probable invasive aspergillosis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Prevalence:</strong></th>
<th>**2%**³</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example Clinical Scenario:</strong></td>
<td>Adults undergoing transplant, no neutropenic patients.</td>
</tr>
<tr>
<td><strong>Pooled Sensitivity:</strong></td>
<td><strong>0.64 (95% CI)</strong></td>
</tr>
<tr>
<td><strong>Pooled Specificity:</strong></td>
<td><strong>0.95 (95% CI)</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)⁴</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>2777 (18)</td>
<td>⊕⊕⊕⊕ High⁵</td>
<td>True positives are important, because they provide improved survival and reduce the need for invasive diagnostic procedures, but patients may suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td></td>
<td></td>
<td>False negatives are important, because they represent missed diagnoses and may lead to delayed treatment with uncertain consequences on mortality.</td>
</tr>
<tr>
<td>True negatives</td>
<td>941 (901 to 960)</td>
<td>2777 (18)</td>
<td>⊕⊕⊕⊕ High⁵</td>
<td>True negatives are important, because they provide reassurance and prevent unnecessary treatment.</td>
</tr>
<tr>
<td>False positives</td>
<td>50 (30 to 89)</td>
<td></td>
<td></td>
<td>False positives are important, because they indicate patients who may be treated unnecessarily and exposed to nephrotoxic drugs.</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group.


**Footnotes:**

1. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
2. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of clinical and histological criteria.
3. Estimate of prevalence of IA was based on the median and range of values in included studies.
4. Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### SoF Table Component: Prevalence – Layout 2 Table 2

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

- **Patients or population**: immunocompromized patients, mostly hematology patients
- **Settings**: mainly hematology or cancer departments, mainly inpatients
- **New Test**: commercial Platelia® sandwich ELISA detecting galactomannan in serum¹
- **Cut-off value**: 1.5 ODI
- **Reference Test**: composite of EORTC/MSG clinical and histological criteria²
- **Threshold**: Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Prevalence:</th>
<th>12%³</th>
<th>Pooled Sensitivity:</th>
<th>0.64 (95% CI 0.50 to 0.77)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example Clinical Scenario:</td>
<td>Adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</td>
<td>Pooled Specificity:</td>
<td>0.95 (95% CI 0.91 to 0.97)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)⁴</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>2777 (18)</td>
<td>•••• High⁵</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td></td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality.</td>
</tr>
<tr>
<td>True negatives</td>
<td>836 (801 to 854)</td>
<td>2777 (18)</td>
<td>•••• High⁵</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td>False positives</td>
<td>44 (26 to 79)</td>
<td></td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


### Footnotes:

1 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

2 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

3 Estimates of prevalence of IA were based on the median and range of values in included studies.

4 Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).

5 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
# Comparative Test SoF Table Format 1

Magnetic resonance imaging compared to computed tomography for detection of acute vascular lesions in patients presenting with acute stroke symptoms

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Pooled Sensitivity CT</th>
<th>Pooled Specificity CT</th>
<th>Pooled Sensitivity MRI</th>
<th>Pooled Specificity MRI</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positives (TP)</td>
<td>195 (80 to 345)</td>
<td>0.39 (95% CI 0.16 to 0.69)</td>
<td>1.00 (95% CI 0.94 to 1.00)</td>
<td>0.99 (95% CI 0.23 to 1.00)</td>
<td>0.92 (95% CI 0.83 to 0.97)</td>
</tr>
<tr>
<td>False negatives (FN)</td>
<td>305 (155 to 420)</td>
<td>0.92 (95% CI 0.83 to 0.97)</td>
<td>0.99 (95% CI 0.23 to 1.00)</td>
<td>0.99 (95% CI 0.23 to 1.00)</td>
<td>0.99 (95% CI 0.23 to 1.00)</td>
</tr>
</tbody>
</table>

### Test Result Details

**Pooled Specificity CT**

- **CT**: 1.00 (95% CI 0.94 to 1.00)
- **MRI**: 0.92 (95% CI 0.83 to 0.97)

**Pooled Sensitivity CT**

- **CT**: 0.39 (95% CI 0.16 to 0.69)
- **MRI**: 0.92 (95% CI 0.83 to 0.97)

### Comments

- **True positives (TP)**: Patients will be correctly classified and treated; mortality and disability will be reduced, quality of life will be improved.
- **False negatives (FN)**: Patients may receive unnecessary treatment and suffer its adverse effects; true cause of symptoms may be missed and correct diagnosis will be delayed.
- **True negatives (TN)**: Patients will be reassured that they do not have stroke and will undergo investigation for other causes of symptoms and likely will be treated accordingly.
- **False positives (FP)**: Patients will be falsely reassured that they do not have stroke, unnecessary further testing for other causes of symptoms may be performed and the potentially beneficial treatment will be delayed.

### Footnotes

1. Prevalence of 50% was assumed to be the average prevalence in a representative population (this assumed prevalence should ideally be based on observational studies done in the target population).
2. Prevalence of 70% was estimated based on the mean prevalence of ischemic stroke in the included studies.
3. In some studies patients whose symptoms lasted less than 24 hours but who had evidence of an ischemic lesion on imaging were counted as having had strokes and hence analyzed as true positives.
## Magnetic resonance imaging compared to computed tomography for detection of acute vascular lesions in patients presenting with acute stroke symptoms

### Comparative Test SoF Table Format 2

**Patients or population:** adult patients suspected of acute stroke within 12 hours of the onset of symptoms  
**Settings:** hospital emergency departments  
**New Test:** diffusion-weighted magnetic resonance imaging  
**Comparison Test:** Non-contrast computed tomography  
**Reference Test:** a combination of clinical and imaging information supported by clinical or imaging follow up (CT or MRI)

<table>
<thead>
<tr>
<th>Test Property</th>
<th>Summary Estimate (95% CI)</th>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CT</td>
<td>MRI</td>
<td>Prevalence 50%(^1)</td>
<td>Prevalence 70%(^2)</td>
<td></td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td></td>
<td></td>
<td></td>
<td>CT</td>
<td>MRI</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.39 (0.16 to 0.69)</td>
<td>0.99 (0.23 to 1.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>195 (80 to 345)</td>
<td>495 (115 to 500)</td>
<td>273 (112 to 483)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>300 more</td>
<td>420 more</td>
<td>(230 fewer to 420 more)</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td></td>
<td></td>
<td></td>
<td>1.00 (0.94 to 1.00)</td>
<td>0.92 (0.83 to 0.97)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>500 (470 to 500)</td>
<td>460 (415 to 485)</td>
<td>300 (282 to 300)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40 fewer</td>
<td>24 fewer</td>
<td>(15 more to 85 fewer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0 (0 to 30)</td>
<td>0 (0 to 18)</td>
<td>0 (0 to 18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>40 more</td>
<td>24 more</td>
<td>(15 fewer to 85 more)</td>
</tr>
</tbody>
</table>

---

\(^1\) Prevalence of 50% was assumed to be the average prevalence in a representative population (this assumed prevalence should ideally be based on observational studies done in the target population)  
\(^2\) Prevalence of 70% was estimated based on the mean prevalence of ischemic stroke in the included studies.  
\(^3\) In some studies patients whose symptoms lasted less than 24 hours but who had evidence of an ischemic lesion on imaging were counted as having had strokes and hence analyzed as true positive cases. In other studies, however, patients with symptom duration less than 24 hours and an ischemic lesion on imaging were analyzed as being false positive cases.  
\(^4\) Serious risk of bias, because of unblinded reference standard results and clinical information available that otherwise would not be available; serious indirectness (not representative population); serious imprecision (only 226 patients and very wide confidence intervals); serious inconsistency (sensitivity of CT varied very widely)  
\(^5\) Serious risk of bias, because of unblinded reference standard results and clinical information available that otherwise would not be available; serious indirectness (not representative population); serious imprecision (only 226 patients and very wide confidence intervals)
**Comparative Test SoF Table Format 3**

**Comparison of diffusion-weighted magnetic resonance imaging with conventional computer tomography for the early detection of ischaemic brain lesions in patients suspected of stroke**

**Patient population:** adults suspected of acute stroke  
**Setting:** hospital departments

**Geographical location:** studies were conducted in Europe (3 studies), the USA (3 studies), and in Australia (1 study)

**Index test:** diffusion-weighted magnetic resonance imaging (MRI) performed within 12 hours of stroke onset

**Alternative test:** computer tomography (CT) performed within 12 hours of stroke onset

**Reference standard:** clinical assessment and imaging follow up

**Included studies:** 7 comparative studies that evaluated MRI and CT in the same patients

**Total number of patients assessed:** 226

<table>
<thead>
<tr>
<th>Limitations of included studies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Limited number of included studies (7 studies); small sample sizes; presence of incorporation bias</td>
<td></td>
</tr>
<tr>
<td>MRI and CT were evaluated in highly selected patient samples (patients with high probability of stroke), which therefore are not representative of the typical population of patients presenting with 'suspected acute stroke' to an emergency department (poor generalisability of results)</td>
<td></td>
</tr>
<tr>
<td>The stroke vascular territory was not reported in the majority of included studies although it is likely that they enrolled patients with anterior circulation stroke</td>
<td></td>
</tr>
<tr>
<td>Only a minority of the studied patients had severe strokes (in whom MRI might be contraindicated)</td>
<td></td>
</tr>
<tr>
<td>The high proportion of mild strokes and reclassification of TIA cases with a positive MRI lesion as strokes might have inflated sensitivity estimate</td>
<td></td>
</tr>
<tr>
<td>In most of the studies stroke mimics were not included</td>
<td></td>
</tr>
<tr>
<td>In all but one study CT was performed before MRI (reducing the sensitivity of CT to detect ischaemia)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CT Results</th>
<th>MRI Results</th>
<th>Summary Effect (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP (73)</td>
<td>TP (147)</td>
<td>CT sensitivity 0.39 (0.16 to 0.69)</td>
</tr>
<tr>
<td>FP (0)</td>
<td>FP (5)</td>
<td>CT specificity 1.00 (0.94 to 1.00)</td>
</tr>
<tr>
<td>TN (88)</td>
<td>TN (14)</td>
<td>MRI sensitivity 0.99 (0.23 to 1.00)</td>
</tr>
<tr>
<td>FN (65)</td>
<td>FN (60)</td>
<td>MRI specificity 0.92 (0.83 to 0.97)</td>
</tr>
</tbody>
</table>

| Total 226 | Total 226 |

**Conclusions and comments**

In the small cohort of included studies, MRI is more sensitive than CT - but not more specific - for the early detection of ischaemic brain lesions. The small amount of data and the presence of methodological biases preclude any reliable calculation - from the sensitivity and specificity estimates - of a positive or negative stroke diagnosis at different rates of stroke prevalence.

**Applicability of tests in clinical practice**

None of the studies addressed practicality. CT is known to be quicker to perform and more readily available in most emergency care settings than magnetic resonance imaging (MRI). MRI is contraindicated in patients with pacemakers and some metal implants. In acutely ill str may be difficult to monitor the patient’s condition while being MR scanned (and this increases the risk of any respiratory difficulty compromise that develops during the scan which passes undetected and may have adverse effects for the patient). If the patient is restless as a result of the stroke, the patient may not be able to co-operate for the longer scan times of MRI.

**Costs**

None of the studies included a cost-effectiveness evaluation. MRI is known to be more expensive than CT.

CI: Confidence interval; CT: computed tomography; MRI: diffusion-weighted magnetic resonance imaging; TP: true positive; FP: false positive; TN: true negative; FN: false negative

**Indirect Comparative Tests: SoF Table 1 – ParaCheck Test**

ParaCheck-Pf compared to microscopy in diagnosis of malaria in outpatients with symptoms suggestive of malaria in P. falciparum endemic areas

<table>
<thead>
<tr>
<th>Patients or population:</th>
<th>patients with symptoms suggestive of malaria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Settings:</td>
<td>ambulatory health facilities in P. falciparum endemic areas</td>
</tr>
<tr>
<td>New Test:</td>
<td>ParaCheck-Pf</td>
</tr>
<tr>
<td>Cut-off value:</td>
<td></td>
</tr>
<tr>
<td>Reference Test:</td>
<td>microscopy or PCR</td>
</tr>
<tr>
<td>Threshold:</td>
<td>Proven or probable malaria</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Pooled Sensitivity</th>
<th>Pooled Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>22,319 (27)</td>
<td>0.93 (0.90 to 0.96)</td>
<td>0.96 (0.93 to 0.98)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prevalence 10%</td>
<td>Prevalence 30%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>93 (90 to 96)</td>
<td>280 (269 to 287)</td>
<td>⊗⊗⊗⊗</td>
<td>Moderate†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>True positives are important, because patients will be treated with improved survival.</td>
<td></td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7 (4 to 10)</td>
<td>20 (13 to 31)</td>
<td>⊗⊗⊗⊗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>True negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on morbidity and mortality.</td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>861 (834 to 878)</td>
<td>670 (649 to 683)</td>
<td>⊗⊗⊗⊗</td>
<td>Moderate†</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>True negatives are important, because patients will not be treated unnecessarily and will be reassured.</td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>39 (23 to 66)</td>
<td>30 (18 to 51)</td>
<td>⊗⊗⊗⊗</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; PCR: Polymerase chain reaction


Footnotes:

† Downgraded for risk of bias. Downgraded based on the assessment of all Type 1 studies included in the review together (only 25% reported an adequate reference standard, only half of the included studies were explicit about patient recruitment involving a consecutive or random series of patients, blinding of the index and reference tests was reported in 65% and 70%, respectively).
### Indirect Comparative Tests: SoF Table 2 – ParaSight Test

**ParaSight-F compared to microscopy in diagnosis of malaria in outpatients with symptoms suggestive of malaria in P. falciparum endemic areas**

**Patients or population:** patients with symptoms suggestive of malaria  
**Settings:** ambulatory health facilities in P. falciparum endemic areas  
**New Test:** ParaSight-F  
**Cut-off value:**  
**Reference Test:** microscopy or PCR  
**Threshold:** Proven or probable malaria

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Pooled Sensitivity</th>
<th>Pooled Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>12,591 (17)</td>
<td>0.94 (0.90 to 0.97)</td>
<td>0.95 (0.90 to 0.97)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives</strong></td>
<td>94 (90 to 97)</td>
<td>283 (269 to 290)</td>
<td>⊕⊕Ο Moderate</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>6 (3 to 10)</td>
<td>17 (10 to 31)</td>
<td>⊕⊕Ο Moderate</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>851 (813 to 872)</td>
<td>662 (632 to 678)</td>
<td>⊕⊕Ο Moderate</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td>50 (28 to 87)</td>
<td>39 (22 to 68)</td>
<td>⊕ΟΟΟ Moderate</td>
</tr>
</tbody>
</table>

**Notes:**  
- **GRADE:** Grading of Recommendations, Assessment, Development, and Evaluation.  
- **Confidence interval (CI):** The range within which the true value is expected to lie.  
- **PCR:** Polymerase chain reaction.

---

**Footnotes:**  
1. Downgraded for risk of bias. Downgraded based on the assessment of all Type 1 studies included in the review together (only 25% reported an adequate reference standard, only half of the included studies were explicit about patient recruitment involving a consecutive or random series of patients, blinding of the index and reference tests was reported in 65% and 70%, respectively).  
2. Low for medium-risk (30% initial probability of malaria) population because of imprecision.

---

Indirect Comparative Tests: SoF Table 3 - Combined Tests

Pracheck-Pf compared to ParaSight-F in diagnosis of malaria in outpatients with symptoms suggestive of malaria in P. falciparum endemic areas

Patients or population: patients with symptoms suggestive of malaria
Settings: ambulatory health facilities in P. falciparum endemic areas
Index Test 1: Pracheck-Pf  
Cut-off value: –
Index Test 2: ParaSight-F  
Cut-off value: –
Reference Test: microscopy or PCR
Threshold: Proven or probable malaria

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prevalence 10%</td>
<td>Prevalence 30%</td>
</tr>
<tr>
<td>True positives (TP)</td>
<td>ParaCheck</td>
<td>ParaSight</td>
</tr>
<tr>
<td></td>
<td>93 (90 to 96)</td>
<td>94 (90 to 97)</td>
</tr>
<tr>
<td>TP absolute difference</td>
<td>1 more</td>
<td>3 more</td>
</tr>
<tr>
<td>False negatives (FN)</td>
<td>7 (4 to 10)</td>
<td>6 (3 to 10)</td>
</tr>
<tr>
<td>FN absolute difference</td>
<td>1 fewer</td>
<td>3 fewer</td>
</tr>
<tr>
<td>True negatives (TN)</td>
<td>861 (834 to 878)</td>
<td>851 (813 to 872)</td>
</tr>
<tr>
<td>TN absolute difference</td>
<td>10 fewer</td>
<td>8 fewer</td>
</tr>
<tr>
<td>False positives (FP)</td>
<td>39 (23 to 66)</td>
<td>50 (28 to 87)</td>
</tr>
<tr>
<td>FP absolute difference</td>
<td>11 more</td>
<td>9 more</td>
</tr>
</tbody>
</table>

CI: Confidence interval; PCR: Polymerase chain reaction

Footnotes:
1. For an indirect comparison of 2 index tests we suggest that the score for each domain of QoE be determined as the lower of the scores for that domain for each of index tests compared against a reference standard. We suggest that the overall QoE for an indirect comparison of 2 index tests be further downgraded by one level for indirectness of comparison.
2. Downgraded based on the assessment of all Type 1 studies included in the review together (only 25% reported an adequate reference standard, only half of the included studies were explicit about patient recruitment involving a consecutive or random series of patients, blinding of the index and reference tests was reported in 65% and 70%, respectively).
# SoF Table Component: Comments Column – Example Table 1

## Galactomannan ELISA for the diagnosis of invasive aspergillosis

**Patients or population:** immunocompromized patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum

**Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria

**Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Result</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 2%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6 (5 to 8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 12%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>77 (60 to 92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>⊕⊕⊕⊕ High^5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>With a prevalence of 12%, 1200 patients will develop IA. Of these correctly classified by the Platelia IA (64% of 1200) and treated with improved survival.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 2%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4 (2 to 5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 12%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>43 (28 to 60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 2%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>941 (901 to 960)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 12%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>836 (801 to 854)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>⊕⊕⊕⊕ High^5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>With a prevalence of 12%, 880 patients will not develop IA. Of these correctly classified by the Platelia as not having IA (36% of 880). Patients will be spared the toxic effects of treatment.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 2%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>50 (30 to 89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prevalence 12%^4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>44 (26 to 79)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**CI:** Confidence interval; **ODI:** Optical density index; **EORTC:** European Organization for Research and Treatment of Cancer; **MSG:** Mycoses Study Group


**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patients a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### Table 2

#### Galactomannan ELISA for the diagnosis of invasive aspergillosis

<table>
<thead>
<tr>
<th>SoF Table Component: Comments Column – Example Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Patients or population:</strong> immunocompromized patients, mostly hematology patients¹</td>
</tr>
<tr>
<td><strong>Settings:</strong> mainly hematology or cancer departments, mainly inpatients</td>
</tr>
<tr>
<td><strong>New Test:</strong> commercial Platelia® sandwich ELISA detecting galactomannan in serum²</td>
</tr>
<tr>
<td><strong>Cut-off value:</strong> 1.5 ODI</td>
</tr>
<tr>
<td><strong>Reference Test:</strong> composite of EORTC/MSG clinical and histological criteria³</td>
</tr>
<tr>
<td><strong>Threshold:</strong> Proven or probable invasive aspergillosis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Participants (Studies)</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prevalence 2%⁴</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>6 (5 to 8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>77 (60 to 92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>⊕⊕⊕⊕ High⁵</td>
<td>Detected by the test and referred for treatment.</td>
</tr>
<tr>
<td></td>
<td>Prevalence 12%⁴</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>4 (2 to 5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>43 (28 to 60)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Not detected by the test, with a delay in treatment.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>941 (901 to 960)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>836 (801 to 854)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>⊕⊕⊕⊕ High⁵</td>
<td>No referral, patient reassurance.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td>50 (30 to 89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>44 (26 to 79)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Unnecessary referral for follow-up testing and/or unnecessary treatment.</td>
</tr>
</tbody>
</table>

**Notes:**

- **CI:** Confidence interval; **ODI:** Optical density index; **EORTC:** European Organization for Research and Treatment of Cancer; **MSG:** Mycoses Study Group


**Footnotes:**

1 One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4 Estimates of prevalence of IA were based on the median and range of values in included studies.
5 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### SoF Table Component: Presentation of Outcomes

**Paracheck-Pf compared to microscopy in diagnosis of malaria in outpatients with symptoms suggestive of malaria in P. falciparum endemic areas**

**Patients or population:** patients with symptoms suggestive of malaria. **Settings:** ambulatory health facilities in P. falciparum endemic areas  
**New Test:** Paracheck-Pf. **Cut-off value:**  
**Reference Test:** microscopy or PCR. **Threshold:** Proven or probable malaria

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prevalence 10%</td>
<td>Prevalence 30%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>93 (90 to 96)</td>
<td>280 (269 to 287)</td>
<td>22,319 (27)</td>
<td>☓ ☓ ☓ ☓ Moderate¹</td>
</tr>
<tr>
<td>False negatives</td>
<td>7 (4 to 10)</td>
<td>20 (13 to 31)</td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on morbidity and mortality.</td>
</tr>
<tr>
<td>True negatives</td>
<td>861 (834 to 878)</td>
<td>670 (649 to 683)</td>
<td>22,319 (27)</td>
<td>☓ ☓ ☓ ☓ Moderate¹</td>
</tr>
<tr>
<td>False positives</td>
<td>39 (23 to 66)</td>
<td>30 (18 to 51)</td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

**Outcome**

<table>
<thead>
<tr>
<th>Inconclusive Results</th>
<th>Uninterpretable, intermediate, or indeterminate test results. They are important because they generate anxiety for the patient, uncertainty as to how to proceed, and likely repeat testing.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complications</td>
<td>Complications of the Paracheck-Pf test are unlikely and usually related to blood sampling as it is a noninvasive test.</td>
</tr>
<tr>
<td>Resource Use</td>
<td>The cost of the test is sufficiently low that its routine use for confirmation of P. falciparum malaria is likely to be cost-effective. Other resource use including staffing and facility costs were not reported.</td>
</tr>
</tbody>
</table>

CI: Confidence interval; PCR: Polymerase chain reaction  
Footnotes:

1 Downgraded for risk of bias. Downgraded based on the assessment of all Type 1 studies included in the review together (only 25% reported an adequate reference standard, only half of the included studies were explicit about patient recruitment involving a consecutive or random series of patients, blinding of the index and reference tests was reported in 65% and 70%, respectively).

2 Inconclusive results are inconsistent as they were reported to range from 1% to 14% in different studies.

3 Indirectness of resource use data as cost of test may vary in different settings that those reported in the studies.
APPENDIX 3.3: EVIDENCE PROFILE WITH INDIVIDUAL DTA OUTCOMES

Galactomannan ELISA for the diagnosis of invasive aspergillosis

**Question:** Should Galactomannan ELISA be used for the diagnosis of invasive aspergillosis?

**Authors of the profile:** Nancy Santesso, Jan Brozek, Holger Schünemann

**Population:** Immunocompromised patients, mostly hematology patients [1]

**Setting:** Mainly hematology or cancer departments, mainly inpatients

**New Test:** Commercial Platelia® sandwich ELISA detecting galactomannan in serum [2]. **Cut-off value:** 1.5 OD

**Reference Test:** Composite of EORTC/MSG clinical and histological criteria [3]

**Threshold:** Proven or probable invasive aspergillosis

**Bibliography:** Leeflang et al. Galactomannan detection for invasive aspergillosis in immunocompromised patients. Cochrane Database of Systematic Reviews 2008, Issue 4. Art. No.: CD007394

<table>
<thead>
<tr>
<th>Number of Participants (studies)</th>
<th>Study design</th>
<th>Risk of bias</th>
<th>Indirectness</th>
<th>Inconsistency</th>
<th>Imprecision</th>
<th>Publication bias</th>
<th>Effect (number per 1000 patients tested)</th>
<th>Quality of the evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives (Patients correctly classified as having invasive aspergillosis) [6]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2777 (18 studies)</td>
<td>cohort, case-control</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>unlikely</td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
</tr>
<tr>
<td><strong>True negatives (Patients correctly classified as not having invasive aspergillosis) [7]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2777 (18 studies)</td>
<td>cohort, case-control</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>unlikely</td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
</tr>
<tr>
<td><strong>False positives (Patients incorrectly classified as having invasive aspergillosis) [8]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2777 (18 studies)</td>
<td>cohort, case-control</td>
<td>no</td>
<td>serious [9]</td>
<td>no</td>
<td>no</td>
<td>unlikely</td>
<td>50 (89 to 30)</td>
<td>44 (79 to 26)</td>
</tr>
<tr>
<td><strong>False negatives (Patients incorrectly classified as not having the disease) [10]</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2777 (18 studies)</td>
<td>cohort, case-control</td>
<td>no</td>
<td>serious [11]</td>
<td>no</td>
<td>no</td>
<td>unlikely</td>
<td>4 (5 to 2)</td>
<td>43 (60 to 28)</td>
</tr>
</tbody>
</table>

**Footnotes:**

[1] One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

[2] A diagnostic test for invasive aspergillosis (1A) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome that a current reference standard.

[3] A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
[8] False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.
[9] There is uncertainty to what extent a delayed diagnosis and treatment of true cause of symptoms would impact on patient-important outcomes.
[10] False negatives important, because of missed diagnosis and either will not be treated with 60–90% mortality rate or treatment will be delayed with uncertain consequences on mortality.
[12] Inconclusive results were not reported. Most likely they would result in repeating the index test and therefore increased cost.
[13] Complications and resource use were not reported.
### Summary of findings table with individual DTA data

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

**Patients or population:** immuno compromised patients, mostly hematology patients. **Settings:** mainly hematology or cancer departments, mainly inpatients  
**New Test:** commercial Plateia® sandwich ELISA detecting galactomannan in serum. **Cut-off value:** 1.5 ODI  
**Reference Test:** composite of EORTC/MSG clinical and histological criteria  
**Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Quality of the Evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Example Clinical Scenario: Adults undergoing transplant, no neutropenic patients.</td>
<td><strong>Prevalence 2%</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>6 (5 to 8)</td>
</tr>
<tr>
<td></td>
<td>Example Clinical Scenario: Adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</td>
<td><strong>Prevalence 2%</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 (2 to 5)</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>High</strong></td>
<td><strong>High</strong></td>
</tr>
<tr>
<td></td>
<td>Patients will be correctly classified and treated; mortality and disability will be reduced, quality of life will be improved.</td>
<td>Patients may receive unnecessary treatment and suffer its adverse effects; true cause of symptoms may be missed and correct diagnosis will be delayed.</td>
</tr>
<tr>
<td><strong>True negatives</strong></td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
</tr>
</tbody>
</table>

**CI:** Confidence interval; **ODI:** Optical density index; **EORTC:** European Organization for Research and Treatment of Cancer; **MSG:** Mycoses Study Group


**Footnotes:**

1 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

2 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

3 Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).

4 Estimates of prevalence of IA were based on the median and range of values in included studies.

5 Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
## APPENDIX 3.4

**SoF Table – 3 Prevalence estimates in rows (sens and spec in table)**

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Test Property</th>
<th>Summar y Estimate (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Test Result</th>
<th>Illustrative Comparative Numbers per 1000 patients tested (95% CI)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sensitivity</strong></td>
<td>0.64 (0.50 to 0.77)</td>
<td>2777 (18)</td>
<td>⚫⚫⚫⚫ High</td>
<td>True positives</td>
<td>Prevalence 1%</td>
<td>True positives are important, because patients will be treated with improved survival and spared the toxicity of invasive diagnostic procedures.</td>
</tr>
<tr>
<td>(Patients correctly classified as having invasive aspergillosis)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td></td>
<td></td>
<td></td>
<td>Prevalence 1%</td>
<td>6 (5 to 8)</td>
<td>False negatives are important, because they are missed cases and patients may not be treated or treated too late.</td>
</tr>
<tr>
<td>(Patients incorrectly classified as not having invasive aspergillosis)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td>0.95 (0.91 to 0.97)</td>
<td>2777 (18)</td>
<td>⚫⚫⚫⚫ High</td>
<td>True negatives</td>
<td>Prevalence 1%</td>
<td>True negatives are important, because patients will not be treated unnecessarily and spared the toxic effects of treatment.</td>
</tr>
<tr>
<td>(Patients correctly classified as not having invasive aspergillosis)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>False positives</strong></td>
<td></td>
<td></td>
<td></td>
<td>Prevalence 1%</td>
<td>50 (30 to 89)</td>
<td>False positives are important, because patients may be exposed to nephrotoxic drugs and possibly delayed diagnosis and treatment.</td>
</tr>
<tr>
<td>(Patients incorrectly classified as having invasive aspergillosis)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group.

---

**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patients that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
**SoF Table - 3 Prevalence estimates in rows (sens and spec in header)**

### Galactomannan ELISA for the diagnosis of invasive aspergillosis

**Patients or population:** immunocompromized patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum

**Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria

**Threshold:** Proven or probable invasive aspergillosis

**Pooled Sensitivity:** 0.64 (95% CI 0.50 to 0.77), **Pooled Specificity:** 0.95 (95% CI 0.91 to 0.97)

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Illustrative Comparative Numbers per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True positives</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| (Patients correctly classified as having invasive aspergillosis) | Prevalence 1%  
6 (5 to 8) | 2777 (18) | High | True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment. |
| | Prevalence 12%  
77 (60 to 92) | | | |
| | Prevalence 44%  
282 (220 to 339) | | | |
| **False negatives** | | | | |
| (Patients incorrectly classified as not having invasive aspergillosis) | Prevalence 1%  
4 (2 to 5) | 2777 (18) | High | False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality. |
| | Prevalence 12%  
43 (28 to 60) | | | |
| | Prevalence 44%  
158 (101 to 220) | | | |
| **True negatives** | | | | |
| (Patients correctly classified as not having invasive aspergillosis) | Prevalence 1%  
941 (901 to 960) | 2777 (18) | High | True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured. |
| | Prevalence 12%  
836 (801 to 854) | | | |
| | Prevalence 44%  
532 (510 to 543) | | | |
| **False positives** | | | | |
| (Patients incorrectly classified as having invasive aspergillosis) | Prevalence 1%  
50 (30 to 89) | 2777 (18) | High | False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms. |
| | Prevalence 12%  
44 (26 to 79) | | | |
| | Prevalence 44%  
28 (17 to 50) | | | |

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


**Footnotes:**
1 One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2 A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3 A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4 Estimates of prevalence of IA were based on the median and range of values in included studies.
**SoF Table – 2 prevalence estimates (sens and spec in header)**

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prevalence 2%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
<td>2777 (18)</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td>False negatives</td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
<td></td>
<td>False negatives are important, because of missed diagnosis and either will not be treated or treatment will be delayed with uncertain consequences on mortality.</td>
</tr>
<tr>
<td>Prevalence 12%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>941 (901 to 960)</td>
<td>836 (801 to 854)</td>
<td>2777 (18)</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td>False positives</td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
<td></td>
<td>False positives are important, because patients will be treated unnecessarily and many will be exposed to nephrotoxic drugs, also because of a possibly delayed diagnosis and treatment of true cause of symptoms.</td>
</tr>
</tbody>
</table>

CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group


### Footnotes:

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
### SoF Table – 2 prevalence estimates (sens and spec in table)

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

<table>
<thead>
<tr>
<th>Test Property (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>True positives</td>
<td></td>
<td>Prevalence 2%(^4)</td>
<td>True positives are important, because patients will be treated with improved survival and spared the invasive diagnostic procedures, however they will suffer the toxicity of the treatment.</td>
</tr>
<tr>
<td><strong>Sensitivity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Prevalence 12%(^4)</td>
<td></td>
</tr>
<tr>
<td>0.64 (0.50 to 0.77)</td>
<td>2777 (18)</td>
<td>⬤⬤⬤⬤ High(^5)</td>
<td></td>
<td></td>
<td>6 (5 to 8)</td>
<td>77 (60 to 92)</td>
</tr>
<tr>
<td>False negatives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4 (2 to 5)</td>
<td>43 (28 to 60)</td>
</tr>
<tr>
<td><strong>Specificity</strong></td>
<td></td>
<td></td>
<td>True negatives</td>
<td></td>
<td>941 (901 to 960)</td>
<td>True negatives are important, because patients will not be treated unnecessarily, will be spared the toxic effects of treatment, and will be reassured.</td>
</tr>
<tr>
<td>0.95 (0.91 to 0.97)</td>
<td>2777 (18)</td>
<td>⬤⬤⬤⬤ High(^5)</td>
<td></td>
<td></td>
<td>836 (801 to 854)</td>
<td></td>
</tr>
<tr>
<td>False positives</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50 (30 to 89)</td>
<td>44 (26 to 79)</td>
</tr>
</tbody>
</table>

**Notes:**
- CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group

**Footnotes:**
1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

**Patients or population:** immunocompromized patients, mostly hematology patients

**Settings:** mainly hematology or cancer departments, mainly inpatients

**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum | **Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria | **Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prevalence 20 per 1000⁵: Which is typically seen in adults undergoing transplant, no neutropenic patients.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>13 (9 to 15)</td>
<td>282 (220 to 339)</td>
<td>2777 (18)</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>7 (4 to 10)</td>
<td>158 (101 to 220)</td>
<td></td>
</tr>
<tr>
<td>Prevalence 440 per 1000⁵: Which is typically seen in adults with hematological disorders that were neutropenic, underwent chemotherapy, had persistent fever despite antibiotics, acute graft versus host disease, or received corticosteroids.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True positives</strong></td>
<td>931 (901 to 960)</td>
<td>532 (510 to 543)</td>
<td>2777 (18)</td>
</tr>
<tr>
<td><strong>False negatives</strong></td>
<td>49 (30 to 89)</td>
<td>28 (17 to 50)</td>
<td></td>
</tr>
</tbody>
</table>

**Sensitivity (95% CI):** 0.64 (0.50 to 0.77)

**Specificity (95% CI):** 0.95 (0.91 to 0.97)

---

**Footnotes:**

1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.

2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.

3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.

4. Assumed numbers with galactomannan ELISA compared to reference test (EORTC/MSG criteria).

5. Estimates of prevalence of IA were based on the median and range of values in included studies.

6. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.

---

### SoF Table – 2 prevalence estimates (sens and spec in table), no clinical scenario

**Galactomannan ELISA for the diagnosis of invasive aspergillosis**

**Patients or population:** immunocompromized patients, mostly hematology patients\(^1\)  
**Settings:** mainly hematology or cancer departments, mainly inpatients  
**New Test:** commercial Platelia® sandwich ELISA detecting galactomannan in serum\(^2\)  
**Cut-off value:** 1.5 ODI

**Reference Test:** composite of EORTC/MSG clinical and histological criteria\(^3\)  
**Threshold:** Proven or probable invasive aspergillosis

<table>
<thead>
<tr>
<th>Test Result</th>
<th>Number of results per 1000 patients tested (95% CI)</th>
<th>Number of participants (studies)</th>
<th>Quality of the Evidence (GRADE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prevalence 20 per 1000(^4)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sensitivity (95% CI): 0.64 (0.50 to 0.77)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True positives</td>
<td>13 (9 to 15)</td>
<td>2777 (18)</td>
<td>⭐⭐⭐⭐⭐ High(^5)</td>
</tr>
<tr>
<td>False negatives</td>
<td>7 (4 to 10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Specificity (95% CI): 0.95 (0.91 to 0.97)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True negatives</td>
<td>931 (901 to 960)</td>
<td>2777 (18)</td>
<td>⭐⭐⭐⭐⭐ High(^5)</td>
</tr>
<tr>
<td>False positives</td>
<td>49 (30 to 89)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\text{CI: Confidence interval; ODI: Optical density index; EORTC: European Organization for Research and Treatment of Cancer; MSG: Mycoses Study Group}\)


**Footnotes:**
1. One study that included only patients with HIV/AIDS was excluded, because this patient group and setting differed from the other included patient groups in such a way that authors of the review regarded the study population as not being representative.
2. A diagnostic test for invasive aspergillosis (IA) needs to be not too invasive or a burden to immunocompromised patients. The galactomannan ELISA test on a serum specimen is less burdensome than the current reference standard.
3. A true reference standard is autopsy combined with culture from autopsy specimens; a clinical reference standard is composed of EORTC/MSG clinical and histological criteria.
4. Estimates of prevalence of IA were based on the median and range of values in included studies.
5. Included studies were valid diagnostic accuracy studies in patients with diagnostic uncertainty and direct comparison of test results with an appropriate reference standard. Quality was not downgraded based on the GRADE criteria.
CHAPTER 4: A SYSTEMATIC REVIEW OF AVAILABLE INSTRUMENTS TO ASSESS THE QUALITY OF EVIDENCE AND STRENGTH OF RECOMMENDATIONS
Decision-making about healthcare related tests and diagnostic strategies: A systematic review of available instruments to assess the quality of evidence and strength of recommendations

Reem A. Mustafa, MD, MPH\textsuperscript{1,2}, Wojtek Wiercioch, MSc\textsuperscript{1}, Maicon Falavigna, MD\textsuperscript{1}, Yuan Zhang, B.Sc.\textsuperscript{1}, Liudmila Ivanova, MD, MPH\textsuperscript{1}, Ingrid Arevalo-Rodriguez, MSc\textsuperscript{3}, Adrienne Cheung, B.H.Sc\textsuperscript{4}, Barbara Prediger, B.Sc.\textsuperscript{5}, Matthew Ventresca, B. Sc.\textsuperscript{1}, Jan Brozek, MD\textsuperscript{1,6}, Nancy Santesso, RD, MLIS\textsuperscript{1}, Patrick Bossuyt, MSc, PhD\textsuperscript{7}, Amit X. Garg, MD, PhD\textsuperscript{1,8}, Nancy Lloyd, MSc, B.Sc.\textsuperscript{1}, Monika Legemmann, MD, MSc\textsuperscript{9}, Diedrich Bühler, MD\textsuperscript{10}, Holger J Schünemann, MD, MSc, PhD\textsuperscript{1,6}.

(22) Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada

(23) Department of Internal Medicine, University of Missouri-Kansas City, Kansas City, USA

(24) Fundación Universitaria de Ciencias de la Salud, Hospital San José & Hospital Infantil de San José, Colombia

(25) Faculty of Medicine, University of British Colombia, Vancouver, Canada

(26) Center for Medical Biometry and Medical Informatics, University of Freiburg, Germany

(27) Department of Medicine, McMaster University, Hamilton, Canada

(28) Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherland

(29) Department of Medicine, Western University, London, Canada

(30) Medizinischer Dienst des Spitzenverbandes Bund der Kranken-kassen e.V. (MDS) Theodor Althoff-Str. 47 45133 Essen, Germany

(31) Abteilung Medizin. GKV— Reinhardtstraße 28 10117 Berlin, Germany
ABSTRACT

Objectives
To identify, describe and compare available instruments, checklists, and critical appraisal tools designed for assessing the quality of evidence (QoE) and/or strength of recommendations (SoR) dealing with healthcare related tests and diagnostic strategies (HCTDS).

Study Design and Settings
We conducted a comprehensive systematic search of electronic databases and websites of major international organizations to identify guidelines making recommendations about use of tests, methods papers, and systematic reviews of test accuracy. We also contacted experts in research about HCTDS to identify additional tools.

Results
We identified 29 tools and 14 modifications of existing tools for assessing QoE and SoR. Twenty three out of 29 tools acknowledge the importance of assessing the QoE and SoR separately but in 8, SoR is based solely on QoE. Most tools include several individual quality criteria but no tool rates all of the quality criteria suggested by the GRADE working group. When making decisions about the use of tests, patient values and preferences and impact on resource utilization were considered in 6 and 8 out of 29 tools respectively. There is also confusion about the terminology that describes the various factors that influence the QoE and SoR.

Conclusions
Most tools identified acknowledge the importance of assessing the QoE and SoR separately but for many of them the SoR is based solely on QoE. Although at least one approach includes all criteria for assessing QoE and determining SoR that we identified, a more
detailed guidance about how to operationalize these assessments and make related judgments would be beneficial. There is a need for a better description of the framework for using evidence from systematic reviews to make decisions and develop recommendations about HCTDS.
WHAT IS NEW

- We identified 43 tools and modifications of existing tools to assess the quality of evidence (QoE) and strength of recommendations (SoR) about HCTDS.

- Although the GRADE approach includes all identified domains, a more detailed guidance about how to operationalize individual criteria and make related judgments would be beneficial.

- Most guideline development frameworks that we evaluated do not include all factors needed to make decisions (recommendations) about the use of tests.

- There is confusion about the terminology that describes various factors/domains/criteria that influence the QoE and SoR.
BACKGROUND

In two previous articles in this series addressing diagnostic decision-making in health care, we provided an overview of known roles and applications of tests, different diagnostic study designs and the main challenges encountered in conducting and applying results of diagnostic research. We also described the research supporting the development of evidence tables for presenting results of systematic reviews of test accuracy. In this third article in a series of eight articles we identify and appraise tools used to assess the quality of evidence (QoE) and/or strength of recommendations (SoR) about health care-related tests and testing strategies (HCTDS). In this article “tests” for simplicity of communication refers to all healthcare related tests and diagnostic strategies that are used for different application and roles and not necessarily to make a diagnosis sensu stricto.

Guidelines pertaining to HCTDS aim to assist health care providers in making decisions about appropriateness of using tests in the management scheme for patients. They also aim to assist policy makers in making decisions about the appropriateness of using tests for the screening of, diagnosis of, staging, or evaluating treatment for specific clinical conditions among other applications, at the population level. The challenges that systematic reviewers and guideline developers face when drawing conclusions and making recommendations about tests are considerably different when compared to treatment decisions. This is particularly true as it relates to challenges in assessing the QoE, also known as strength of evidence, certainty in the evidence or confidence in effects, and the frequent lack of a direct link between test accuracy results and patient important outcomes.

In 2002, the US Agency for Health Care Research and Quality (AHRQ) published a systematic review ¹ that summarized the evidence, available until 2000, on tools for evaluating the quality of individual studies, including diagnostic studies, and evidence
grading systems. An update of the evidence available until 2007 was conducted by the Canadian Agency for Drugs and Technologies in Health (CADTH)\(^2\). In this review we compile the most up to date list of available tools to assess QoE. We focus our work on tools that evaluate evidence about HCTDS. Our objective was to identify and summarize available instruments, checklists, critical appraisal tools, and frameworks published in the literature that are designed for assessing the QoE and/or SoR dealing with HCTDS and compare these tools.
METHODS

Search strategy

The AHRQ and the CADTH reviews described the challenges in conducting searches of the literature for this purpose. Originally, the AHRQ review questioned the usefulness of searching bibliographic databases for individual quality assessment and evidence grading systems, and documented most systems were not identified through this process (only 30 of the final 121 systems that were reviewed). Identifying the key weakness of electronic searching for individual quality assessment and evidence grading systems, as described in the AHRQ review, is the relatively undeveloped state of the National Library of Medicine's Medical Subject Headings (MeSH) in expressing methodological concepts inherent to quality assessment and evidence-based medicine more generally. In 2005, the Canadian Optimal Medication Prescribing and Utilization Services (COMPUS) team's preliminary searches confirmed the assessment of the AHRQ report. Therefore, they concluded that any MeSH-based search remains largely ineffective as a means of identifying individual systems, resulting in large numbers of false hits and unknown levels of comprehensiveness, even when running a very sensitive search. In these two reviews the groups depended on consultation processes to identify relevant articles and publications.

Building upon the search strategies used by the AHRQ and the CADTH teams, we worked with the author of the CADTH report and with two experienced information specialists who developed a comprehensive systematic search strategy to identify grading systems for evidence concerning the use of tests. We have modified the CADTH search strategy to perform an even more sensitive search for available systems and used very extensive consultation strategies to address the issues raised by AHRQ and CADTH (the detailed search strategy is provided in Appendix 1).
Data sources

We applied the following strategies to identify relevant tools:

1. Performed a systematic review of published methods for grading diagnostic evidence
2. Reviewed prior systematic reviews of test accuracy published between January 2011 and July 2012 to identify systems used in the existing literature
3. Compiled a list of systems already identified in the AHRQ and CADTH reports and compared their criteria
4. Reviewed grading systems used in guidelines making recommendations about tests (we searched websites of professional societies in the areas of radiology, laboratory medicine and pathology
5. Reviewed the reference lists of methodology papers and systematic reviews that listed quality assessment and grading systems
6. Consulted experts in the area of HCTDS

Inclusion and exclusion criteria

We included articles that 1) discussed methods for evaluating QoE and/or SoR about HCTDS or 2) rated, graded, or assessed QoE and/or SoR about HCTDS (i.e. diagnostic evidence syntheses including systematic reviews, health technology assessments, guidelines among others). We included articles published in any language and translated tools published in languages other than English. We excluded tools that have not been developed or used to assess evidence and make decisions about HCTDS
**Data abstraction**

For the systematic reviews, we conducted title and abstract screening followed by full text screening and data abstraction in duplicate. Quality assessment tools were identified at data abstraction, and once named, an Internet search was conducted to identify additional sources of information such as methodology papers, manuals, or handbooks for the tools. Additionally, we contacted professional societies and organizations requesting information on any updates, whether published or not. Following data abstraction, we requested clarifications about inconsistencies or ambiguity in the manuals or methods papers. Inevitably, we identified modifications of existing tools. When we identified a modification, we carefully evaluated the rationale by fully extracting information. We then organized our findings into themes that summarized any differences.
RESULTS

Assessment of Identified Tools

From the systematic searches and consultations, we identified 29 tools and 14 modifications that are used to assess the QoE and/or SoR about HCTDS. Figure 1 summarizes the results of the search. We screened 5534 titles and abstracts, 1004 full text articles, and abstracted data from 205 references for tools and 125 references for their modifications. We reviewed the quality assessment tools and evidence grading systems identified in the AHRQ and CADTH reports, and included those that met our inclusion criteria. Appendix 2 summarizes these tools and provides reasons for exclusion for those not meeting our criteria. Additionally, we identified 10 methodological papers and systematic reviews that discussed multiple tools and we searched their reference lists 3-12. A complete list of the identified tools with references assessed per tool is available in Appendix 3. A complete list of excluded articles with reasons for exclusions and references is available in Appendix 4.

Table 1 summarises the 29 tools for rating QoE and SoR. Out of the 205 references that were included, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach was discussed in 52 references, Oxford in 24, United States Preventive Services Task Force (USPSTF) in 16, Scottish Intercollegiate Guidelines Network (SIGN) in 13, and the Australian National Health and Medical Research Council (NHMRC) in seven. While many of the tools identified take into consideration the same, or similar, criteria or factors for assessing the QoE and determining the SoR, each of the tools differed slightly in the overall approach despite the overlap in the factors considered.

We demonstrate this in the description that follows of the 5 commonly used tools. Assessment of the QoE using the GRADE approach begins with the study design, and QoE is
downgraded for risk of bias, imprecision, indirectness, inconsistency, and publication bias, or upgraded for magnitude of effect, effect of plausible confounding, and the presence of a dose-response gradient. Rating of the SoR then takes into account the QoE, the balance of desirable and undesirable consequences, patients’ values and preferences, and resource implications among other factors.

For the 2009 version of the Oxford Centre for Evidence-based Medicine tool, assessment of the QoE includes consideration of study design, risk of bias, imprecision, and inconsistency, with the SoR based on the level of evidence, consistency of the evidence, and indirectness. The 2011 update the Oxford tool no longer includes criteria for rating SoR, and instead includes consideration of inconsistency and indirectness as part of the QoE assessment, with the additional consideration of the magnitude of effect for downgrading or upgrading the quality.

The assessment tool from SIGN takes into consideration the study design and risk of bias for assessment of QoE, and the rating of SoR follows the considered judgment process, which takes into consideration a number of factors including the number of studies (volume of evidence, as described by the SIGN tool), consistency of results, applicability to the target population, generalizability, the potential clinical impact of an intervention, potential harms of implementation, effects on equality groups, and implementability.

The tool developed by the USPSTF, frames the assessment of the QoE as the level of certainty regarding the net benefit. The assessment consists of multiple criteria related to studies that form the body of evidence, including study design, risk of bias, consistency of results, the number and size of individual studies, magnitude and direction of the effect, and additional factors such as dose-response effects, as well as a broader assessment of the body of evidence that includes consideration of gaps in the chain of evidence, generalizability of the findings to practice, and lack of information on important health
outcomes. Rating of SoR takes into consideration the level of certainty (i.e. the quality of the evidence) and the magnitude of the net benefit of an intervention.

The tool from the NHMRC includes a number of criteria that are considered in the tools described above as part of the assessment of QoE, including study design, quantity of the evidence, risk of bias, consistency, clinical impact, directness, and applicability. The assessment of the SoR then includes as the sole factor the strength of the body of evidence.

Although 23 out of 29 tools acknowledged the importance of assessing both the QoE and SoR, in 8 SoR is based solely on QoE. Moreover, for 12 of the 29 tools the QoE assessment consisted of hierarchies of evidence determined by study design and evaluation of risk of bias, and in some tools the number of studies, without consideration of factors such as consistency, directness, or precision of study results. In 4 of these 12 tools consideration of consistency, directness, or precision was included as part of the rating of SoR, but no tool included all three factors. When making recommendations about use of tests, balance of benefits and harms as a factor were explicitly considered in 8 tools\(^{17,25-31}\), patient values and preferences were considered in 6 tools and resource implications were considered in 8 tools.

When evaluating modifications of existing tools, we failed to identify rationale for these modifications. Also, we were not able to find guidance for the minimum requirements needed to apply, and reasons to adopt or adapt a certain tool. Our work shows examples of use or “misuse” of modifications of existing tools. We observed violations of the conceptual principles of the original tools. In one of the examples, for a tool that claimed using similar but not identical criteria to the GRADE approach\(^ {32}\), the authors described guidelines for interpreting statistical test results from meta-analyses based on \(p\)-values and the number of studies included. This was not only a clear violation of the conceptual principles of GRADE which do not suggest relying on \(p\)-values, but it was also not clear how one can use
a \( p \)-value to assess the precision of the effect of HCTDS. This was particularly confusing in the context of the observations that the meta-analyses do not address patient important outcomes directly but only summarize diagnostic test accuracy results (e.g. sensitivity and specificity). Given the low yield, we terminated evaluation of the remaining modified tools after reviewing 14 modified tools in duplicate. Table 2 summarizes examples of the themes we identified for the main modifications of existing tools.
DISCUSSION

In this systematic review we identified, summarized and compared instruments, checklists, critical appraisal tools, and frameworks designed for assessing the QoE and SoR dealing with HCTDS. The most commonly referenced tools are GRADE, Oxford, USPSTF, SIGN and NHMRC. Although most tools acknowledged the importance of assessing the QoE and SoR separately, for many of them the SoR is based solely on QoE. When moving from evidence to recommendations, patient values and preferences and resource implications are infrequently considered.

Our review has multiple strengths. First, we used a very comprehensive and inclusive literature search that included electronic databases, societies and organizations’ websites, and grey literature. Second, we contacted authors and representatives of societies to clarify any questions about major tools. Third, we included all tools regardless of language of publication. Fourth, we collected information about each tool from multiple sources and we requested unpublished updates and modifications.

Our review has a few limitations. Abstracting data about each tool required judgment. This was magnified by inconsistencies within manuals, partial updates of manuals, the availability of multiple versions of the same manual without clear guidance on which one to follow, and updates of the tools without updating the manuals at the same time. This sometimes led to confusion about the most accurate and up-to-date description of the tool. To minimize subjectivity and to ensure our abstraction was accurate, we contacted experts from the societies and groups representing or using these tools to confirm our judgments.

In more than one instance, these representatives appreciated our comments and correspondence as they felt our extensive review of the documents addressing their tools served as an independent review of these documents.
When assessing how tools were applied in guidelines and systematic reviews, we commonly observed deviations from manuals which also create confusion about how users apply these tools when they make decisions and develop recommendations. Also, the lack of clear guidance about processes and the steps on how to use the tools in practice and the absence of regulations for the minimum requirements needed to apply certain tools have likely contributed to the variability in implementing these tools. The inconsistency in terminology and vagueness of concepts might have also contributed to the confusion in the literature. For example, “balancing benefits and harms” was a common consideration mentioned in the tools but it was often not clear how this balancing might be performed when only evidence about diagnostic accuracy of a test was available. In tools that provided guidance for “balancing benefits and harms” different factors were considered. For example, some tools included assessment of cost effectiveness or referred to direct harms from the test itself, like radiation exposure or complications.

Most tools include individual factors/domains/criteria for assessing QoE suggested in the GRADE approach but none rated all of them. Although the GRADE approach to rating the QoE and SoR about tests does not provide mature guidance about all the quality criteria and factors to consider when making recommendations, we believe it appears to be the most complete of the approaches that we reviewed. Our results, similar to the review by Gopalakrishna et al\textsuperscript{33}, show that some tools address certain domains in more detail and more explicitly describe how to consider clinical pathways. Our review also confirm that many tools share similar concepts and criteria but the approach and the processes to make a decision differ significantly.

When assessing modifications of existing tools, we failed to understand the rationale underlying such alterations and encountered inconsistency and vague reporting. Our work shows examples of use or “misuse” of modifications of existing tools. Thus, with the
exception of the tools that just changed the labeling of the domains or categories of the QoE or SoR to a more familiar labeling to their users, we observed clear misrepresentation of original tools.

The field of decision making about HCTDS is rapidly evolving and those making decisions about the use of tests in health care need to keep up to date with new methodological research. They also need to be explicit about the process of developing recommendations about tests. Assessment of the adherence of decision makers and guideline developers to their own manuals is essential to ensure appropriate application of the tools. To reduce confusion attention should also be paid to keeping manuals updated, consistent and easy to understand, clear, while removing or clearly archiving out-dated manuals and documents. Additionally, there is a need for a better description of the framework for using evidence from systematic reviews to make decisions and develop recommendations about HCTDS.
Acknowledgement

We thank Dr. Susan Jack, associate professor of nursing, McMaster University, for her assistance with methods of qualitative data analysis. We also thank Mr. Vijay Shukla and Ms. Hayley Fitzsimmons from CADTH for their feedback about our search strategy. This work was partially funded by the German Insurance Fund agency as part of a larger project about decision-making for HCTDS. The views presented here are those of the authors and should not be attributed to the funding agency or its staff.
Figure 4.1: PRISMA Flow Diagram of Studies
### TABLE 4.1: TOOLS SUMMARY WITH DOMAINS IN RATING QOE AND SOR

<table>
<thead>
<tr>
<th>Tool/Organization</th>
<th>Domains for Evaluating Quality of the Body of Evidence</th>
<th>Domains for Rating Strength of Recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GRADE</td>
<td>• Study Design&lt;br&gt;• Risk of Bias&lt;br&gt;• Imprecision&lt;br&gt;• Indirectness&lt;br&gt;• Inconsistency&lt;br&gt;• Publication Bias&lt;br&gt;Upgrading for:&lt;br&gt;• Magnitude of effect&lt;br&gt;• Effect of all plausible confounding&lt;br&gt;• Dose-response gradient</td>
<td>• Quality of evidence&lt;br&gt;• Balance of desirable and undesirable consequences&lt;br&gt;• Values and Preferences&lt;br&gt;• Resource Implications</td>
</tr>
<tr>
<td>2. Oxford 2011</td>
<td>• Study design&lt;br&gt;• Risk of bias/study quality (e.g. reference standard and blinding)&lt;br&gt;• Magnitude of effect for</td>
<td><em>Not applicable, no rating of strength of recommendations</em></td>
</tr>
<tr>
<td>Observational Study with ‘dramatic effect’</td>
<td>Downgrading for:</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>• Imprecision</td>
<td>• Indirectness</td>
<td></td>
</tr>
<tr>
<td>• Indirectness</td>
<td>• Inconsistency between studies</td>
<td></td>
</tr>
<tr>
<td>Expert opinion as LoE 5</td>
<td>(“mechanism-based reasoning”)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Risk of Bias/study quality (e.g. good reference standard, &gt;80% follow-up)</td>
</tr>
<tr>
<td></td>
<td>• Imprecision (“wide CI”)</td>
</tr>
<tr>
<td></td>
<td>• Inconsistency (“troublesome heterogeneity”)</td>
</tr>
<tr>
<td>Expert opinion without explicit critical appraisal as LoE 5</td>
<td>• Level of evidence</td>
</tr>
<tr>
<td></td>
<td>• Consistency</td>
</tr>
<tr>
<td></td>
<td>• Indirectness (“extrapolation from studies where data is used in a situation that has potentially clinically important differences than the original study situation”)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. SIGN</th>
<th>• Study design</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Risk of bias</td>
</tr>
<tr>
<td></td>
<td>• Volume of evidence</td>
</tr>
<tr>
<td></td>
<td>• Consistency of results</td>
</tr>
<tr>
<td>5. USPSTF</td>
<td>Level of certainty regarding net benefit (high, moderate, low):</td>
</tr>
<tr>
<td></td>
<td>• Study design (“appropriate study design to answer question”)</td>
</tr>
<tr>
<td></td>
<td>• Risk of Bias</td>
</tr>
<tr>
<td></td>
<td>• Consistency of results</td>
</tr>
<tr>
<td></td>
<td>• Lack of coherence</td>
</tr>
<tr>
<td></td>
<td>• Number of studies (i.e. volume of evidence)</td>
</tr>
<tr>
<td></td>
<td>• Size of individual studies (i.e.</td>
</tr>
<tr>
<td></td>
<td>Level of certainty regarding net benefit (i.e. level of evidence)</td>
</tr>
<tr>
<td></td>
<td>Magnitude of net benefit (i.e. balance of benefit vs. harm)</td>
</tr>
</tbody>
</table>

- **Applicability** (“directly applicable to target population”)
- **Generalizability**
- **Clinical Impact** (Size of patient population (i.e. disease burden), magnitude of effect, relative benefit, resource implications, balance of risk and benefit)
- **Potential Harms of Implementation**
- **Equality Groups**
- **Implementability**
<table>
<thead>
<tr>
<th></th>
<th>precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Gaps in chain of evidence</td>
<td></td>
</tr>
<tr>
<td>• Generalizability of findings to routine primary care practice</td>
<td></td>
</tr>
<tr>
<td>• lack of information on important health outcome</td>
<td></td>
</tr>
<tr>
<td>• Magnitude and direction of effect</td>
<td></td>
</tr>
<tr>
<td>• Additional factors that assist in drawing conclusions, e.g. presence or absence of dose-response effect, fit within biological model</td>
<td></td>
</tr>
</tbody>
</table>

| 6. NHMRC    | Study design |
|            | • Study design |
|            | • Quantity of evidence (e.g. one or more studies) |
|            | • Risk of bias |
|            | • Consistency |
|            | • Clinical impact |
|            | • Directness (i.e. generalizability/external validity; population in studies) |
|            | Strength of the body of evidence underpinning the recommendation |
| 7. Strength of Recommendation Taxonomy (SORT) | Similar to target population)  
• Applicability (directly applicable to Australian healthcare context) |  
• Study design  
• Risk of bias  
• Directness (i.e. patient-oriented evidence)  
• Consistency in study findings |  
• Level of evidence (i.e. strength of evidence) |
| 8. Agency for Health Care Policy and Research (AHCPR) (as used by organizations such as the American Burn Association)* | Study design  
• Sample size |  
• Number of studies  
• Clear cut results  
• Consistency across studies |
| 9. American Academy of Neurology (AAN) | Study design  
• Risk of bias |  
• Clinical impact (e.g. “adherence expected to affect”)  
• Variation in patient preferences  
• Cost  
• Availability  
• Value of benefit relative to risk  
• Strength/level of evidence (i.e. “confidence in evidence”)  
• Strength of principle-based inferences |
| 10. American Academy of Paediatrics (AAP) | • Study design  
• Risk of bias  
• Consistency of studies  
• Magnitude of effect that the studies detect  
• Individual and aggregate sample sizes of these studies | • Evidence quality/strength  
• Benefit-harm assessment  
• Cost of adherence to recommendation |
|------------------------------------------|-------------------------------------------------------------------------------------------------|
| 11. American Academy of Sleep Medicine (AASM) | • Study design  
• Risk of bias | • Level of evidence  
• High degree of clinical certainty  
• Directness |
| 12. American Association of Clinical Endocrinologists (AACE) | • Study design  
• Level of evidence  
• Risk of bias  
• Interpretation of results: generalizability, logical, incompleteness, validity  
• Two-thirds consensus  
• Directness  
• Other qualifiers: cost-effectiveness, risk-benefit analysis, evidence gaps, alternative physician preferences, alternative recommendations |
| 13. American College of Radiology (ACR) | • Study design  
• Risk of bias ("conclusions of study valid, inconclusive, not valid") | Radiologic Procedure Appropriateness:  
• Balance of risk and benefits  
• "Supplemented by expert opinion" |
| 14. American Heart Association (AHA) | • Study design  
• Risk of bias  
• Quantity of studies | • Estimate of certainty (precision) of treatment effect (Level of Evidence)  
• Size of treatment effect: Benefit vs. risk assessment |
| 15. American Urological Association (AUA) | • Study design  
• Risk of bias  
• Consistency of findings across studies ("conflicting information")  
• Precision (i.e. "adequacy of sample sizes")  
• Generalizability of samples, settings, and treatments | • Strength of body of evidence  
• Balance between benefits and risks/burdens |
| 16. Canadian Diabetes Association (CDA) | • Risk of bias | • Level of Evidence |
| 17. Canadian Task Force on Preventive Health Care (CTFPHC) | • Study design  
• Number of studies  
• Risk of bias  
• Sample Sizes  
• Direction, Magnitude and Significance of Effects  
• Consistency of the Results | • Relative Strength of Studies  
Downgraded for:  
• Applicability to Canadian population and representation of sub-groups  
• Consensus of the committee members of concerns about the recommendation  
| 18. Diagnostic Imagining Pathways | • Study design  
• Risk of bias  
• quality of statistical analysis  
• quality of reporting  
• Generalizability | • Level of evidence  
• Availability of technology  
• Compliance with procedure  
• Potential for harm  
• Burden of disease  

Not applicable, no rating of strength of recommendations  
| 19. Dutch Institute for Healthcare Improvement  
(Centraal Begeleidings Orgaan; CBO) | • Sample size  
• Methodological assessment (Risk of bias)  
• Number of studies | • Level of evidence  
• Clinical relevance  
• Harms/side effects |
| 20. Evaluation of Genomic Applications in Practice and Prevention (EGAPP) | Quality of evidence for the chain of evidence:  
• Study design  
• Risk of bias  
• Number of studies  
• Consistency  
• Generalizability | • Availability  
• Organizational issues  
• Costs  
• Patient and professional perspectives  
• Legal aspects and ethics |
|---|---|---|
| 21. European League Against Rheumatism (EULAR) | • Study design  
• quantity of studies  
• validity  
• reliability  
• DTA results (e.g.: sensitivity, specificity, likelihood ratio)  
• reference standard  
• age range of patients  
• prevalence in the source population | EULAR uses term 'propositions’  
• Level of evidence  
• Clinical expertise (logistics, patient perceived acceptance, and tolerability) |
<table>
<thead>
<tr>
<th>22. European Federation of Neurological Societies (EFNS)</th>
<th>Study design</th>
<th>Level of evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of bias</td>
<td>Consistency between studies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of studies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>23. German Federal Joint Committee (Gemeinsame Bundesausschuss; G-BA)</th>
<th>Study design</th>
<th>Not applicable, no rating of strength of recommendations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of bias</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>24. International Liaison Committee on Resuscitation (ILCOR) Consensus Document on the Science of Cardiopulmonary Resuscitation with Treatment (CoSTR)</th>
<th>Study design</th>
<th>Do not rate SoR, but in formulating recommendations, consider:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Risk of bias</td>
<td>Magnitude of effect</td>
</tr>
<tr>
<td>Felipe</td>
<td>Directness</td>
<td>Precision</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The outcome affected</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Generalizability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Potential barriers to implementation (cost, education, logistics)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>25. International Scientific Society on Scoliosis Orthopaedic and Rehabilitation Treatment (SOSORT)</th>
<th>Study design</th>
<th>Clinical importance of the recommendation as determined by consensus of guideline panel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of studies</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Risk of bias</td>
<td></td>
</tr>
</tbody>
</table>

Additional considerations (not used consistently):
- risk ratio
- incremental cost-effectiveness
- perspective
<table>
<thead>
<tr>
<th>Study</th>
<th>Categories</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two levels of expert agreement: 90% agreement, 70-89% agreement</td>
<td>Balancing all typical factors involved in a clinical decision (patients, professional, social) (not clear what types of factors and how considered.</td>
<td></td>
</tr>
</tbody>
</table>

26. Italian National Guidelines System (Programma Nazionale Linee Guida; PNLG)  
- Study design  
- Risk of bias  
- Number of studies  
- Quality of evidence

27. Let Evidence Guide Every New Decision (LEGEND)  
- Study design  
- Number of studies  
- Risk of bias  
- Consistency of Evidence  
- Not applicable, no rating of strength of recommendations

28. New Zealand Guidelines Group (NZGG)  
- Study design  
- Risk of bias  
- Volume of evidence  
- Consistency  
- Applicability  
- Clinical impact (potential clinical impact given size of population, magnitude of effect, balance of benefit vs risk, resource implications)  
- Strength of the overall evidence
<table>
<thead>
<tr>
<th>Other factors (not specified)</th>
<th>Study design</th>
<th>Level of evidence (Study design, Risk of bias, Study sample size)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oncology Nursing Society (ONS)</td>
<td>Risk of bias</td>
<td>Number of studies</td>
</tr>
<tr>
<td></td>
<td>Study sample size</td>
<td>Studies conducted at more than one institution</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Consistency</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benefit versus Harm and Burdens/Costs</td>
</tr>
</tbody>
</table>

* AHCPR is former Agency for Healthcare, Research & Quality (AHRQ), which is now using an adaptation of GRADE-EPC.
### TABLE 4.2: MAIN THEMES OF THE MODIFICATION OF EXISTING TOOLS TO ASSESS QOE AND SOR FOR HCTDS

<table>
<thead>
<tr>
<th>Theme</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decrease the number of levels of QoE</td>
<td>ACP adaptation of GRADE&lt;sup&gt;34&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>ACCP adaptation of GRADE&lt;sup&gt;35&lt;/sup&gt;</td>
</tr>
<tr>
<td>Change the name of the levels from letters to numbers or vice versa</td>
<td>ACCP adaptation of GRADE&lt;sup&gt;35&lt;/sup&gt;</td>
</tr>
<tr>
<td>Changing the individual QoE or SoR domains</td>
<td>GRADE NACB&lt;sup&gt;36&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>GRADE Adaptation – Working Group on Prosthetic Joint Infections (Gruppo Italiano di Studio sulle Infezioni Gravi; GISIG)&lt;sup&gt;37&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Oxford ICUD&lt;sup&gt;38,39&lt;/sup&gt;</td>
</tr>
<tr>
<td>Change the definition of the domains but keep the same name</td>
<td>GRADE CHERG&lt;sup&gt;32&lt;/sup&gt;</td>
</tr>
<tr>
<td>Change the name of the tools with no change in any domains, or present tool with no name or reference to original tool</td>
<td>Journal of bone and joint surgery (Oxford)&lt;sup&gt;40&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>North American Spine Society (oxford)&lt;sup&gt;41&lt;/sup&gt;</td>
</tr>
<tr>
<td>Combine different tools, using one for QoE and one for SoR</td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX 4.1: SEARCH STRATEGIES FOR PUBLISHED LITERATURE ON EVIDENCE GRADING SYSTEMS

Database: Ovid MEDLINE(R) without Revisions <1996 to June Week 4 2012>

1 evidence based medicine/ (46020)
2 practice guidelines as topic/ (64802)
3 research design/ (429)
4 meta-analysis as topic/ or review literature as topic/ (13288)
5 1 or 2 or 3 or 4 (153868)
6 (evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates).ti,ab. (2232073)
7 (evidence or recommendation or recommendations or conclusions or conclusion or systematic review or review or reviews or meta-analysis).ti,ab. (2827550)
8 (strength or level or levels or grade or grades or hierarch* or quality).ti,ab. (1920185)
9 ((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (evidence or recommendation or recommendations or conclusions or conclusion or systematic review or review or reviews or meta-analysis))).ti,ab. (47289)
10 ((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (strength or level or levels or grade or grades or hierarch* or quality) adj10 (evidence or recommendation or recommendations or conclusions or conclusion or systematic review or review or reviews or meta-analysis))).ti,ab. (14228)
11 9 or 10 (57410)
Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations <July 06, 2012>

1. ((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (evidence or recommendation or recommendations or conclusions or conclusion or systematic review or review or reviews or meta-analysis)).ti,ab. (4332)

2. ((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (strength or level or levels or grade or grades or hierarch* or quality) adj10 (evidence or recommendation or recommendations or conclusions or conclusion or systematic review or review or reviews or meta-analysis)).ti,ab. (1689)

3. 1 or 2 (5523)

4. 3 and (guidelines or guideline).ti,ab. (401)

5. limit 4 to yr="2007 -Current" (349)

Database: Embase <1996 to 2012 Week 27>

1. practice guideline/ (196678)

2. review/ (1373473)

3. meta-analysis/ (59931)

4. evidence based medicine/ (73706)

5. methodology/ or method*.tw. (3826722)

6. 1 or 2 or 3 or 4 (1592135)
((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (evidence or recommendation or recommendations)).ti,ab. (18325)

((evaluate or evaluates or evaluation or evaluating or grading or grade or grades or assess or assesses or assessment or assessing or rate or rating or rates) adj4 (strength or level or levels or grade or grades or hierarch* or quality) adj10 (evidence or recommendation or recommendations)).ti,ab. (5540)

8 or 9 (20941)

10 10 and 7 (5180)

12 limit 11 to yr="2007 -Current" (3001)

Total citations: 6139

Total after duplicates removed: 5217

Published Diagnostic Systematic Reviews

Database: Embase <1996 to 2012 Week 26>, Ovid MEDLINE(R) without Revisions <1996 to June Week 4 2012>, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations <July 05, 2012>

Search Strategy:

1  systematic review.ti. (44771)

2  meta analysis.ti. (42153)

3  1 or 2 (78336)

4  (sensitivity and specificity).ab. (202140)

5  likelihood ratio*.ab. (14120)

6  (receiver operator characteristic or receiver operating characteristic or receiver operator...
characteristics or receiver operating characteristics or roc or roc curve).ab. (65907)

7  predictive value*.ab. (102052)

8  or/4-7 (300001)

9  3 and 8 (3643)

10  limit 9 to yr="2011 -Current" (1100)

11  remove duplicates from 10 (688)

**Total citations after duplicates removed: 688**
# APPENDIX 4.2: QUALITY RATING AND EVIDENCE GRADING SYSTEMS IDENTIFIED IN AHRQ AND CADTH REPORTS

<table>
<thead>
<tr>
<th>Tool Name or Reference</th>
<th>Tool Name in review</th>
<th>Included (I) or Excluded (E) in Review</th>
<th>Reason for Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>AHRQ (2002) Systems to Rate The Quality of Individual Diagnostic Studies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sheps and Schechter 1984; Arroll 1988</td>
<td>Not identified; not identified</td>
<td>E; E</td>
<td>Both articles are systematic reviews of diagnostic test accuracy studies. They use a checklist to assess whether the studies fulfill 7 ‘well-known’ methodological criteria for appraising diagnostic tests (e.g. a well-defined gold standard).</td>
</tr>
<tr>
<td>Hoffman 1991</td>
<td>Not identified</td>
<td>E</td>
<td>Full text not available. Review of diagnostic accuracy studies on thermography for diagnosing lumbar radiculopathy. Reported methodological flaws in studies according to checklist.</td>
</tr>
<tr>
<td>Cochrane Methods Working Group 1996</td>
<td>Identified as Risk of Bias tools: Cochrane Handbook for Systematic Reviews of Interventions Risk of Bias Tool; Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Reference</td>
<td>Description</td>
<td>Method</td>
<td>Results</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------</td>
<td>--------</td>
<td>---------</td>
</tr>
<tr>
<td>Khan 2000</td>
<td>Khan 2001 – Centre for Reviews and Dissemination Report No. 4</td>
<td>E</td>
<td>Not a tool. Provides a list of potential biases in diagnostic test accuracy studies (e.g. biases in test methods, in application of reference standard, in interpretation and analysis).</td>
</tr>
<tr>
<td>NHMRC 2000</td>
<td>NHMRC</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Harbour and Miller 2001</td>
<td>SIGN</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Working Group on methods for prognosis and decision making 1990</td>
<td>Köbberling Framework</td>
<td>E</td>
<td>Not a tool</td>
</tr>
<tr>
<td>Pinson 1991</td>
<td>Not identified</td>
<td>E</td>
<td>SR of 99mTc-RBC venography in diagnosis of DVT. Reviewed studies according to 7 methodological standards (e.g. avoidance of workup bias).</td>
</tr>
<tr>
<td>Carruthers 1993</td>
<td>Sackett 1989</td>
<td>I¹</td>
<td>n/a</td>
</tr>
<tr>
<td>Jaeschke 1994</td>
<td>Sackett 1994²</td>
<td>I¹</td>
<td>n/a</td>
</tr>
<tr>
<td>Irwig 1994</td>
<td>Not identified</td>
<td>E</td>
<td>Not a tool. Guidance paper on conducting a meta-analysis of a</td>
</tr>
<tr>
<td>Author</td>
<td>Evaluation Tool</td>
<td>Evaluation Type</td>
<td>Rating</td>
</tr>
<tr>
<td>--------------------------------</td>
<td>--------------------------</td>
<td>-----------------</td>
<td>--------</td>
</tr>
<tr>
<td>Reid 1995</td>
<td>Not identified</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>Bruns 1997</td>
<td>STARD (draft work)</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>McCrory 1999</td>
<td>AHCPR</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Ross 1999</td>
<td>AHCPR</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Goudas 2000; Lau 2000</td>
<td>AHCPR</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td><strong>AHRQ (2002) Systems to Rate The Strength of a Body of Evidence</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canadian Task Force on the Periodic Health Examination 1979</td>
<td>Canadian Task Force on Preventive Health Care (formerly the Canadian Task Force on the Periodic Health Examination)</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Anonymous 1981 – McMaster University Department of Clinical Epidemiology and Biostatistics</td>
<td>Not identified</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Methodology</td>
<td>Level</td>
<td>Notes</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>--------------------------------------</td>
<td>-------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Cook et al. 1992; Sackett 1989</td>
<td>Sackett 1989</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>U.S. Preventive Services Task Force 1996</td>
<td>USPSTF</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Ogilvie et al. 1993</td>
<td>Sackett 1989</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Gross et al. 1994; Bartlett et al. 1998</td>
<td>Not identified</td>
<td>E</td>
<td>Infectious Diseases Society of America (IDSA) 1994 QoE and SoR classification. Generic tool based on study design. IDSA currently using GRADE.</td>
</tr>
<tr>
<td>Gyorkos et al. 1994</td>
<td>Not identified</td>
<td>E</td>
<td>Level of evidence based on study validity, strength of association, precision, variability of findings between studies. Not specific to diagnosis. Full text not available.</td>
</tr>
<tr>
<td>Guyatt et al. 1995</td>
<td>Not identified</td>
<td>E</td>
<td>Article from the Users’ Guides to the Medical Literature specific to grading recommendations for therapy.</td>
</tr>
</tbody>
</table>
| Granados et al. 1997                                                 | Not identified                        | E     | EUR-ASSESS Subgroup on Dissemination and Impact recommendations for HTA agencies. Simple 3-level strength of evidence hierarchy (strong-based on empirical evidence,
<table>
<thead>
<tr>
<th>Reference</th>
<th>Methodology</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gray 1997</td>
<td>Not identified</td>
<td>E</td>
<td>Book on evidence-based healthcare. Suggests strength of evidence based on study design and execution.</td>
</tr>
<tr>
<td>Djulbegovic and Hadley 1998</td>
<td>Not identified</td>
<td>E</td>
<td>Propose quality of evidence scoring for guidelines based on study design and error rate. Full text not available.</td>
</tr>
<tr>
<td>Edwards et al. 1998</td>
<td>Not identified</td>
<td>E</td>
<td>Propose a classification system that assesses weight of evidence based on signal (effect size) to noise (methodological weaknesses) rather than a simple hierarchy based on study design.</td>
</tr>
<tr>
<td>Bril et al. 1999</td>
<td>Not identified</td>
<td>E</td>
<td>Evidence grades based on trial design (e.g. RCT, controlled trial, open trial) adapted from USPSTF.</td>
</tr>
<tr>
<td>Chesson et al. 1999</td>
<td>AASM – Old Tool</td>
<td>E</td>
<td>Adapted from Sackett 1989. Classification of Evidence and Grades of Recommendation based on study design.</td>
</tr>
<tr>
<td>Clarke and Oxman 1999</td>
<td>Cochrane Handbook for Systematic Reviews of Interventions 1999</td>
<td>E</td>
<td>Guidance on the strength of inference about the effectiveness of an intervention, including study quality, magnitude of effect, consistency, dose-response relationship, supporting indirect evidence, and no other plausible explanation.</td>
</tr>
<tr>
<td>Hoogendoorn et al. 1999</td>
<td>Not identified</td>
<td>E</td>
<td>AHCPR adaptation with 3 levels of evidence based on study quality, quantity and consistency.</td>
</tr>
<tr>
<td>Working Party for Guidelines for the</td>
<td>Not identified</td>
<td>E</td>
<td>Grading of evidence based on study design and expert consensus. Full text not available.</td>
</tr>
<tr>
<td>Study Description</td>
<td>Evidence Quality</td>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------------------</td>
<td>==================================================================================================================</td>
<td></td>
</tr>
<tr>
<td>Management of Heavy Menstrual Bleeding 1999</td>
<td>E</td>
<td>Methods paper on guideline development. Suggests simple scheme for categories of evidence based on quantity and study design and strength of recommendations scheme based on category of evidence and directness.</td>
<td></td>
</tr>
<tr>
<td>Shekelle et al. 1999</td>
<td>E</td>
<td>American Academy of Ophthalmology guideline regarding prevention or retinal detachment. Includes strength of evidence classification based on study design, study quality, directness, and expert opinion.</td>
<td></td>
</tr>
<tr>
<td>Wilkinson 1999</td>
<td>E</td>
<td>Systematic review of physical risk factors for neck pain. Developed observational study quality checklist and levels of evidence based on study design, quality, quantity, and consistency.</td>
<td></td>
</tr>
<tr>
<td>Briss et al. 2000</td>
<td>E</td>
<td>Methods article about the Institute for Clinical Systems Improvement (ICSI) evidence grading system. Developed system tailored to evidence on therapy. Based on study design, risk of bias, generalizability, consistency, adequate sample size and statistical</td>
<td></td>
</tr>
<tr>
<td>Author(s)</td>
<td>Grade</td>
<td>Grade description</td>
<td></td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-------</td>
<td>-----------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Guyatt et al. 2000</strong></td>
<td>E</td>
<td>Not identified. Introductory article from the Users’ Guides to the Medical Literature on Evidence Based Medicine.</td>
<td></td>
</tr>
<tr>
<td><strong>NHS Centre for Evidence Based Medicine 2001</strong></td>
<td>I</td>
<td>Oxford.</td>
<td></td>
</tr>
<tr>
<td><strong>New Zealand Guidelines Group 2000</strong></td>
<td>I</td>
<td>NZGG.</td>
<td></td>
</tr>
<tr>
<td><strong>Sackett et al. 2000</strong></td>
<td>E</td>
<td>Not identified. Book on evidence based medicine. Suggests levels of evidence based on study design.</td>
<td></td>
</tr>
<tr>
<td><strong>Harris et al. 2001</strong></td>
<td>I</td>
<td>USPSTF.</td>
<td></td>
</tr>
<tr>
<td><strong>Chestnut et al. 1999</strong></td>
<td>I</td>
<td>AHCPR.</td>
<td></td>
</tr>
<tr>
<td><strong>West et al. 1999</strong></td>
<td>I</td>
<td>AHCPR.</td>
<td></td>
</tr>
<tr>
<td><strong>McNamara et al. 1999</strong></td>
<td>I</td>
<td>AHCPR.</td>
<td></td>
</tr>
<tr>
<td><strong>Ross et al. 2000</strong></td>
<td>I</td>
<td>AHCPR.</td>
<td></td>
</tr>
<tr>
<td><strong>Levine et al. 2000</strong></td>
<td>I</td>
<td>AHCPR.</td>
<td></td>
</tr>
<tr>
<td><strong>CADTH (2012) Evidence Grading Systems Identified</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Carruthers</strong></td>
<td></td>
<td>Sackett 1989.</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Author/Source</td>
<td>Quality</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------</td>
<td>---------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1993</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gyorkos 1994</td>
<td>Not identified</td>
<td>E</td>
<td>LoE based on study validity, strength of association, precision, variability of findings between studies. Not specific to diagnosis. Full text not available.</td>
</tr>
<tr>
<td>Liddle 1996</td>
<td>SIGN – MERGE Methodology Checklists</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>The Joanna Briggs Institute 1999</td>
<td>Not identified</td>
<td>E</td>
<td>The Joanna Briggs Institute FAME Levels of Evidence. Generic system with four levels of evidence and grades of recommendation based on assessment of feasibility, appropriateness, meaningfulness, effectiveness, and economic evaluation.</td>
</tr>
<tr>
<td>NHMRC 1999</td>
<td>NHMRC</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>NHMRC 2000</td>
<td>NHMRC</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>NHMRC 2005</td>
<td>NHMRC</td>
<td>I</td>
<td>n/a</td>
</tr>
</tbody>
</table>
specific to diagnosis. Tool based on study design, study execution, number of studies, consistency, effect size, and expert opinion.

<table>
<thead>
<tr>
<th>Author</th>
<th>Publication Year</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellis 2000</td>
<td>Not identified</td>
<td>E</td>
<td>Article on clinical practice benchmarking. Presents simple classification of evidence based on study design.</td>
</tr>
<tr>
<td>Greer 2000</td>
<td>Not identified</td>
<td>E</td>
<td>Methods article about the Institute for Clinical Systems Improvement (ICSI) evidence grading system. Developed system tailored to evidence on therapy. Based on study design, risk of bias, generalizability, consistency, adequate sample size and statistical power.</td>
</tr>
<tr>
<td>Guyatt 2000</td>
<td>Not identified</td>
<td>E</td>
<td>Not a tool. Introductory article from the Users’ Guides to the Medical Literature on Evidence Based Medicine.</td>
</tr>
<tr>
<td>ACCP-Guyatt 2001</td>
<td>GRADE Adaptation by ACCP</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Eccles 2001</td>
<td>AHCPR</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Harris 2001</td>
<td>USPSTF</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>NHS Centre for Evidence Based Medicine 2001</td>
<td>Oxford</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>OCEBM 2001</td>
<td>Oxford</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>CTFPHC 2003</td>
<td>Canadian Task Force on Preventive Health Care</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>GRADE Working Group</td>
<td>GRADE</td>
<td>I</td>
<td>n/a</td>
</tr>
<tr>
<td>Year</td>
<td>Reference</td>
<td>Grade</td>
<td>Type</td>
</tr>
<tr>
<td>-----------</td>
<td>--------------------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>2004</td>
<td>Guyatt 2004</td>
<td>E</td>
<td>Not identified</td>
</tr>
<tr>
<td>2004</td>
<td>SIGN 50 2004</td>
<td>I</td>
<td>SIGN</td>
</tr>
<tr>
<td>2005</td>
<td>Soldani 2005</td>
<td>E</td>
<td>Not identified</td>
</tr>
</tbody>
</table>

1 Note: Bolded references/tool names were identified in both AHRQ and CADTH reviews.

2 Sackett 1994 tool\(^43\) is named after article by Jaeschke et al. 1994 with Sackett as the final author. The Sackett references are not described as separate tools in our report but are discussed under the AHCPR tool.
## APPENDIX 4.3: LIST OF TOOLS INCLUDED WITH REFERENCES

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. GRADE</td>
<td>44-94</td>
</tr>
<tr>
<td>2. Oxford</td>
<td>95-118</td>
</tr>
<tr>
<td>3. USPSTF</td>
<td>119-134</td>
</tr>
<tr>
<td>4. SIGN</td>
<td>13,135-146</td>
</tr>
<tr>
<td>5. National Health and Medical Research Council (NHMRC)</td>
<td>147-153</td>
</tr>
<tr>
<td>7. Agency for Health Care Policy and Research (AHCPR)</td>
<td>14,150,159</td>
</tr>
<tr>
<td>8. American Academy of Neurology (AAN)</td>
<td>15,160-166</td>
</tr>
<tr>
<td>10. American Academy of Sleep Medicine (AASM)</td>
<td>16,169-172</td>
</tr>
<tr>
<td>11. American Association of Clinical Endocrinologists (AACE)</td>
<td>17</td>
</tr>
<tr>
<td>12. American College of Radiology (ACR) Appropriateness Criteria</td>
<td>18,173,174</td>
</tr>
<tr>
<td>13. American Heart Association (AHA)</td>
<td>175-184</td>
</tr>
<tr>
<td>15. American Urological Association (AUA)</td>
<td>29,188-191</td>
</tr>
<tr>
<td>16. Canadian Diabetes Association (CDA)</td>
<td>20</td>
</tr>
<tr>
<td>17. Canadian Task Force on Preventive Health Care (CTFPHC)</td>
<td>192-195</td>
</tr>
<tr>
<td>18. Diagnostic Imaging Pathways (DIP)</td>
<td>196</td>
</tr>
<tr>
<td>19. Dutch Institute for Healthcare Improvement (Centraal BegeleidingsOrgaan; CBO)</td>
<td>21,197,198</td>
</tr>
<tr>
<td>20. Evaluation of Genomic Applications in Practice and Prevention (EGAPP)</td>
<td>30,199-205</td>
</tr>
<tr>
<td>21. European League Against Rheumatism (EULAR)</td>
<td>206-213</td>
</tr>
<tr>
<td>22. European Federation of Neurological Societies (EFNS)</td>
<td>22,214,215</td>
</tr>
<tr>
<td>23. German Federal Joint Committee (Gemeinsame Bundesausschuss; G-BA)</td>
<td>23,216-222</td>
</tr>
<tr>
<td>24. International Liaison Committee on Resuscitation (ILCOR) Consensus Document on the Science of Cardiopulmonary Resuscitation with Treatment (CoSTR)</td>
<td>223-225</td>
</tr>
<tr>
<td>25. International Scientific Society on Scoliosis Orthopaedic and Rehabilitation Treatment (SOSORT)</td>
<td>226,227</td>
</tr>
<tr>
<td>26. Italian National Guidelines System (Programma Nazionale Linee Guida; PNLG)</td>
<td>24,220-230</td>
</tr>
<tr>
<td>27. Let Evidence Guide Every New Decision (LEGEND)</td>
<td>231,232</td>
</tr>
<tr>
<td>28. New Zealand Guidelines Group (NZGG)</td>
<td>233,234</td>
</tr>
<tr>
<td>29. Oncology Nursing Society (ONS)</td>
<td>31,235</td>
</tr>
<tr>
<td>30. European Society of Cardiology (ESC)</td>
<td>236</td>
</tr>
<tr>
<td>31. Mixed Tools, Adaptations, and Other</td>
<td>34,36-39,41,42,237-351</td>
</tr>
</tbody>
</table>
## APPENDIX 4.4: EXCLUDED STUDIES WITH REASONS FOR EXCLUSION AND REFERENCES

<table>
<thead>
<tr>
<th>Reason for Exclusion</th>
<th>Reference(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Risk of Bias Tools (including Adaptations)</td>
<td>352-395</td>
</tr>
<tr>
<td>2. Translation not possible</td>
<td>396</td>
</tr>
<tr>
<td>3. Papers could not be obtained in full</td>
<td>597-599</td>
</tr>
<tr>
<td>4. No discussion or application of quality of evidence or strength of recommendation tool</td>
<td>35,600-832</td>
</tr>
<tr>
<td>5. Not pertaining to diagnostic test or diagnosing a clinical condition or disease</td>
<td>19,32,833-1010</td>
</tr>
<tr>
<td>6. Systematic review of tools</td>
<td>3-12,1011,1012</td>
</tr>
</tbody>
</table>
References


137. Carlsen KH, Anderson SD, Bjermer L, et al. Exercise-induced asthma, respiratory and allergic disorders in elite athletes: Epidemiology, mechanisms and diagnosis: Part I of the report from the Joint Task Force of the European Respiratory Society (ERS) and the European Academy of Allergy and Clinical Immunology (EAACI) in cooperation with GA 2LEN. *Allergy: European Journal of Allergy and Clinical Immunology*. April 2008;63 (4):387-403.


151. 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.


216. Brandenburg S, Kranig A, Letzel S, von Mittelstaedt G, Palfner S, Selbmann HK. Joint recommendations of AWMF and DGUV in co-operation with DGAUM and DGSMP for the


220. German Institute for Quality and Efficiency in Healthcare (IQWiG). PET, PET/CT for malignant melanoma (Nr. 86). Cologne, Germany2011.

221. German Institute for Quality and Efficiency in Healthcare (IQWiG). PET, PET/CT for head and neck tumors (Nr. 82). Cologne, Germany2011.


253. Deutsche Gesellschaft für Bipolare Störungen (DGBS), Deutsche Gesellschaft für Psychiatrie Psychotherapie und Nervenheilkunde (DGPPN). Leitlinie zur Diagnostik und Therapie Bipolarer


269. German Institute for Quality and Efficiency in Healthcare (IQWiG). *Standard for diagnosing bronchial asthma of children in the age of 2 to 5 (Nr. 29).* Cologne, Germany 2008.

270. German society for digestive and metabolic diseases (DGVS). *Diagnosis and therapy of adenocarcinoma of the stomach and gastroesophageal passage.* Berlin, Germany 2011.


355. Al-Sukhni E, Milot L, Fruitman M, et al. Diagnostic Accuracy of MRI for Assessment of T Category, Lymph Node Metastases, and Circumferential Resection Margin Involvement in Patients with


582. Yin JY, Ho KM. Use of plethysmographic variability index derived from the Massimo([REGISTERED]) pulse oximeter to predict fluid or preload responsiveness: a systematic review and meta-analysis. *Anaesthesia*. Jul 2012;67(7):777-783.


248


728. Monegon FP, Pal RS, Wersching H, Kwong RY, Jerosch-Herold M, Coelho-Filho OR. Reduced predictive value of scar imaging by cardiac MRI to predict death or myocardial infarction in


806. Treglia G, Stefanelli A, Cason E, Cocciliillo F, Di Giuda D, Giordano A. Diagnostic performance of iodine-123-metaiodobenzylguanidine scintigraphy in differential diagnosis between Parkinson’s...


846. Arlt SP, Heuwieser W. Training students to appraise the quality of scientific literature. Journal of Veterinary Medical Education. 01 Jan 2011;38 (2):135-140.


959. McLaughlin SL, Canfield SE, Fesperman SF, Dahm P. Evaluating the evidence: A critical appraisal of the methodological quality of systematic reviews published in the urological literature (1998-


CHAPTER 5: ASSESSING INTERNATIONAL GUIDELINES
Decision-making about healthcare related tests and diagnostic strategies: Assessing international guidelines.

Reem A. Mustafa, MD, MPH¹,², Wojtek Wiercioch, MSc¹, Ingrid Arevalo-Rodriguez, MSc³, Adrienne Cheung, B.H.Sc⁴, Barbara Prediger, B.Sc.⁵, Liudmila Ivanova, MD, MPH¹, Matthew Ventresca, B. Sc.¹, Jan Brozek, MD¹,⁷, Nancy Santesso, RD, MLIS¹, Patrick Bossuyt, MSc, PhD⁸, Amit X. Garg, MD, PhD¹,⁹, Nancy Lloyd, MSc, B.Sc.¹, Monika Lelgemann, MD, MSc¹⁰, Diedrich Bühler, MD¹¹, Holger Schünemann, MD, MSc, PhD¹,⁷.

(1) Department of Clinical epidemiology and Biostatistics, McMaster University, Hamilton, Canada

(2) Department of Internal Medicine and Nephrology, University of Missouri-Kansas City, Kansas City, USA

(3) Fundación Universitaria de Ciencias de la Salud, Hospital San José & Hospital Infantil de San José, Colombia

(4) Faculty of Medicine, University of British Colombia, Vancouver, Canada

(5) Center for Medical Biometry and Medical Informatics, University of Freiburg, Germany

(6) Faculty of Applied Health Sciences, Brock University, St. Catharines, ON

(7) Department of Medicine, McMaster University, Hamilton, Canada

(8) Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherland

(9) Department of Medicine, Western University, London, Canada

(10) Medizinischer Dienst des Spitzenverbandes Bund der Kranken-kassen e.V. (MDS) Theodor Althoff-Str. 47 45133 Essen, Germany

(11) Abteilung Medizin. GKV— Reinhardtstraße 28 10117 Berlin, Germany
ABSTRACT

Objective
To describe and compare current practices in developing guidelines about the use of health care related tests and diagnostic strategies (HCTDS)

Method
We obtained a random sample of 37 public health and clinical practice guidelines about HCTDS published in any language.

Results
Detailed descriptions of the systems used to assess the quality of evidence and develop recommendations were challenging to find within guidelines. We observed much variability among and within organizations with respect to how they develop recommendations about using tests. 24% of the guidelines making recommendations about using tests did not consider health benefits and harms but based the decisions solely on test accuracy. We did not identify any guidelines that described the main potential care pathways for a healthcare problem including use of tests. In addition, we did not identify any guidelines that systematically assessed, described and referenced the evidence that linked diagnostic test accuracy and patient-important outcomes.

Conclusion
There is considerable variability among the processes used and factors considered in developing recommendations about the use of tests that led, in some instances, to disagreement in recommendations about the use of the same test for the same condition.
WHAT IS NEW?

- It was challenging to locate information about methods used to appraise evidence and develop recommendations pertaining to HCTDS due to poor and inconsistent reporting.
- We identified considerable variability among the processes used and factors considered in developing recommendations about the use of tests in health care.
- We did not identify any guidelines that systematically assessed, described and referenced the evidence that linked test accuracy and patient-important outcomes.
- Disagreement about the use of the same test for the same condition may be the result of variability in the process of developing recommendations, evidence considered about the test, or the care pathways and the treatment effects that are considered following a specific test result.
BACKGROUND

In the three previous articles in this series addressing diagnostic decision-making in health care, we provided an overview of known roles and applications of tests, different study designs for evaluating tests and the main challenges encountered in conducting diagnostic research and applying it in health care decision-making. We also described our research on presenting information for test accuracy and summarized all currently available instruments to assess the quality of evidence and strength of recommendations about the use of health care-related tests and diagnostic strategies (HCTDS). In this fourth of a series of eight articles we describe how guidelines about HCTDS are developed. Specifically we address the following issues about HCTDS guidelines: How do international organizations develop guidelines about HCTDS? How are test accuracy results considered in making recommendations? How do different guideline development processes for the same or similar test compare?
METHODS

Data sources

We obtained a sample of 37\textsuperscript{1-28} public health and clinical practice guidelines using purposeful random sampling from various sources.

a. Electronic sources

We electronically searched: 1. The National Guideline Clearinghouse (NGC)(29), 2. The TRIP database\textsuperscript{30}, 3. The Canadian Medical Association (CMA) Infobase\textsuperscript{31}, and 4. The Guideline International Network (GIN) library\textsuperscript{32}. The remainder of our sample came from directly searching the websites of societies and groups known to produce guidelines about using tests. These groups included: The American College of Radiology (ACR)\textsuperscript{33}, the National Academy of Clinical Biochemistry (NACB) for the Academy of American Association for Clinical Chemistry (AACC)\textsuperscript{34}, the National Institute for Health and Clinical Excellence (NICE)\textsuperscript{35}, the Scottish Intercollegiate Guideline Network (SIGN)\textsuperscript{36}, the U.S. Preventive Services Task Force (USPSTF)\textsuperscript{37}, the Australian National Health and Medical Research Council (NHMRC)\textsuperscript{38}, Cancer Care Ontario (CCO)\textsuperscript{39}, the American College of Chest Physicians (ACCP)\textsuperscript{40}, the Association of the Scientific Medical Societies in Germany (AWMF)\textsuperscript{41}, and the World Health Organization (WHO)\textsuperscript{42}.

In addition, we searched the National Institute of Health Research Economic Evaluation Database, the National Institute of Health Research Health Technology Assessment Database, and the National Institute of Environmental Health Systems\textsuperscript{43}, however we did not identify guidelines about using tests from these sources.

b. Other sources

We contacted experts in guideline methodology and development from Australia, Canada, Chile, China, Colombia, the Netherlands, Norway, South Africa and the United States and
those from professional societies, associations or organizations known to follow rigorous methods in producing guidelines.

**Inclusion and exclusion Criteria**

**Inclusion criteria**

1. Guidelines which provided recommendations, guidance, or practice points related to the use of HCTDS and their applications including screening or surveillance, risk assessment and classification, diagnosis (ruling in and/or ruling out specific conditions), treatment triage, treatment monitoring, grading and staging, or determining prognosis.

2. In Germany, there is a classification to differentiate between methodological quality of guidelines, which includes S1, S2 and S3 levels. We purposefully sampled five S3 guidelines (representing the highest level of methodological rigour in guideline development).

3. We included all languages.

**Exclusion criteria**

1. Guidelines developed using an obsolete process that was explicitly replaced with an updated one used by the same group in subsequent guidelines.

2. Guidelines with focus not on supporting decisions related to the use of tests.

3. Guidelines published before 2005. We aimed to examine guidelines that followed the most up to date methods in guideline development.

**Data abstraction and reporting**

We developed a data abstraction form using the criteria summarized in the AGREE II instrument for assessing rigor of development and quality of reporting of guidelines.
the WHO guide for guidelines. We used an iterative process to modify the form based on pilot testing and feedback from data abstractors. The first part of the abstraction form consisted of fields to describe the practice guideline, including the scope of the guideline (e.g. diagnosis, screening, staging, monitoring, treatment) and the diagnostic tests covered in the guideline. The second part of the form included fields to assess the guideline development methodology, such as whether the health questions, target population, and target audience were specifically described, and whether the guideline development group included all relevant professional groups and stakeholders. The third part of the form was used to obtain information about the evidence search and synthesis in the guideline, such as whether systematic methods were used to search for evidence, how the evidence was analyzed and synthesized, and whether a system or tool was used to evaluate the quality of the evidence. The fourth part of the form was used to assess whether links were made between diagnostic test accuracy and patient outcomes in the guideline. The fifth and sixth parts of the form were used to record how recommendations were formulated and presented in the guideline, such as whether the methods used to formulate recommendations were clearly described, whether there was an explicit link between the recommendations and the supporting evidence, and whether the recommendations were clearly presented. The last section of the form was used to capture abstractors’ comments. The final data abstraction form is provided in Appendix 1.

We sampled at least two guidelines from each of the previously described sources, and we sampled guidelines about the same test from different groups to compare processes. All relevant data were abstracted independently and in duplicate. Discrepancies were then resolved by discussion. Pairs of abstractors were mixed to avoid any systematic biases in abstraction. When uncertainty arose, we contacted representatives from organizations to confirm our findings and to clarify any confusion. We summed the abstracted data from all
sampled guidelines to present the results according to the domains of the AGREE II tool (Table 1). In the abstraction form we used 5 points likert like scale with specific instructions per question. For answers “clearly yes” or “probably yes” on the abstraction form we reported them as “yes” in table 1. For answers “clearly no” on abstraction form we reported them as “No” in table 1. The other response options were reported under “additional results” with clear explanation for what do the responses mean for the specific domain.
RESULTS

Overview of Identified Guidelines
While we encountered minimal challenges with identifying guidelines themselves, detailed descriptions of the systems used to assess the quality of evidence and develop recommendations were harder to find within guidelines. We identified a diverse sample of guidelines from Australia, Europe, North America, and South America. These guidelines were generally applicable to the countries in which they were developed with the exception of WHO guidelines with worldwide applicability and certain geographic considerations, particularly around resource implications. Our sample of guidelines was also variable in clinical areas, diseases and tests under evaluation.

With regards to possible applications of tests, we identified guidelines about using tests for diagnosis, screening, staging, determining prognosis, risk assessment and monitoring treatment. A summary of the guideline characteristics, the tests and applications is provided in Appendix 2.

Guideline development methodology
We observed much variability among and within organizations with respect to how they develop recommendations about using tests. Table 1 summarizes guideline development methodologies. The majority of guidelines (87%) specified the population to whom the guideline applies. However, only half sought the views and preferences of the target population while developing recommendations.

None of the guidelines described using a different process for developing recommendations about tests and about treatments.
24% of the guidelines making recommendations about using tests did not consider final health benefits and harms but based the decisions solely on test accuracy. 54% of the guidelines summarized the evidence either in tables or narratively while 11% failed to reference the evidence used to support the recommendations. In 84% of the guidelines, recommendations were separated from the main text, in bold, in larger font, summarized in a table, presented at the beginning of the guideline, or some combination thereof. Although 76% of guidelines noted editorial independence, 24% did not report the guideline development group’s financial or intellectual competing interests.

Assessing the quality of evidence (QoE) and strength of recommendations (SoR)

Seventy eight percent of the guidelines used a formal system for rating the QoE. However, only 30% clearly described the strengths and limitations of the body of evidence. Quality of evidence rating scales varied considerably with some groups using letters, categories, or a numeric system. The most common rating scheme was a 4-point scale, though many different versions were identified.

Additionally, 68% of the guidelines used some system to determine the SoR. Nonetheless, only six guidelines\(^2\) provided guidance for proper interpretation of the SoR and its implications for practice. There was also inconsistency in terminology among guidelines developed by different organizations, the same organization or even within individual guidelines. Hence, we were unable to determine whether and what was the difference in meaning in wording specific recommendations, e.g. phrases “is recommended”, “should be used”, or “is indicated”. Another inconsistency involved variation among groups about when to, or not to, develop recommendations. Although some groups refrained from providing a recommendation when they perceived the evidence to be lacking or
insufficient, others chose to recommend against using a test when there was insufficient evidence to assess benefit.

**Linking evidence about test accuracy to patient-important outcomes**

We did not identify any guidelines that described the main potential care pathways for a health care problem including use of tests. In addition, we did not identify any guidelines that systematically assessed, described and referenced the evidence that linked test accuracy and patient-important outcomes.

In guidelines addressing more than one application for a given test (e.g. diagnosis and staging), it was not clear whether the same or different evidence was used to inform recommendations about different applications. To assess linking test accuracy results to patient-important outcomes, we looked thoroughly for any consideration of this link in any part of the guideline document. For example, if a guideline stated that a false positive result could lead to patient anxiety or unnecessary further diagnostic testing, we took this as evidence that this link was indeed considered. Although none of the groups clearly described the links between all four possible outcomes of a test (true and false positives and negatives) and patient-important outcomes, 35% described the link between a true positive test result and patient outcomes. 27% and 19% of guidelines described the link between a false positive and false negative test results respectively and patient outcomes, while only 8% explicitly considered implications of true negative results when making recommendations (Table 2).

**Guidelines making recommendations about the same test for the same condition: comparison of guideline methods and recommendations**
In our sample of 37 guidelines we identified two topics for which more than one group produced a separate guideline. The first example involves diagnosing, staging and treating lung cancer using various tests, including chest x-rays and CT scans\textsuperscript{8,26}. While both groups followed similar transparent methods including disclosing COI, considering patients’ values and preferences and using systematic methods to search for and summarize evidence, they reached different conclusions. Both guidelines described parts of the link between test accuracy and patient-important outcomes. However, they did not agree on treatment consequences; one guideline explained “an early diagnosis of lung cancer can lead to a better chance for treatment” while the other described the consequences as “a diagnosed lung cancer is often inoperable”. However, neither referenced the evidence supporting the treatment consequences.

The second example of guidelines addressing the same topic focused on diagnosing and staging renal cell carcinoma using multiple tests including CT scans and chest x-rays, among others \textsuperscript{24,51}. One of the groups did not provide the search strategy and did not summarize the evidence informing each recommendation separately. In addition, the two guidelines used different methods and tools to assess the QoE and SoR but they concluded with similar recommendations. More details about these guidelines and comparing them is provided in Appendix 3.
DISCUSSION

In this article, based on a systematic approach, we identified 37 international guidelines developed by multiple organizations that addressed a variety of tests with different applications for a wide range of conditions. We did not identify any guidelines that described the main potential care pathways for a healthcare problem including use of tests. When direct evidence about the effect of a test on patient-important outcome was lacking, we did not find any guidelines describing using a different process for developing recommendations about tests from the process used for developing recommendations about treatments. Overall, guideline developers did not provide evidence to demonstrate that the links between test accuracy and patient-important outcomes were thoroughly and systematically considered as they formulated their recommendations.

Our review has multiple strengths. We used a rigorous iterative method to identify up-to-date guidelines pertaining to HCTDS. We searched websites of international societies and guideline clearinghouses and consulted with experts from around the world. We also included guidelines published in any language and translated them. In this setting, it is unlikely that we missed any major guidelines in our sample. Our method allowed us to compare recommendations addressing the same test for the same condition to explore potential reasons for variation in recommendations among different groups. Additionally, we were able to explore new developments for any of the tests’ applications and roles as we did not restrict our search to guidelines about using tests for diagnosis only. Tests are used for many different applications: screening or surveillance, risk assessment and classification, diagnosis (ruling in), ruling out diagnosis, treatment triage, treatment monitoring, staging and determining prognosis. Tests are also used in different roles in the
Care pathways: triage, add-on, replacement, parallel testing and replacement of a reference test (ref paper 1).

This review has a few limitations. It was, at times, not clear whether certain processes were in fact not followed or not reported. To address this limitation, we contacted representatives from organizations to confirm our findings and to clarify any confusion when possible. In most instances, partial updates of methodology manuals, where some sections were not updated and old information was still used, led to discrepancies within the same manual. It was more difficult to clarify instances when specific guideline panels did not follow the methods described in the manual. We believe that discrepancies between the manuals and the actual processes followed in specific panels relate likely to panelists' and chair's unfamiliarity with the original method and process. We also believe that discrepancies were more likely encountered when there was no methodologist involved in the process of developing the guidelines. While one of the primary goals of this work was to determine the various methods of developing recommendations about tests used by different organizations, it became a challenge to draw conclusions about that because of the reporting limitations we encountered. Most guidelines did not clearly describe the method for formulating the recommendations about using tests. Even when guideline groups did report their methods, oftentimes information was scattered, vague and inconsistent. Additionally, our findings were based on abstractors' judgments, which may have varied from one abstractor pair to another. To address this limitation, we scheduled frequent meetings for the abstractors to discuss their judgments. In addition, we mixed abstractor pairs to minimize potential biases by a specific pair of abstractors. We also standardized the abstraction form with explanation for each question and its responses. Lastly, the main author was a common third abstractor for all guidelines to minimize inter-rater variability.
It was challenging to determine what methods were followed, when only a generic statement such as “benefits and harms were considered” was given. Although most guideline developers stated that they considered benefits and harms, there were usually no details about how this was actually done. This is particularly important in guidelines about using tests, as assessing benefits and harms from using a test requires laying out the potential care pathways and determining the downstream consequences related to particular test result in the care pathway as well as the complications of testing itself. Including only vague statements about benefits and harms rather than a complete assessment and description of considerations in the care pathway is less useful to guideline users.

Another challenge is inadequate reporting of strengths and limitations of the overall body of evidence supporting recommendations. This limitation makes it difficult to determine the rationale for and appropriateness of a given recommendation. It was difficult to find all the information about methods used in developing guidelines given inadequate descriptions of methods in manuals, poor reporting in guideline documents, and having information scattered among multiple documents without clear referencing. In some instances we were able to identify multiple versions of the same manual or methods paper on websites without clear guidance which version is the most recent and how it differed from the previous versions. These discrepancies and disconnect between methods described in manuals and the actual methods applied in the guidelines precluded proper assessment and description of the methodology. While groups may provide separate documents outlining the general methods for development of practice guidelines in their organization it is often difficult to ascertain whether these methods were indeed followed during the development of specific guidelines. These generic descriptions of methods are even more problematic in the context of recommendations about using tests. These
methods were likely originally intended for producing treatment guidelines and do not account for major differences in factors being considered and approaches to making recommendations about treatments and tests.

Presumed consequences after positive test results (both true and false) were more commonly considered when making recommendations about testing. Consequences of false negative results were rarely considered. This is likely due to the difficulty in assessing what actually happens to patients who are misclassified as not having the condition when they actually have it. It also highlights the need for additional information about natural progression of the condition, which may not be identified with the same body of evidence used for assessing the effects of testing.

There were discrepancies among recommendations about using the same test for the same condition but developed by different organizations. In the guidelines addressing lung cancer, the discrepancy was the result of assuming different effectiveness of the following treatment. This highlights the importance of transparency in describing the presumed downstream consequences of particular test results in the care pathway. It is critical for information about these consequences to be based on a systematic review of the best available evidence.

In this fourth article in a series of eight, we identified and analyzed 37 guidelines making recommendations about the use of tests. In general methods of developing recommendations about HCTDS were inadequately and inconsistently reported. There is considerable variability among the processes being used and factors being considered in developing recommendations about tests that is the most likely explanation for disagreement among recommendations about the use of the same test for the same condition but made by different guideline groups.
Acknowledgment

This work was partially funded by the German Insurance Fund agency as part of a larger project about decision-making for HCTDS. The views presented here are those of the authors and should not be attributed to the funding agency or its staff.
References


30. http://www.g-i-n.net.


34. http://www.wihn.ac.uk.


42. The Association of the Scientific Medical Societies in Germany (Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften e.V., AWMF).


50. German association of urology (DGU). Interdisciplinary guideline with S3 quality for screening, diagnosis and therapy of different stages of prostate carcinoma. Berlin, Germany; 2011.

### TABLE 5.1. SUMMARY OF RESULTS BASED ON AGREE II DOMAINS

<table>
<thead>
<tr>
<th>Items</th>
<th>Yes</th>
<th>No</th>
<th>Additional results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain: Scope and Purpose</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The health question(s) covered by the guideline is (are) specifically described</td>
<td>24 (65%)</td>
<td>8 (22%)</td>
<td>Incomplete description: 5 (13%)</td>
</tr>
<tr>
<td>The population to whom the guideline is meant to apply is specifically described</td>
<td>32 (87%)</td>
<td>2 (5%)</td>
<td>Vague description: 3 (8%)</td>
</tr>
<tr>
<td><strong>Domain: Stakeholder Involvement</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The guideline development group includes individuals from all the relevant professional groups/specialities</td>
<td>27 (73%)</td>
<td>4 (11%)</td>
<td>Insufficient information: 6 (16%)</td>
</tr>
<tr>
<td>The views and preferences of the target population have been sought</td>
<td>18 (49%)</td>
<td>16 (43%)</td>
<td>Insufficient information: 3 (8%)</td>
</tr>
<tr>
<td>The target users of the guideline are clearly defined</td>
<td>33 (89%)</td>
<td>4 (11%)</td>
<td></td>
</tr>
<tr>
<td><strong>Domain: Rigor of Development</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Systematic methods were used to search for evidence</td>
<td>Specific description: 23 (62%)</td>
<td>1 (3%)</td>
<td>No mention of search method: 6 (16%)</td>
</tr>
<tr>
<td></td>
<td>Generic description: 7 (19%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The criteria for selecting the evidence are clearly described</td>
<td>16 (43%)</td>
<td>21 (57%)</td>
<td></td>
</tr>
<tr>
<td>Case-control studies are included</td>
<td>21 (57%)</td>
<td>1 (3%)</td>
<td>Insufficient information: 15 (41%)</td>
</tr>
<tr>
<td>The methods for formulating the recommendations are clearly described</td>
<td>19 (51%)</td>
<td>7 (19%)</td>
<td>Vague or incomplete description provided: 11 (30%)</td>
</tr>
<tr>
<td>The health benefits, side effects, and risks have been considered in formulating the recommendations</td>
<td>Clearly described and/or quantified: 15 (40%)</td>
<td>Clearly not considered: 9 (24%)</td>
<td>Unclear: 5 (14%)</td>
</tr>
<tr>
<td>Domain: Clarity of Presentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>The different options for diagnosing the condition or health issue are considered</td>
<td>Comprehensive list of options: 25 (68%) Few options: 10 (27%)</td>
<td>2 (5%)</td>
<td></td>
</tr>
<tr>
<td>The different options for diagnosing the condition or health issue are clearly presented**</td>
<td>Clearly listed (e.g. in table): 17 (49%) Clearly presented in text: 9 (26%)</td>
<td>6 (17%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Scattered in different parts of text: 3 (8%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Key recommendations are easily identifiable</td>
<td>31 (84%)</td>
<td>6 (16%)</td>
<td></td>
</tr>
<tr>
<td>The evidence used in developing recommendations is well summarized</td>
<td>24 (65%) 20 (54%) in tables 4 (11%) narratively</td>
<td>8 (22%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Evidence referenced but not summarized: 5 (13%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain: Applicability</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>The potential resource implications of applying the recommendations have been considered</td>
<td>16 (43%) 6 with economic analysis 10 without economic analysis</td>
<td>13 (35%) 8 do not mention resources 5 explicitly do not consider resources</td>
</tr>
<tr>
<td></td>
<td>Mention but do not consider resources in recommendations: 8 (22%)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Domain: Editorial Independence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Competing interests of guideline development group members have been recorded</td>
<td>28 (76%) Financial and Intellectual: 16</td>
<td>9 (24%)</td>
</tr>
</tbody>
</table>
### Assessing the Quality of the Evidence and the Strength of the Recommendations

<table>
<thead>
<tr>
<th>Description</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>A system for evaluating the quality of evidence is used</td>
<td>29 (78%)</td>
<td>8 (22%)</td>
</tr>
<tr>
<td>The strengths and limitations of the body of evidence are clearly described</td>
<td>11 (30%)</td>
<td>26 (70%)</td>
</tr>
<tr>
<td>A system for evaluating the strength of recommendations is used</td>
<td>25 (68%)</td>
<td>12 (32%)</td>
</tr>
</tbody>
</table>

* Out of 37 guidelines

** Out of 35 guidelines (excludes two guidelines that did not consider different options)
### TABLE 5.2. LINKING DIAGNOSTIC TEST ACCURACY TO PATIENT-IMPORTANT OUTCOMES

<table>
<thead>
<tr>
<th>Domain</th>
<th>Yes*</th>
<th>Likely yes**</th>
<th>No***</th>
<th>Examples</th>
</tr>
</thead>
</table>
| The link between false positive (FP) and patient outcomes is considered | 9    | 1            | 27    | “Women should be informed that MRI does not reliably distinguish benign from malignant findings and that they may be at risk of additional investigations or unnecessary surgery, as a result of the MRI findings (7)”.  
“In patients who did not have additional malignancy on histology (false positive detection), a change from wide local excision (WLE) to mastectomy of 1.1% (95% CI, 0.3-3.6%) and from WLE to more extensive surgery of 5.5% (95% CI, 3.1-9.5%) was made”.  
“Patients who are misdiagnosed with DVT will be prescribed unnecessary anticoagulants and some will suffer major bleeding as a result (49)”. |
| The link between false negative (FN) and patient outcomes is considered | 6    | 1            | 30    | “A negative result cannot be conclusively interpreted as meaning the person does not carry a mutation known to cause familial hypercholesterolaemia. Rather, it could simply mean that the person does not have one of the mutations tested for in that specific test. This uncertainty would affect patient management and would not permit people with a true negative result to discontinue treatment for a condition they do not have NICE (15)” |
| The link between true positive (TP) and patient outcomes is considered | 11   | 2            | 24    | “This resulted in a change in surgical management for patients with multifocal/multicentric histologically-proven cancer from wide local excision (WLE) to mastectomy of 8.1% (95% CI, 5.9-11.3%) and from WLE to more extensive surgery (i.e. wider/additional excision or mastectomy) of 11.3% (95% CI, 6.8-18.3%)”.  
“MRI was shown to detect the tumour in more than two thirds of patients and provided the possibility of breast conserving surgery in one third of patients” |
| The link between true negative (TN) and patient outcomes is considered | 2    | 1            | 34    | “a negative result may be useful to exclude recurrence and prompt a search for alternative diagnoses”  
“TN result would mean that patients are spared from unnecessary treatment and they would have the benefit of reassurance and an alternative diagnosis” |

*Yes: Guidelines described, presented, or quantified the consequences of a FP, FN, TP or TN test result on patient outcomes
### APPENDIX 5.1: GUIDELINES DATA ABSTRACTION FORM

#### Section 1: Guideline Identification and Description

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guideline Source</td>
<td>Write name of source</td>
<td>Where was the guideline retrieved from (e.g. guideline clearinghouse or the guideline developer website)?</td>
</tr>
<tr>
<td>Guideline Developer or Organization</td>
<td>Write name of organization</td>
<td>-</td>
</tr>
<tr>
<td>Year</td>
<td>Write year of most recent update</td>
<td>-</td>
</tr>
<tr>
<td>Is guideline an update of a prior version?</td>
<td>Yes or No; Provide date of previous version</td>
<td>-</td>
</tr>
<tr>
<td>Was an expiration date given?</td>
<td>Yes or No; Provide date of next update if stated</td>
<td>Does the guideline provide an expiration date or a planned date for update? Does the organization have an explicit policy to update guidelines every few years?</td>
</tr>
<tr>
<td>Guideline Title</td>
<td>Write title</td>
<td>-</td>
</tr>
<tr>
<td>Guideline to be used to aid in:</td>
<td>Screening for a disease; Diagnosis of a disease; Staging of a disease; Monitoring of a disease</td>
<td>-</td>
</tr>
<tr>
<td>Condition Investigated</td>
<td>List all conditions guideline covers</td>
<td>-</td>
</tr>
<tr>
<td>Diagnostic Tests used</td>
<td>List all diagnostic tests guideline covers</td>
<td>-</td>
</tr>
<tr>
<td>Supplementary materials available online or in separate document?</td>
<td>Yes or No; If Yes, list all documents used and provide website links</td>
<td>Additional documents such as a methodology handbook, evidence synthesis, or conflict of interest disclosure are often provided separately from the main guideline document and needed to be identified.</td>
</tr>
<tr>
<td>Is guideline an adaptation of other guidelines?</td>
<td>Yes or No</td>
<td>Some guidelines are developed based only on existing guidelines and not a separate review of the evidence, and form recommendations based on adaptation of these existing guidelines.</td>
</tr>
<tr>
<td>To what country does the guideline apply?</td>
<td>Write name of country</td>
<td>-</td>
</tr>
<tr>
<td>Source of funding</td>
<td>Write name of source</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Section 2: Guideline Methodology

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The health question(s) covered by the guideline is (are) specifically described</td>
<td>1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree;</td>
<td>Strongly Agree – A health question statement was provided with all PICO components Agree – A health question statement was provided but was missing one or more PICO components Neither Agree nor Disagree – A health question statement was not provided.</td>
</tr>
</tbody>
</table>
and follows PICO. | 4-Disagree; 5-Strongly Disagree | PICO components; or, no health question statement but all PICO components described in introduction/preamble  
Disagree – No health question statement but some PICO components described  
Strongly Disagree – No health question statement or mention of PICO components anywhere in guideline

The population (patients, 
public, children, etc.) to 
whom the guideline is 
meant to apply is 
specifically described. | 1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree; 4-Disagree; 5-Strongly Disagree | Strongly Agree – Population is clearly described in guideline introduction/preamble  
Agree – Population is described only within the recommendation statements  
Disagree – Population is not described but there is some discussion about different populations (e.g. paragraph about children) in the guideline  
Strongly Disagree – No mention of population

What is the target audience of the guideline? | Free text | 

Is an interpretation aid provided? | Yes or No | An interpretation aid is intended to help users determine what a recommendation means and what to do with a recommendation. This is different from the strength of recommendation rating scheme.

The guideline development group includes individuals from all the relevant professional groups. | Yes; Likely Yes; Likely No; No; Insufficient Information | Guideline development groups should include members from different clinical specialties, different health care providers, technicians, etc. If there was some uncertainty about group membership, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as the guideline developer's policy statement or methodology handbook. If group membership was not provided or described there was insufficient information to answer this question.

List all the stakeholders that were included in this guideline development. | Free text or Insufficient information | Stakeholders include clinicians, professional groups, consumer groups, policy makers, etc. If stakeholders involved in guideline development or consultation were not described, there was insufficient information to answer this question.

Does the guideline development group disclose conflicts of interest? | Yes or No  
If Yes, note the type of conflict of interest disclosure: financial and/or intellectual | It was not sufficient for the organization to state that they have a conflict of interest policy and that members disclose conflict of interest. In order to answer ‘Yes’, disclosure had to be presented in the specific guideline or a link was provided to the disclosure.

The views and preferences of the target population (patients, public, etc.) have been sought. | Yes; Likely Yes; Likely No; No | To answer ‘Yes’, patients/public must have been members of the guideline development group or formal consultation was sought from patients or the public during the guideline development process. If there was uncertainty about patient/public involvement, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as the guideline developer's policy statement or methodology handbook.

### Section 3: Evidence Search and Synthesis

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Systematic methods were used to search for evidence.</td>
<td>Yes; Likely Yes; Likely No; No</td>
<td>Did the guideline document state that systematic methods were used or provide a search strategy? If there was uncertainty, 'likely yes' or 'likely no'</td>
</tr>
</tbody>
</table>
The criteria for selecting the evidence are clearly described (e.g. search strategy, study design, inclusion and exclusion criteria, and outcomes).

<table>
<thead>
<tr>
<th>Yes or No</th>
<th>Were the search strategy, inclusion and exclusion criteria, and outcomes of interest described in the guideline? In other words, would you be able to replicate the search and come up with the same evidence?</th>
</tr>
</thead>
</table>

When was the search conducted?

<table>
<thead>
<tr>
<th>Provide dates the search covered</th>
<th>-</th>
</tr>
</thead>
</table>

Did they include case-control studies?

<table>
<thead>
<tr>
<th>Yes; No; Insufficient Information</th>
<th>This was determined based on the search strategy, inclusion/exclusion criteria, and final evidence included. If not available, answered 'insufficient information'.</th>
</tr>
</thead>
</table>

Methods used to analyze/synthesize evidence

<table>
<thead>
<tr>
<th>Free text</th>
<th>After completing search and selecting evidence, what did the guideline development group do with it? Was a meta-analysis conducted, was evidence summarized narratively, was it summarized in evidence tables?</th>
</tr>
</thead>
</table>

The evidence used in developing the guidelines is well summarized.

<table>
<thead>
<tr>
<th>1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree; 4-Disagree; 5-Strongly Disagree</th>
<th>Strongly Agree – There was a detailed summary of all evidence in a narrative format or in evidence tables, or both. Agree – There was a summary of the evidence but not fully comprehensive. Disagree – Only some parts of the evidence were summarized. Strongly Disagree – There was no summary of the evidence.</th>
</tr>
</thead>
</table>

A system for evaluating the quality of evidence was used.

<table>
<thead>
<tr>
<th>Yes or No</th>
<th>-</th>
</tr>
</thead>
</table>

If Yes, please describe the system.

<table>
<thead>
<tr>
<th>Free text</th>
<th>Provide an overview about the system for evaluating quality of evidence used. e.g. System used categories 1 (low quality) to 4 (high quality) to rate evidence</th>
</tr>
</thead>
</table>

The strengths and limitations of the body of evidence are clearly described.

<table>
<thead>
<tr>
<th>Yes or No</th>
<th>Did the guideline authors discuss specific strengths and limitations of the evidence such as the availability of many high quality studies or limitations in study design, such as studies using surrogate outcomes, lacking the gold standard, etc.</th>
</tr>
</thead>
</table>

### Section 4: Link to Patient Outcomes

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The link between false positive (DTA) and patient outcomes is considered.</td>
<td>Yes; Likely Yes; Likely No; No</td>
<td>To answer 'Yes', the guideline must describe, present, or quantify the consequences of a false positive test result on patient outcomes (e.g. unnecessary treatment, patient anxiety) to demonstrate that this link was considered in the guideline development process.</td>
</tr>
</tbody>
</table>
The link between false negatives and patient outcomes is considered

| Yes; Likely Yes; Likely No; No |

To answer 'Yes', the guideline must describe, present, or quantify the consequences of a false negative test result on patient outcomes (e.g., disease progression) to demonstrate that this link was considered in the guideline development process. If there was uncertainty about consideration of the link in the guideline development, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as some discussion of these concepts in the guideline document or guideline methodology handbook.

The link between true positive and patient outcomes is considered

| Yes; Likely Yes; Likely No; No |

To answer 'Yes', the guideline must describe, present, or quantify the consequences of a false positive test result on patient outcomes (e.g., timely treatment) to demonstrate that this link was considered in the guideline development process. If there was uncertainty about consideration of the link in the guideline development, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as some discussion of these concepts in the guideline document or guideline methodology handbook.

The link between true negatives and patient outcomes is considered

| Yes; Likely Yes; Likely No; No |

To answer 'Yes', the guideline must describe, present, or quantify the consequences of a false positive test result on patient outcomes (e.g., no wasted resources, patient assurance) to demonstrate that this link was considered in the guideline development process. If there was uncertainty about consideration of the link in the guideline development, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as some discussion of these concepts in the guideline document or guideline methodology handbook.

### Section 5: Formulation of Guideline Recommendations

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>The methods for formulating the recommendations are clearly described.</td>
<td>1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree; 4-Disagree; 5-Strongly Disagree</td>
<td>This information will often be provided in an external document, such as a methodology handbook. The following was considered for the answer options: Strongly Agree – All steps of the development process described in detail Agree – Methods are described but lack some detail Disagree – Only some mention of methods or very general description Strongly Disagree – Methods not described at all</td>
</tr>
<tr>
<td>If Yes, describe the method used to formulate recommendations</td>
<td>Free text</td>
<td>Provide summary of methods used by the guideline development group (e.g., commission evidence review, consensus meetings to review evidence, development of draft recommendations, stakeholder consultation, publication)</td>
</tr>
<tr>
<td>Is the guideline peer-reviewed?</td>
<td>Yes or No</td>
<td>Peer-review consists of external review by experts not involved in the guideline development. May be described in external document, such as a methodology handbook.</td>
</tr>
<tr>
<td>The health benefits, side effects, and risks have been considered in formulating the recommendations.</td>
<td>Yes; Likely Yes; Likely No; No;</td>
<td>To answer 'Yes', the guideline must describe, present, or quantify benefits and risks. If there was uncertainty about consideration of the health benefits and risks, 'likely yes' or 'likely no' were used as answer options in consideration of other information such as the guideline developer's policy statement or</td>
</tr>
<tr>
<td>Question</td>
<td>Answer Options</td>
<td>Answer Option Explanation</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>There is an explicit link between the recommendations and the supporting evidence.</td>
<td>Yes or No</td>
<td>To answer 'Yes', the evidence used must be cited with the recommendation statement or described/presented with, or prior to, the recommendation statement in the guideline document. It was not sufficient to only provide the strength of evidence along with the recommendations; it was necessary to know which studies/body of evidence were used to form the specific recommendations.</td>
</tr>
<tr>
<td>The potential resource implications of applying the recommendations have been considered.</td>
<td>1-Clearly stated that did not consider; 2-Not mentioned; 3-Mentioned but not considered; 4-Yes, without economic analysis; 5-Yes, with economic analysis</td>
<td>Consideration of resource implications could vary between guidelines. Some will clearly state that this is not considered in the guideline development. Some will mention resource use or importance of consideration of resource use, but not incorporate this evidence into development of the guideline. Other guidelines will describe and consider resource implications without conducting or reviewing an economic analysis, or consider resource use and implications by including economic analysis in the guideline development process.</td>
</tr>
</tbody>
</table>

**Section 6: Presentation of Guideline Recommendations**

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
</table>
| The different options for diagnosing the condition or health issue are considered. | 1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree; 4-Disagree; 5-Strongly Disagree | Strongly Agree – The guideline document describes/presents a comprehensive list of all available options/diagnostic tests for the condition  
Agree – A few different options/diagnostic tests are described/presented  
Disagree – There is little mention of different options/diagnostic tests  
Strongly Disagree – There is no mention of different options/diagnostic tests  |
| The different options for diagnosing the condition or health issue are clearly presented. | 1-Strongly Agree; 2-Agree; 3-Neither Agree nor Disagree; 4-Disagree; 5-Strongly Disagree | Strongly Agree – The different options are clearly presented in a table  
Agree – Different options are presented together and clearly labeled in paragraphs  
Disagree – Different options are dispersed throughout the document and presented in different paragraphs  
Strongly Disagree – Different options are not clearly presented and difficult to identify |
| Key recommendations are easily identifiable.                         | Yes or No                                                                      | Were recommendations easy to find in the document, such as being differentiated from other text by being boxed or using bold font? |
| There is a rating scheme for the strength of the recommendations      | Yes; No; Unclear                                                               | 'Unclear’ was used as an answer option in rare circumstances where it was difficult to determine whether a scheme was used or whether a scheme described strength of recommendations. |
| If Yes, please describe the rating scheme                             | Free text                                                                      | Provide an overview about the scheme for rating strength of recommendation. e.g. System classifies recommendations as 'strong' or 'weak' |
| CPG has used a visual aid to describe the links between               | Yes or No                                                                      | Was there any visual aid such as a diagram or flow chart used to demonstrate the link between diagnostic test accuracy (e.g. true positive, true negative, false |
Section 7: Comments about Guideline Methodology

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer Options</th>
<th>Answer Option Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is guideline available from clearing house?</td>
<td>Yes or No</td>
<td>Guideline clearinghouses, such as the National Guideline Clearinghouse (NGC), prepare their own summary and presentation of recommendations from a guideline document prepared by another organization.</td>
</tr>
<tr>
<td>If the CPG are summarized in one of the clearing houses, is there any discrepancy between the original CPG and the information on the clearing house site?</td>
<td>Yes or No; If Yes, specify which components are different</td>
<td>Did the guideline clearinghouse summary differ in the description of the guideline purpose/audience, development methods, presentation of recommendations as compared to the original document from the guideline developer?</td>
</tr>
<tr>
<td>Did you have difficulty identifying/finding the guideline from the source (i.e. the organization or clearing house)?</td>
<td>Yes or No; If Yes, elaborate (free text)</td>
<td>How difficult was it to locate the guideline on the website of the guideline developer? Was it difficult to find the webpage with the guideline, did it involve navigating through many pages, were all guidelines clearly presented in a list, were guidelines dispersed throughout the organization’s website, were documents clearly labeled as guidelines, were all the supporting documents clearly visible and grouped together?</td>
</tr>
<tr>
<td>Comments about discrepancies between Methods planned and Methods used</td>
<td>Yes or No, If Yes, elaborate (free text)</td>
<td>Did any of the methods used to formulate the guideline differ from those outlined by the organization or the guideline development manual? Did the guideline development group adhere to the process outlined by the manual?</td>
</tr>
<tr>
<td>Comments</td>
<td>Free text</td>
<td>Any other comments or observations about the guideline?</td>
</tr>
</tbody>
</table>
### APPENDIX 5.2: OVERVIEW OF GUIDELINES WE REVIEWED ON HEALTHCARE RELATED TESTS AND THEIR APPLICATIONS

<table>
<thead>
<tr>
<th>Source</th>
<th>Guideline Development Group (year)</th>
<th>Disease or Condition</th>
<th>Application</th>
<th>Diagnostic Test(s) Investigated</th>
</tr>
</thead>
<tbody>
<tr>
<td>CMA Infobase</td>
<td>Alberta Health Services 2012</td>
<td>Breast cancer</td>
<td>Screening, monitoring, staging</td>
<td>MRI</td>
</tr>
<tr>
<td>CMA Infobase</td>
<td>Canadian TB Committee for the PHAC 2010</td>
<td>Latent TB infection</td>
<td>Diagnosis</td>
<td>IGRA</td>
</tr>
<tr>
<td>NGC</td>
<td>European Association of Urology 2012</td>
<td>Urolithiasis</td>
<td>Diagnosis</td>
<td>US, radiography of KUB, regular and low dose NCCT; Enhanced CT; IVU, blood and urine tests with IRS or XRD analysis</td>
</tr>
<tr>
<td>NGC</td>
<td>European Society of Cardiology 2008</td>
<td>Acute pulmonary embolism</td>
<td>Diagnosis, risk assessment</td>
<td>D-dimer, compression US, CT venography, V/Q scintigraphy, CT, pulmonary angiography, echocardiography; diagnostic strategies by suspected high-risk/suspected non-high-risk PE</td>
</tr>
<tr>
<td>TRIP</td>
<td>Australasian Society for Infectious Diseases 2011</td>
<td>Clostridium difficile infection</td>
<td>Diagnosis</td>
<td>Stool culture, PCR assays, cell-culture cytotoxicity assays, enzyme immunoassays detecting C. diff glutamate dehydrogenase and/or toxin A and/or B</td>
</tr>
<tr>
<td>TRIP</td>
<td>Digestive Health Foundation of the GSA 2011</td>
<td>Gastro-oesophageal reflux disease</td>
<td>Diagnosis</td>
<td>High dose PPI, endoscopy, barium swallow and meal, 24-hr ambulatory oesophageal pH</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>Source</th>
<th>Organisation</th>
<th>Condition</th>
<th>Methodology</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIN</td>
<td>Society of Obstetricians and Gynaecologists of Canada 2008</td>
<td>Hypertensive disorders of pregnancy (hypertension, proteinuria)</td>
<td>Diagnosis, evaluation, prognosis</td>
<td>Measurement of BP (mercury sphygmomanometer, automated BP device) and proteinuria (urinary dipstick testing, urinary protein:creatinine ratio, 24 hour urine collection)</td>
</tr>
<tr>
<td>GIN</td>
<td>American College of Physicians 2008</td>
<td>Osteoporosis</td>
<td>Screening and risk factors</td>
<td>Non-DXA and gold standard (either DXA-defined osteoporosis or occurrence of osteoporotic fracture)</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>NACB for the Academy of American Association for Clinical Chemistry 2009</td>
<td>Metabolic disorders in newborns</td>
<td>Screening, diagnosis</td>
<td>Tandem mass spectrometry</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>NACB for the Academy of American Association for Clinical Chemistry 2011</td>
<td>Diabetes mellitus</td>
<td>Diagnosis, monitoring, prognosis</td>
<td>Urinary glucose, blood glucose, ketone testing, HbA1c, genetic markers, autoimmune markers, albuminuria</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>NICE 2011</td>
<td>Familial hypercholester-olaemia</td>
<td>Screening, diagnosis</td>
<td>Elucigene FH20 and LIPOchip genetic tests, LDL concentration, MLPA, CGA, targeted gene sequencing</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>NICE 2011</td>
<td>Lung cancer</td>
<td>Diagnosis, staging</td>
<td>PET, PET-CT, MRI, SPECT, Scintigraphy, Bronchoscopy ± biopsy, Endoscopic ultrasound fine needle aspiration (EUS-FNA), EBUS-</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>SIGN 2011</td>
<td>Colorectal cancer</td>
<td>Screening, diagnosis, monitoring, staging</td>
<td>TBNA, Non ultrasound-guided TBNA, Cutting needle biopsy, Mediastinoscopy, VATS</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>------------------------------------------</td>
<td>---------------------------------------------------------------------</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>SIGN 2006</td>
<td>Peripheral arterial disease</td>
<td>Diagnosis, monitoring</td>
<td>In Primary Care: ankle brachial pressure index, toe pressure measurement, pulse oximetry, near-IR spectroscopy. In Secondary Care: DSA, duplex US, MRA, CTA</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>U.S. Preventive Services Task Force 2009</td>
<td>Depression</td>
<td>Screening</td>
<td>Screening instruments for depression (none specified)</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>U.S. Preventive Services Task Force 2008</td>
<td>COPD</td>
<td>Screening</td>
<td>Spirometry</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>American College of Radiology 2011</td>
<td>Renal Cell Carcinoma</td>
<td>Staging</td>
<td>CT, X-Ray, MRI, US, FDG PET, Tc-99m bone scan</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>American College of Radiology 2010</td>
<td>Abnormalities of abdominal aorta</td>
<td>Screening and diagnosis</td>
<td>US, US with color Doppler imaging and/or spectral Doppler with waveform analysis</td>
</tr>
<tr>
<td>Obtained direct</td>
<td>Australian National Health and Medical Research Council</td>
<td>Glaucoma</td>
<td>Screening, diagnosis, prognosis</td>
<td>Ophthalmoscopy, optic disc photography, fundus/optical coherence tomography, nerve fibre photography, scanning laser,</td>
</tr>
<tr>
<td>Year</td>
<td>Source</td>
<td>Test/Procedure</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>----------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>Australian National Health and Medical Research Council 2005</td>
<td>Ophthalmoscopy, polarimetry, visual field testing, biomicroscopy, gonioscopy, IOP measurement, central corneal thickness</td>
<td>Ophthalmoscopy, polarimetry, visual field testing, biomicroscopy, gonioscopy, IOP measurement, central corneal thickness</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>ACCP 2012</td>
<td>Deep vein thrombosis</td>
<td>Diagnosis, monitoring, staging</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>WHO 2011</td>
<td>Tuberculosis</td>
<td>Diagnosis, monitoring, staging</td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td>WHO 2011</td>
<td>Active Pulmonary/Extrapulmonary TB</td>
<td>Diagnosis, monitoring, staging</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>CCOPCBE 2009</td>
<td>Head and neck cancer</td>
<td>Diagnosis, staging, recurrence, restaging</td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>Informant from Health Ministry of Chile</td>
<td>Bronchial asthma in adults</td>
<td>Screening, diagnostic uncertainty and assessment, Wright fluorimetry, spirometry, thoracic x-ray, prick test, airway hyperactivity assessment, assessment of blood levels of IgE,</td>
<td></td>
</tr>
</tbody>
</table>

FNA, other cytological specimens, NCB, endoscopic/surgical biopsy, BMA and trephine, flow cytometry, immunohistochemistry, prognostic markers, molecular and cytogenetic studies, B-L/T-cell clonality testing, MRDDM, testing for chromosomal translocations, virus detection, standardization of molecular tests.

Clinical probability assessment, D-dimer, proximal compression US, whole-leg US, venography (CT scan or MRI).

Xpert MTB/RIF assay (PCR).


PET, PET/CT, MRI, CT.
<table>
<thead>
<tr>
<th>Informant from Health Ministry of Chile</th>
<th>Health Ministry of Chile</th>
<th>Cervical cancer</th>
<th>Confirmation, follow up</th>
<th>Secretion of eosinophils</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informant from University of Columbia</td>
<td>National University of Colombia, University of Antioquia, Pontifical Javeriana University</td>
<td>Hodgkin and non- Hodgkin's lymphoma, acute myeloid leukemia, acute lymphoblastic leukemia in boys, girls, and adolescents from 0-18 yrs</td>
<td>Diagnosis</td>
<td>Hematological profile, myelogram, cytometry, cytogenetic test, peripheral blood smear, flow cytometry</td>
</tr>
<tr>
<td>Informant from National University of Colombia</td>
<td>National University of Columbia, Institute of Clinical Research - Faculty of Medicine, HTA and Health Policy Group</td>
<td>Infection, neuroarthropathy, neuropathic foot and ischemia (diabetic foot)</td>
<td>Diagnosis</td>
<td>Conventional x-ray, ultrasound, CT, MRI, nuclear medicine</td>
</tr>
<tr>
<td>Informant from Comprehensive Cancer Care of the Netherlands (IKNL)</td>
<td>Urological Tumours National Working Group</td>
<td>Renal cell carcinoma</td>
<td>Screening, diagnosis, follow-up</td>
<td>History, PE, laboratory tests, diagnostic imaging including: CT, MRI, skeletal scintigraphy, FDG-PET, pathological assessment (biopsy, frozen section analysis, classification, grading, staging,</td>
</tr>
<tr>
<td>Informant from Comprehensive Cancer Care of the Netherlands (IKNL)</td>
<td>The Dutch Society for Radiology</td>
<td>Bladder carcinoma</td>
<td>Diagnosis, follow up</td>
<td>Symptoms, PE, urine cytology and tests, imaging, urethral cystoscopy, biopsy, fluorescence cystoscopy, pathology</td>
</tr>
<tr>
<td>---------------------------------------------------------------</td>
<td>--------------------------------</td>
<td>------------------</td>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Informant from Australian Government</td>
<td>Australian Government MSAC</td>
<td>Thyroid carcinoma</td>
<td>Diagnosis of recurrence</td>
<td>Thyrotropin alfa-rch, THT-withdrawal stimulated preparation</td>
</tr>
<tr>
<td>Informant from Australian Government</td>
<td>Australian Government MSAC</td>
<td>Growth hormone deficiency II</td>
<td>Diagnosis, monitoring</td>
<td>Diagnosis: ITT, arginine + GHRH, GHRH + GHRP, and glucagon alone (stated that arginine alone not considered due to less established diagnostic value); Screening: IGF-I; monitoring: dual X-ray absorptiometry, clinical examination</td>
</tr>
<tr>
<td>AWMF-database</td>
<td>German association of urology 2011</td>
<td>Prostate carcinoma</td>
<td>Screening, diagnosis, staging</td>
<td>PSA-test, digital-rectal examination, biopsy, US, US-elastography, histoscanning, MRI, PET/CT, CT,</td>
</tr>
<tr>
<td>AWMF-database</td>
<td>German society of osteology 2009</td>
<td>Osteoporosis</td>
<td>Diagnosis</td>
<td>Osteodensitometry (dual x-ray-absorptiometry, US), X-rays, basic laboratory test, CT, MRI, Scintigraphy</td>
</tr>
<tr>
<td>AWMF-database</td>
<td>German respiratory society and German cancer society 2010</td>
<td>Lung cancer</td>
<td>Screening, diagnosis, staging</td>
<td>X-rays, CT, sputum test, bronchoscopy, tumour marker, MRI, needle aspiration, thoracoscopy, mediastinoscopy, US, FDG-PET, PET/CT, scintigraphy, histopathology</td>
</tr>
<tr>
<td>AWMF-database</td>
<td>German society for digestive and metabolic</td>
<td>Stomach and gastroesophogeal</td>
<td>Screening, diagnosis, staging</td>
<td>Serological marker, endoscopy, biopsy, pathohistology, US, CT, X-rays, MRI, PET/CT, bone-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>diseases (DGVS), German cancer society (DKG), 2011</th>
<th>carcinoma</th>
<th>scintigraphy, laparoscopy, laboratory test</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWMF-database</td>
<td>German society for general medicine and family medicine 2011</td>
<td>Chest pain</td>
</tr>
</tbody>
</table>
APPENDIX 5.3: DETAILS ABOUT GUIDELINES MAKING RECOMMENDATIONS ABOUT THE SAME TEST FOR THE SAME CONDITION: COMPARISON OF GUIDELINE METHODS AND RECOMMENDATIONS

a. American College of Radiology (ACR) 2011 vs. Comprehensive Cancer Care of the Netherlands (IKNL) 2010 about Renal cell carcinoma staging

The two guidelines looked at diagnosis of renal cell carcinoma using various diagnostic tests. Recommendations on CT, MRI, chest x-rays, and FDG-PET were compared. Box 1 summarises these recommendations. Although the recommendations made by the two groups were quite similar, it is difficult to determine whether their recommendations were based on the same evidence since the ACR does not provide details of the literature search.

When looking at the methodologies used by both groups, IKNL better reported their overall processes and were more transparent. The main difference between the guidelines was their presentation of the recommendations. The ACR, unlike any other group we evaluated, provided recommendations in the form of ‘appropriateness criteria’ using a 9-point scale.

Some other notable differences included: statement of the health question, whether/not patient views and preferences were considered, whether there was stakeholder involvement, and conflict of interest disclosure.
Box 1.

IKNL recommended that ‘routine work-up for staging of renal cell carcinoma includes a multiphase contrast CT and a chest x-ray’ and the ACR assigned an appropriateness rating of 8, meaning that chest x-rays and CT are ‘usually appropriate’ in staging renal cell carcinoma. Further, both groups have similar recommendations for patients with neurological symptoms or suspicion of brain metastases in that a CT scan and/or MRI of the brain is recommended. Both guideline groups state that MRI is a suitable substitute test when the patient cannot undergo contrast CT. One difference in the groups’ recommendations is that the ACR guideline groups their recommendations according to tumour size (< or >3cm), while the IKNL does not make this distinction. Both guidelines suggest that skeletal scintigraphy, or bone scan, to detect bone metastases “is not a routine part of staging patients with renal cell carcinoma” (IKNL) or is “usually not appropriate” (ACR), except in symptomatic patients. Similar recommendations are also made by both groups with respect to the use of 18F-FDG-PET, which state that there is limited evidence in the role of FDG-PET and it is not a standard part of the primary staging of renal cell carcinoma.

b. Cancer Council Australia Guideline on diagnosis and management of lymphoma (2005) approved by NHMRC versus IQWiG guideline on PET or PET/CT for malignant lymphoma (2009)

The IQWiG guideline focuses on the use of PET and PET/CT for the purpose of re-staging or diagnosing recurrence of lymphoma and uses other tests as references or comparators. The Australian guideline covers a range of diagnostic tools and their use in diagnosis, staging,
monitoring and management of lymphoma. In terms of methodology and overall quality, both guidelines provided clear summaries of the evidence used to produce their recommendations. Neither of the guidelines used a rating scheme for the strength of the recommendations and although the IQWIG guideline made conclusions regarding the comparative effectiveness of various tests, they did not make clear recommendations due to a weak evidence base. The IQWIG guideline fully reported financial and intellectual conflicts of interest of each panel member. While the NHMRC guideline provided only a general statement. Additionally, IQWIG guideline used systematic methods, reported all selection criteria and discussed the overall strength and limitations of the body of evidence while this was not clear in the NHMRC guideline. They also did not report clear selection criteria or discuss the overall strength and limitations of the body of evidence. However, the NHMRC guideline did include an economic analysis and considered the views and preferences of the target population including patients, the public and consumers, which IQWIG did not. The actual recommendations developed by the two groups are in box 2.

c. Cancer Care Ontario (CCO) – 2009 vs. IQWiG –2011 Head and Neck Tumours

<table>
<thead>
<tr>
<th>Box 2.</th>
</tr>
</thead>
</table>

**IQWiG:** The value of PET in primary staging is not yet clear. Many studies report that the accuracy of PET is higher than that of gallium scintigraphy; however, it is unclear if this superior accuracy has any effect on patient important outcomes. [translation]

**Cancer Council Australia:** FDG-PET scanning or, if unavailable, gallium scanning, are recommended for staging in all cases. Positron emission tomography (PET) is superior to gallium. [for Hodgkin lymphoma] PET scanning rather than gallium scanning is recommended for response assessment after treatment for Hodgkin lymphoma.
These two peer reviewed guidelines make recommendations about PET and PET/CT as compared to conventional imaging (CT or MRI). Both guideline development utilized systematic methods for their literature search and provided the search strategy. General description of the target population was provided in both documents. Neither group sought patient input and stakeholder involvement was not clear. The link between DTA and patient important outcomes was not discussed by either guideline and neither were resource implications. Additionally, a rating scheme for the strength of recommendations was not used by either guidelines. IQWiG guideline was different in that it used a system to evaluate the quality of evidence, and discussed strengths and limitations of the overall body of evidence. Specific recommendations were provided by CCO guideline while IQWiG guideline gave only general guidance in a narrative summary due to the lack of evidence to impact on patient relevant outcomes (details in box 3).

**Box 3.**

Both guidelines state that PET is useful or recommended after conventional imaging when it is not possible to locate the primary tumour with other methods. However, the IQWiG guideline further adds that although PET may be useful in this application, there is a lack of evidence to prove this, and it does not make a clear recommendation as the CCO guideline does. Similarly for restaging/recurrence, the CCO guideline recommends PET for patients who are being considered for major salvage treatment, while the IQWiG guideline states that PET might be useful, with high sensitivity, but very low specificity with a high rate of false positives. The IQWiG guideline further adds that PET might be slightly better than CT or MRI for this application, but there is not enough evidence to prove this.
d. NICE guideline on diagnosis and treatment of lung cancer (2011) versus the German respiratory society and German cancer society guideline on prevention, diagnosis, therapy and follow-up of lung cancer (2010)

Both practice guidelines were developed to aid in diagnosing, staging and treatment of lung cancer. Both looked at various diagnostic tests, specifically the German guideline considered X-Rays, CT, sputum test, Tumor marker, thoracoscopy and histopathology while the English guideline looked at SPECT, needle biopsy and thoracic surgery. There was a lot of supplementary material available for both guidelines. NICE used a general methods handbook, while the German societies developed a methods paper only for this guideline. PICO questions can be identified clearly in NICE, though in an additional document while German societies guideline described parts of the PICO in its publication. Financial COI was disclosed from both groups but intellectual COI was reported only by NICE. Both considered patient views in developing the guidelines and used systematic methods to search for evidence and described it in details. Additionally, both groups used evidence tables to summarize the evidence. NICE also used cost effectiveness analysis and a narrative summary. In order to formulate the recommendations, both groups used similar methods and described them well. NICE guideline refers to benefit and harms consistently while the German societies consider them less consistently. Though both guidelines described parts of the link between DTA and patient outcomes, they did not agree on consequences or the recommendation itself in some instances (Box 4).
Box 4.

The two guidelines don’t agree on major concepts such as the consequences of TP. NICE guideline describes it as: “an early diagnosis of lung cancer can lead to a better chances for treatment” while the German societies guideline described it as: “a positive diagnosed lung cancer is often inoperable”.

We also compared two recommendations on the use of MRI. NICE: “Magnetic resonance imaging (MRI) should not routinely be performed to assess the stage of the primary tumour (T-stage) in NSCLC. MRI should be performed, where necessary to assess the extent of disease, for patients with superior sulcus tumours.” While the German societies recommendation states “Since CT assessment of mediastinal infiltration or infiltration of the chest wall may not be sufficient, additional methods such as thoracic ultrasound or MRI is recommended. (B) ... In case of a superior sulcus tumor or tumor of the lung apex, a MRI is strongly recommended for assessment of tumor extension such as plexus involvement (A)”
CHAPTER 6: WHAT DO EXPERTS SAY?
Decision-making about healthcare related tests and diagnostic strategies: what do experts say?

Reem A. Mustafa, MD, MPH\(^1\,2\), Wojtek Wiercioch, B.H.Sc\(^1\), Matthew Ventresca, B. Sc.\(^1\), Holger J Schünemann, MD, MSc, PhD\(^1,3\) for the DU-Diagnosis expert group.

(32) Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Canada

(33) Department of Medicine, University of Missouri-Kansas City, Kansas City, USA

(34) Department of Medicine, McMaster University, Hamilton, Canada

DU-Diagnosis expert group (listed alphabetically)

Hanan Bell, PhD, Michael Bettman, Patrick Bossuyt, MSc, PhD, George Browman, Jan Brozek, MD, Diedrich Bühler, MD, Roger Chou, MB.BCh, MD, Andrew Don-Wauchope, Glyn Elwyn, Amit X. Garg, MD, PhD, Davina Gershi, Massimo Gion, Paul Glasziou, MD, Robin Harbour, Andrea Rita Horvath, MD, Ina Kopp, Murray Krahn, Rosanne Leipzig, Monika Lelgemann, MD, MSc, Nancy Lloyd, MSc, B.Sc., Saijonkari Maija, Marjukka Mäkelä, MD, PhD, M.Sc (ClinEpi), Richard Mendelson, MD, Michelle Mujoomdar, Susan Norris, Martin Reed, Denis Remedios, FRCR, Nancy Santesso, RD, MLIS, Stefan Sauerland, MD, MPH, Karen Steingart, MD, MPH, Toni Tan, Amir Qaseem, MD, PhD, MHA
ABSTRACT

Objective
To identify the essential factors to consider when developing recommendations and making decisions about health care related tests and diagnostic strategies (HCTDS).

Methods
We used a qualitative descriptive approach and conducted semi-structured in-depth interviews with 24 international key informants in assessing evidence and making decisions about HCTDS.

Results
Although test accuracy (TA) was the factor most commonly considered by organizations when developing recommendations, experts agreed that TA is rarely, if ever, sufficient and may be misleading when solely considered. Informants identified factors related to test accuracy, factors related to clinical decision making, overall benefits and harms, resource implications, patients’ and societies’ values and preferences, quality of evidence, in addition to feasibility, applicability, ethical, legal and organizational factors as essential to consider when developing recommendations about HCTDS. Informants mostly agreed that laying out the potential care pathway based on the test result is an essential early step in developing recommendations about HCTDS but is rarely done unless decision analysis and mathematical modeling are utilized and even then it may not be explicit. Most informants also agreed that decision analysis and mathematical modeling could be useful for organizing the clinical, cost, and preference data relevant to the use of tests in the absence
of direct evidence. However, they noted that utilizing models is limited by the resources and expertise required.

**Discussion**

Developing guidelines about the use of HCTDS requires consideration of factors beyond solely TA, but implementing this may be a challenge. Further development and testing of “frameworks” that can guide this process is a priority for decision makers and guideline developers.
WHAT IS NEW

• In this qualitative descriptive study we explore current practices, challenges and potential solutions for developing recommendations and making decisions about healthcare related tests and diagnostic strategies (HCTDS).

• Informants agreed that test accuracy (TA) results are rarely, if ever, sufficient and may be misleading if considered alone for making decisions about the appropriate use of tests.

• Informants identified the essential factors in making decisions about HCTDS including: factors related to test accuracy, factors related to clinical decision making, overall benefits and harms, resource implications, patients’ and societies’ values and preferences, quality of evidence, in addition to feasibility, applicability, ethical, legal and organizational factors.

• Although outlining care pathways to illustrate the use of a test, with treatment options, if any, and outcomes of the different options helps improve transparency and facilitate the process of developing recommendations, guideline panels rarely explicitly lay out the care pathway.

• Although decision modeling requires additional resources, it may be useful for integrating different relevant factors as well as identifying evidence gaps and high priority research areas.
BACKGROUND

In the four previous articles in our series addressing decision-making about healthcare related tests and diagnostic strategies (HCTDS), we provided an overview of known roles and applications of tests, different diagnostic study designs, and the main challenges encountered in conducting and applying research about diagnostic interventions (Reference paper 1). We also described the research supporting the development of evidence tables for the presentation of test accuracy information from systematic reviews (Reference paper 2). Additionally, we identified tools used to assess the quality of evidence (confidence or certainty of the evidence) and strength of recommendations of HCTDS and described how guidelines about HCTDS are developed (Reference papers 3 & 4). This fifth of a series of eight articles describes results of interviews with key informants in developing guidelines about HCTDS regarding the essential factors in making decisions about appropriate use of HCTDS and how to integrate these factors.

There is general acceptance that the challenges for decision makers about HCTDS may be considerably different from decisions addressing therapeutic interventions alone. Current practices and considerations when making decisions about the use of HCTDS vary among organizations and groups making these decisions (Reference papers 3 & 4). This might be due to the frequent lack of direct evidence about the effect of HCTDS on patient and population health.

The purpose of this qualitative descriptive study was to explore current practices, challenges and potential solutions when making decisions and developing recommendations about the appropriate use of HCTDS.
METHOD

Approach
We used a qualitative descriptive approach to guide the conduct of this study\(^1\). A qualitative research approach allowed for an in-depth exploration of a range of concepts to understand the experience developing recommendations and making decisions about HCTDS from the perspective of the study sample (informants and experts in the field)\(^2\). We used a descriptive approach to avoid predetermined theoretical frameworks that could bias the findings. In addition, we were aiming to clarify the factors used when making decisions including developing recommendations about HCTDS and how they are integrated together, and to explore suggestions by informants for how to develop this field further.

The study was reviewed and approved by the Hamilton Integrated Research Ethics Board. All informants had to provide oral consent to be interviewed and to be audio-recorded. To allow informants to express their opinion freely, we did not request official endorsement from the organizations. Hence, the results of this research should not be viewed as an official statement from organizations.

Participants and organizations
We used maximum variation purposeful sampling by contacting international organizations and groups that assess evidence and produce guidelines and make decisions about HCTDS. We asked leads in these organizations to identify informants in each of the organizations who have most expertise with processes when developing recommendations about HCTDS. We describe the details about how we identified these organizations in articles three and four of this series (Reference paper 3 & 4). If more than one informant
was identified in the same organization, we purposefully contacted informants with a variety of backgrounds. Although originally we aimed for 12-15 interviews, our sample included all informants that agreed to the interview. Some informants suggested others who also had expertise in decision-making about HCTDS and who could enrich the variability among our sample, and we contacted all those suggested.

Data collection

We conducted semi-structured in-depth interviews for 90 minutes with participants either in-person or virtually using electronic video conferencing. During the interviews, we followed a detailed and standardized interview guide that we pilot tested extensively (Appendix 1) with members of the research team who were not involved in the original development of the guide. For the majority of questions, we solicited factual information about processes used to produce guidelines and make decisions about HCTDS, the challenges encountered and how different organizations and groups handled them. Other questions drew on informants’ views, for example their impression about characteristics of ideal tools and how best to handle challenges specific to guidelines and decisions addressing HCTDS. Box 1 summarises the specific overarching topics covered in the interviews. We also collected background information including organization/s, position and type of work or responsibilities, education in research methods, clinical area, experience in developing and/or leading guideline development and number of decisions including guidelines in which they participated. Between October 2012 and February 2013, one investigator (RAM) conducted the interviews.
Box 1:

Topics addressed during interviews with experts about developing guidance and making decisions about HCTDS

1. Advantages and disadvantages of available tools to assess the quality (certainty) of the evidence and strength of recommendations

2. Differences between HCTDS and therapeutic recommendations.

3. Factors considered and how can decision makers integrate these factors.

4. Considering care pathways by different organizations and groups.

5. When are test accuracy results sufficient for decision-making?

6. Advantages and disadvantages of using decision analysis and mathematical

Qualitative analysis

We used basic content data analysis using a data coding system that corresponds to the data collection. We used conventional content analysis as the goal of our study was descriptive and there is little existing theory to guide our analysis. Also this allowed the coding categories to come from the text to produce a rich description of what are the current practices of developing guidelines about HCTDS based on the impressions of the key informants. We used a deductive coding to summarize the finding under each of the topic highlighted in Box 1. We also used an open coding procedure followed by a higher level of synthesis especially for the two topics: factors considered and when are test accuracy results sufficient for decision-making. We then grouped codes into categories and abstracted meanings.
A research assistant transcribed the interviews verbatim with identifying information removed; transcripts cleaned; then read in their entirety. We used a combination of deductive and inductive coding process. Two investigators (RAM and WW) completed all data coding independently and in duplicate. After reading through the transcriptions several times to obtain a sense of the whole interview, RAM and WW derived codes that captured key concepts from reading the data word by word. The initial codes were determined after coding two interviews and comparing results. Coding was refined again after completing each subsequent groups of 5-6 interviews. The establishment of themes and sub-themes came later, and reflected patterns in the data. While the initial topics reflected questions and key concepts from the interview guide (Box 1; itself informed by our previous work (Reference papers 1-4) the evolving codes, especially for the two topics; Factors considered and when are test accuracy results sufficient for decision-making, reflected the reflexive nature of qualitative content analysis, which is data-derived. When data analysis was almost completed, members of the research team (RAM, WW and HJS) discussed refinements to the coding schema.

**Data Trustworthiness**

We enhanced credibility and confirmability through several strategies. The investigators (RAM and HJS) had previously met and worked with several of the informants which allowed for the building of trust and rapport as well as provided the researcher with a greater understanding of the context. Peer debriefing, in the form of ongoing discussions with members of the research team allowed for the testing of new ideas and insights throughout the research process. We conducted member checking by sharing a written synthesis of the overall study data with the informants and inviting them to comment on the accuracy of the findings. RAM and HJS conducted a 90-minute conference call with
informants to discuss the compiled findings and further clarify any questions that emerged during coding. Finally, we included representative quotes from participants to further enhance credibility⁹.
RESULTS

Participants and organizations

We interviewed 24 informants (20 identified by contacting organizations and 4 suggested by other informants) from Australia, Canada, Finland, Germany, Italy, the United Kingdom, and the United States. Some had experience developing guidelines about HCTDS in low and middle-income countries as well as in high-income countries. Participants had a broad range of backgrounds, formal training and experience. They worked in different fields including clinical research, community and public health, general clinical practice, family medicine, internal medicine, geriatrics, oncology, pharmacology, radiology, pathology, laboratory medicine, biochemistry, nursing, decision support tools development, health technology assessment and health economics. In varying capacities, all interviewed informants have been involved in the decision-making processes with international agencies, societies, or organizations such as the American College of Physicians, the Royal Australian College of Physicians, the Canadian Association for Radiology, the Canadian Association of Medical Biochemists, the Canadian Partnership Against Cancer, the Diagnostic Imaging Pathways, the European Society of Radiology, the German Institute for Quality and Efficiency in Health Care (IQWiG), the Global Summit on Radiological Quality and Safety, National Academy of Clinical Biochemistry, the National Agency of Regional Health Services in Italy (Age.na.s), the National Health and Medical Research Council of Australia (NHMRC), the National Institute for Health and Clinical Excellence (NICE), the Royal College of Radiologists, the Scottish Intercollegiate Guideline Network (SIGN), the United States Preventive Services Task Force (USPSTF), and the World Health Organization.
(WHO) among others as almost all informants had experience with multiple groups over the years.

**Topics and subtopics**

Table 1 summarizes the final topics that emerged from our data analysis.

**Guidelines about HCTDS and therapeutic options**

**Similarities and differences**

Informants agreed that the primary aim of using both HCTDS and therapeutic interventions should be to improve overall patient and population health.

Informants also agreed that formulating questions and developing recommendations about HCTDS is generally more complex than for therapeutic ones. The paucity of direct evidence evaluating the whole care pathway, including HCTDS’ effect on patient outcomes, adds to this complexity since the entire care pathway from diagnosis through treatment must be considered. Additionally, informants recognized that while methods for rigorously studying therapeutics (particularly pharmaceuticals) are well established, there is a lack of methodological rigor in primary TA studies and a lack of awareness or implementation of standard methods to evaluate the risk of bias of TA results. Also, there is generally more evidence available evaluating a single test’s performance and fewer studies directly comparing the effects of tests or their accuracy in the same population.

While implementation of new treatments typically follows extensive evaluation by different regulatory bodies, informants noted that in some jurisdictions, there is a lack of regulatory standards prior to wide dissemination of “new” tests. This can lead to inappropriate use of tests, an issue that frequently requires addressing during the development of recommendations about HCTDS.
HCTDS recommendations, similar to therapeutic ones, can address many questions including: testing versus no testing, which test to use initially or for follow up in the care pathway, timing of testing, and test sequences. HCTDS recommendations, as opposed to treatment ones, less frequently compare different types of tests (e.g. radiology versus biomarkers). Informants noted that this is true as many panels are based on the interests of professional societies and groups and not on clinical conditions (radiology groups, biomarker groups, etc).

Tests may have different applications (screening, classifying baseline risk, confirming or ruling out a diagnosis, staging, triaging, monitoring treatment, and determining prognosis). Informants generally noted that although recommendations about HCTDS can address all these applications, they usually focus on testing versus no testing or comparing two tests. Informants pointed out that the rest of the applications and questions, generally follow an even less rigorous process. Additionally, recommendations about HCTDS can address other aspects about using a test such as the best cut-off value to use in practice, an issue that does not apply to treatment recommendations.

**Guidelines about HCTDS - Using care pathways**

Informants noted that creating flowcharts or decision trees to illustrate the use of a HCTDS, with treatment options, if any, and outcomes of the different options following a specific test result could help improve transparency and facilitate the process of developing recommendations but is rarely done. They also noted the lack of formal guidance on how to formulate these care pathways.

Informants described that it is important to have discussions about care pathways among panel members and to incorporate it into protocols during early stages of the guideline.
development process. This applies regardless of the test’s role or application, as it is essential to clarify the integration and sequencing of the test with other HCTDS and interventions. Informants explained that using informal frameworks and abstract scenarios could facilitate discussions to identify potential care pathways. Some informants also clarified that outlining the appropriate care pathways is the first step to build realistic decision analyses. It was also notable that considering the complete care pathway helps in the development of decision rules that aid clinicians in practice, e.g. to provide guidance on what to do if a test is negative or if a test is positive. Additionally, informants noted that following an iterative process supported by discussion with panel members to clarify care pathways also helps identify the studies that qualify to be included in the body of evidence informing each recommendation.

**Guidelines about HCTDS - Test accuracy results are rarely sufficient for decision-making**

Although informants often responded that TA results could be sufficient for making decisions about HCTDS in some situations, it proved very challenging to find practical examples of these situations. More in-depth discussion also highlighted that TA alone is rarely, if ever, sufficient; “there normally has to be consideration of how the test can be applied in practice, acceptability to patients, cost impact…. etc.” Table 2 summarizes the situations where TA results may be sufficient to extrapolate about overall benefits and harms.
**Guidelines about HCTDS – Essential factors to be considered**

Table 3 summarizes the factors identified as essential in making decisions about HCTDS. It is important to note that all or most of these factors are usually considered simultaneously when developing recommendations about HCTDS. However, informants noted that the process of how or which ones are considered is not usually explicit. Additionally, informants noted that one or more factors might play a more important role compared to the rest in specific circumstances or decisions.

**Guidelines about HCTDS – Approaches to integrate factors**

In most organizations, it was not clear how the factors in developing recommendations are specifically considered, unless there was a formal framework or modeling involved which, as we describe below, is rarely used.

We identified the following patterns in assessing benefits and harms of using a test: 1. Take test accuracy into consideration, but do not specifically consider the links to benefits, harms and patient outcomes. 2. Consider the links from a broad perspective in assessing the benefits and harms of using a test in the clinical pathway without specifying how they were considered. 3. Make assertions about what might be the consequences of FP, FN, TP, TN test results by using the elements of a typical 2x2 table and consider the flow of patients in the care pathway with treatment options and patient outcomes. 4. Use decision analyses and mathematical models to determine quantitatively the benefits and harms for patients and populations.
Decision analyses and mathematical models

Organizations varied in how they utilized decision analysis and mathematical modeling to develop recommendations about HCTDS, but overall these were not commonly used. The variability ranged between not using models at all, to formally developing them by trained health economists who model patient outcomes and costs to produce cost effectiveness or cost utility analyses. Also, few groups adapt already existing models to their local settings or use simple decision trees to support their decisions without going through fully developing or adapting models. Informants mostly viewed decision analyses as useful for organizing the clinical, cost, and preference data relevant to the use of tests. However, they noted that these models required intensive financial, time and human resources, which are not always feasible. Lack of clear guidance on when and how to use modeling was another challenge identified by informants. However, informants reported that an advantage for modeling is that it allows panels and decision makers to assess variations in clinical practice that have the potential to influence recommendations about HCTDS.

Informants noted that the quality of decision analyses and mathematical models must be taken into consideration as well as the source of information used for different variables in the models. Additionally, informants agreed that guidelines, policies and decision support tools, should all be based on systematic synthesis of the best available evidence whether models are used or not.

Some informants also pointed out that some “degree” of modeling is unavoidable when developing recommendations about HCTDS, but that it is frequently done implicitly and informally and that following a more formal process makes it more transparent and rigorous. However, some informants noted that there are different “degrees” of modeling and their applications. Some informants noted that some applications of models are
“weaker” (decreases our confidence in the effect of using a test on patient health) than others. For example, applying simple models to consider baseline risk is different than when modeling is used to extrapolate the effect of testing in a population that has never used that test. Both are examples of applying models to inform recommendations about HCTDS.

- Analytical and conceptual “frameworks”

Informants noted that “frameworks” to better incorporate/account for all or most of the different factors is needed regardless of the use of a mathematical model to quantify the trade-offs of benefits and harms. Informants noted that decision analyses and mathematical models may be helpful for elucidating some (but not all) of these factors in a more formal way. Informants also noted that the lack of formal “frameworks” to guide integrating multiple factors including feasibility, acceptability and ethical and legal considerations may allow panel members, policy makers and stakeholders to influence or bias the recommendations and decisions about HCTDS.

Value of tools to assess the quality of evidence (QoE) and strength of recommendations (SoR) about HCTDS

Most informants agreed that using a tool to rate the QoE is useful in addressing the variability inherent in individuals’ judgments and the potential biases that can affect it. Using a tool to assess QoE was also noted to enhance transparency in systematically evaluating the evidence and developing recommendations.

On the other hand, informants’ views about the usefulness of using a system to rate the SoR varied. While most agreed that using such a system allows guideline panels to express
limitations of their decisions and how strongly they recommend that course of action, many were sceptical for different reasons. Some felt that a tool that allows for strong recommendations to be made in the absence of high quality evidence might lead to lack of incentive for researchers to gather higher quality evidence. Some believed that strength of recommendation might not be suitable when providing guidelines about using a test in the patient care pathway as a whole.

Informants also noted that the availability of multiple tools evaluating QoE and/or SoR without clear guidance on how to use them was associated with confusion among users. This highlights the importance of experience in health research methods as well as the need for specific training about the tool itself. Furthermore, the concepts of QoE and SoR are often entangled and misinterpreted. One informant feared that using a tool in a “formulaic way” to assign QoE and SoR may limit critical thinking about the evidence and the recommendations, particularly with weak/conditional recommendations. Informants also explained that this is further complicated by the fact that most grading systems are targeted for therapeutic or management questions, rather than HCTDS ones. Also noted in the interviews that these systems are often tailored to groups producing guidelines as opposed to those using them.
DISCUSSION

In this study, we interviewed 24 international informants and experts in developing recommendations, producing decisions support tools and making decisions about the use of health care related tests and diagnostic strategies. Informants identified factors related to test accuracy, factors related to clinical decision making, overall benefits and harms, resource implications, quality of evidence, patient and society values and preferences, in addition to feasibility, applicability, ethical, legal and organizational factors as essential to consider when developing recommendations about HCTDS. We also outlined four potential cases when TA is sufficient to make conclusions about patient benefits and harms when developing recommendations about HCTDS.

This study has multiple strengths. We obtained input from informants and experts involved with major international organizations in making decisions and developing recommendations about HCTDS. Interviews with these experts were quite extensive and addressed different aspects of the decision-making process. Also, many of the informants had worked for multiple organizations and were able to report on differences in approaches that further enriched the interviews. We used a semi-structured pilot tested script that while maintaining a common structure allowed flexibility to dig deep into the specifics about what is actually happening in different organizations and what the contributors believed, based on experience and background, should be happening. Also, we contacted contributors and clarified questions about their responses to confirm that our results were accurate. Additionally, we organized a conference call to discuss the findings and facilitated additional discussion about areas that needed further clarification. We also circulated an early draft of the results and received feedback and comments which we then
incorporated into our article. Two investigators separately coded all interviews to minimize bias.

This study has a few limitations. Our findings are not officially endorsed by the organizations represented by the informants that we interviewed. However, this may be viewed as a strength given that the experts were able to express their honest opinion and observations freely. Additionally, one may argue that the views of our sample of contributors may have not represented all experts in this area. We contacted major organizations known to be active in decision-making about HCTDS. We also asked an international group of informants to suggest additional organizations or individuals with expertise in this area. Hence, we ended up with a large, and we believe, representative sample of experts to summarize current practices in making decisions about HCTDS and reflect on what they believe should be done to move this field forward.

Although TA was the factor most commonly considered by organizations when developing recommendations about HCTDS, experts and informants agreed that TA is rarely if ever sufficient and may be misleading. They also agreed that the effects of the HCTDS on patient important outcomes and what would be materially relevant to patients if they were having a test are the important considerations when developing recommendations and making decisions. However, organizations rarely address tests effect on patient and population health as a result of lack of direct evidence. This is inflated by lack of guidance on the mandatory and minimum acceptable factors to be considered and how to integrate them for decision-making about HCTDS.

Informants identified four cases for when TA is likely sufficient to extrapolate about tests’ effect on patient important outcomes. These cases include: “win win situation” when all else is equal, when diagnostic non-inferiority is sufficient for a decision when inferences
can be made about the impact on patient important outcome, when the TA of one test is equivalent or better than the combined accuracy of two tests, and when the primary goal is to establish a diagnosis for a condition or rule out a condition. This last case was the most controversial as many informants highlighted that while theoretically it may be possible, realistically in all the examples identified there is an assumed link to patient important outcomes like relieving anxiety, family planning or satisfaction because of the tests’ results. Our results support that developing recommendations about the use of HCTDS may be more complex than for developing treatment recommendations. It requires considering multiple factors beyond TA results. However, addressing this complexity and actually considering all the factors identified in this qualitative study is challenging. In the first article of this series we highlighted the main challenges in the area of developing recommendations about HCTDS including the lack of direct evidence assessing tests’ effect on patient important outcomes, the methodological challenges in TA studies, the fast technology development with the lack of regulatory standards to control their use before proven beneficial among others. We believe the lack of direct and high quality evidence evaluating HCTDS effect on patient important outcomes should not lead to more lenient standards about assessing them. Rather it should be the bases for demanding more informative and higher quality evidence to support decision making in this area. And if that is not feasible, users should be aware of the limitations and our reduced confidence in the available evidence.

Developing recommendations about HCTDS involves correctly placing the HCTDS in the care pathway by conceptually linking different potential options early in the process of guideline development. Although clear guidance about how to outline these care pathways is lacking, experts generally believed that creating flowcharts or decision trees helps
improve transparency and facilitates the process of developing recommendations. Outlining care pathways has multiple advantages including identifying evidence needed for decision making at the early planning phases of the process. This clear mapping of the needed evidence and how it all relates to each other will also typically assist in detecting evidence gaps and priority research areas.

It appears that resource implications are always taken into account when developing recommendations about HCTDS even in jurisdictions that prohibit considering cost. Although resource implications are not commonly explicitly laid out in guidelines unless economic analyses are utilized, the unit and total volume cost of the test are commonly the drive to launch the guidelines about HCTDS in the first place.

We believe guidelines, policies and decision support tools about HCTDS should all be based on the best available evidence. Hence, they can inform each other. Performing formal decision analysis and mathematical modeling is limited by the time and resources required; in addition the frequent lack of high quality evidence to inform these models may also lead to inaccurate conclusions. Rules and criteria are needed to determine when to use and when not to use decision analyses and modeling.

Formulaic use of tools assessing the QoE and SoR of HCTDS is problematic and should be avoided. The solution is not to avoid the use of these tools and accept suboptimal alternatives but rather to enhance training and correct misunderstandings about the tool. This is supported by informants’ views that following a clear framework to assess the QoE and SoR based on appropriate understanding of a tool limited variability inherent in individuals’ judgment and the potential biases that can affect it.

In summary, it appears that there is general agreement among experts that multiple factors beyond TA results should simultaneously be considered to optimally develop
recommendations and make decisions about HCTDS. Ideally, such factors should also be systematically reviewed. Outlining realistic care pathways is essential for properly developing recommendations about HCTDS and identifying evidence gaps and research priorities. Formal decision analyses and mathematical modeling may be useful tools to inform guidelines about HCTDS. However, there remains considerable variability in practices among leading organizations highlighting that practical solutions to current challenges facing decision makers about HCTDS (such as lack of expertise and resources) are needed. Also, further development and testing of frameworks to guide this process are needed.
Acknowledgement

We thank Dr. Susan Jack, Associate Professor on nursing for her support and expert advice on qualitative data analysis. This work was partially funded by the German Insurance Fund agency as part of a larger project about decision-making for healthcare related tests and diagnostic strategies. The views presented here are those of the authors and should not be attributed to the funding agency or its staff.
**TABLE 6.1: OUTLINE OF TOPICS AND SUBTOPICS**

<table>
<thead>
<tr>
<th>Topics</th>
<th>Subtopics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guidelines about HCTDS and therapeutic options</td>
<td>- Similarities and differences</td>
</tr>
<tr>
<td>Guidelines about HCTDS</td>
<td>- Using care pathways</td>
</tr>
<tr>
<td></td>
<td>- Test accuracy are rarely sufficient</td>
</tr>
<tr>
<td></td>
<td>- Essential factors to be considered</td>
</tr>
<tr>
<td></td>
<td>- Approaches to integrate factors</td>
</tr>
<tr>
<td></td>
<td>-- Decision analyses and mathematical modeling</td>
</tr>
<tr>
<td></td>
<td>-- “Frameworks”</td>
</tr>
<tr>
<td>Available tools to assess quality of evidence and strength of recommendations about HCTDS</td>
<td>- Advantages and disadvantages</td>
</tr>
</tbody>
</table>
TABLE 6.2: TEST ACCURACY RESULTS MAY BE SUFFICIENT TO EXTRAPOLATE ABOUT OVERALL BENEFITS AND HARS

<table>
<thead>
<tr>
<th>Theme</th>
<th>Example</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Win-win” situation when all else is equal</td>
<td>Replacing culture by PCR tests to identify a microorganism. Both tests are based on a blood draw so complication profile is similar while the accuracy (both sensitivity and specificity) of PCR is higher.</td>
<td>“All else being equal [(e.g. side effects, invasiveness, resources considerations, timing of test, location of test in care pathway, feasibility)], you can make a decision if accuracy [of the new test] is better.”</td>
</tr>
<tr>
<td>Diagnostic non-inferiority is sufficient for a decision when inferences can be made about the impact on patient important outcomes</td>
<td>“New” HIV diagnostic test is non-inferior to an existing HIV diagnostic test, given that benefits of treatment of HIV are already established beyond doubt. It is crucial that diagnosis by the “new” test identifies the same spectrum of the condition and hence it can be expected that treatment effectiveness is the same for those cases identified by the existing test and those identified by the “new” test.</td>
<td>“When diagnostic non-inferiority is sufficient for decision-making because other benefits of the “new” test are beyond doubt... when there are clearly established and effective treatments for the condition to be diagnosed”</td>
</tr>
<tr>
<td>When the test accuracy of one test is equivalent or better than the combined accuracy of two tests (one of which is the test evaluated separately)</td>
<td>Plain radiography does not add any clinically significant advantage to multidetector row computed tomography in diagnosing cervical spine injuries in blunt trauma patients.</td>
<td>“When comparing a single test A with the combination of two tests A + B, it is clear that the single test is to be preferred, when offering the same TA. In this scenario, all other issues such as safety, feasibility and resources don’t need to be considered, because A is included in A+B.”</td>
</tr>
</tbody>
</table>
| When the primary goal is to establish a diagnosis for a condition or rule out a diagnosis * | -Genetic testing: Some people want to know about a diagnosis of Huntington’s even though nothing can be done for it. However, this information may have impact on future decisions like family planning and will also affect patient important outcomes like increasing or decreasing anxiety.
- Some people want to look at their babies before they are born with ultrasound even if they have no intention of doing anything about abnormalities they might detect. | “The test results are informational i.e. to be used for shared decision-making.. etc and not to drive a specific therapeutic decision ... or to inform use of subsequent interventions. However, this rationale could be used to justify a lot of unnecessary tests e.g. MRI for low back pain, because “the patient just wants to know”[which is not acceptable as a reason]” |

* This theme was the most controversial among informants, while theoretically it may be possible, realistically in all the examples there is an assumed link to patient important outcome like relieving anxiety, family planning or satisfaction and inv because of the tests' results.
### TABLE 6.3: ESSENTIAL FACTORS IN MAKING DECISIONS ABOUT HCTDS

<table>
<thead>
<tr>
<th>Themes*</th>
<th>Subthemes</th>
<th>Explanations</th>
<th>Quotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors related to test accuracy</td>
<td>Test accuracy results</td>
<td>Which can be expressed in relative terms like sensitivity, specificity, likelihood ratios, diagnostic odds ratios among others or absolute terms like true positive and negatives and false positives and negatives.</td>
<td>“Test accuracy is important but is only the first step for considering recommendations about a test. Formulating a recommendation about a test is a multifactorial process where accuracy is necessary but not sufficient.”</td>
</tr>
<tr>
<td></td>
<td>Pre-test probability, prevalence, baseline risk</td>
<td>Pre-test probability can impact the usefulness of a test. It can be assessed based on population demographics and patients' comorbidities.</td>
<td>“In the most extreme case, prevalence may be 100%. Imagine diagnostic imaging in patients with cancer of unknown primary; if it is certain that a primary cancer is growing somewhere in the body, most clinicians try all available tests until the cancer is located”</td>
</tr>
<tr>
<td></td>
<td>Frequency and effect of inconclusive results</td>
<td>These are the results that do not answer your question (E.g, results that do not role-in or role-out a diagnosis)</td>
<td>“For example, in certain patients gallbladder does not show very well on ultrasound, you have to go on and do further testing.”</td>
</tr>
<tr>
<td>Factors related to clinical decision making</td>
<td>Alternatives in the care pathway</td>
<td>Alternatives may include no testing, other tests, or performing the test at a different time; each of which can be associated with additional benefits or harms. It is important to address how the test results affect these alternatives.</td>
<td>“Diagnostic testing is part of the bigger picture and needs to be considered, [with its alternatives], in the management strategy”</td>
</tr>
<tr>
<td></td>
<td>Application of the test</td>
<td>Including screening, classifying baseline risk, confirming or ruling out a diagnosis, staging, triaging and monitoring treatment, and determining prognosis</td>
<td>“The importance of test accuracy will vary given the question...[For some applications] like screening, sensitivity is more important than specificity.”</td>
</tr>
<tr>
<td></td>
<td>Frequency and effect of incidental findings</td>
<td>Typically tests that have high chance of identifying incidental findings are less preferred</td>
<td>“A test that identifies incidental illness will be less preferred over an equally accurate test that does not identify incidental illness.”</td>
</tr>
</tbody>
</table>
| Overall benefits and harms | The link between the test results and patient important outcomes | This includes multiple factors like diagnostic and treatment impact and natural progression of a disease. Benefits and harms may include feeling more or less anxious due to test results. | “A test may be accurate but it has to be useful to the patient [before it is recommended].” “As much as we consider accuracy, we don’t specifically look at the links. In some cases you could afford more FP than in other scenarios, but we don’t think about that very
| Additional benefits of the test | For example, performing colonoscopies has the additional benefit of removing polyps over capsule endoscopy | “Some tests may have therapeutic effects” |
| Direct harms and safety considerations | May or may not relate to the invasiveness of the test | “It is very important that those of us in a position to advise about radiation safety do bring it into the equation and then offer to draw attention to that factor... particularly in children.” |
| Values and preferences | Patient and society values and preferences | It is important to specify whose perspective is being considered, as patient values may be different to individuals who do not have the condition. | “Not just values and preferences in relation to the test itself but also in relation to the treatment options.” |
| Physician values and preferences | These may vary from patient values and preferences | “[If physicians do not prefer the test they will not order it]” |
| Resource implications | Direct resource requirements due to testing | Even when costs are not formally considered, they will be considered informally and not systematically. For example, cost may be a factor in deciding to commission a review of the evidence and develop a guideline about a test in the first place. | “Both [unit cost and total volume cost] will be considered...For example, if it’s a cheap vitamin D [test], that volume will be considered as well” |
| | Indirect resource requirements | Including downstream resource requirements | “Diagnostic tests may have lower direct costs than drugs, but higher indirect costs due to inappropriate requests and use.” |
| Quality of evidence (certainty about the effect estimate) | Certainty about test accuracy results | Including certainty about baseline risk |  |
| | Certainty about linked evidence | Including certainty about treatment effect and natural progression history among others | “Typically, there is more uncertainty when considering patient-important outcomes and implementation factors than there is about TA. This should not be the reason to refrain from looking at the whole picture when making decisions about diagnostic tests.” |
| | Certainty about other factors like resource implications and | There is variability in resources considerations by different groups including: explicit consideration based on |  |
| values and preferences. | systematic evidence syntheses and implicit consideration based on panel discussion which affect our certainty about the resource implications. This also applies to systematic syntheses of values and preferences vs assumptions about the values by panel members. |
| Overall certainty about the effect of the test on patient outcome | Panel members for the most part agreed with this subtheme but there was some variability in opinion about how to express our overall certainty; whether to include an overall estimate of our certainty or to keep the certainty for each of the factors separate. Another point of view presented is that “the key consideration in rating the QoE (our confidence in the effect estimates) is uncertainty about each of the factors considered in decision-making, and trying to make an overall summary score is a methodological mistake...there are no hard rules and there are too many aspects that can be taken into account. It is difficult to determine a priori which QoE criteria will have the most weight for a specific problem or diagnostic question. Outlining the uncertainty around all the factors considered allow a judgment about how secure (i.e. strong) the final decision is.” |
| Additional factors | Ethical and legal considerations | Which are typically setting specific and may vary considerably between different legislations and cultures. "Typically, there is more uncertainty when considering patient-important outcomes and implementation factors than there is about test accuracy. This should not be the reason to refrain from looking at the whole picture when making decisions about diagnostic tests.” |
| Feasibility, applicability and organizational considerations | Including availability of the test, feasibility of conducting the test, ease of use of the test, availability of skills required to apply the test, ease of interpreting the test results, access to treatment, access to test, training required for interpretation of results. "Am I going to get the results in a reasonable time?” “Availability and practicality in a rural setting.” “Space for test-positive patient in clinics or hospitals when the test identifies more patients.” |

* These themes and subthemes are not mutually exclusive; some are interrelated.
References

CHAPTER 7: CONCLUSIONS
CONCLUSIONS AND RECOMMENDATIONS

The process by which treatment and diagnostic recommendations and coverage decisions are developed should be relatively similar. However, we agree with the majority of experts who stated that diagnostic recommendations and coverage decisions were generally more complex and difficult to produce. Experts identified the following primary reasons for this complexity: First, unlike treatment recommendations diagnostic ones had to be cognizant of the entire care pathway including subsequent management and the consequences on patient important outcomes. These pathways are often long and very complex including several diagnostic and treatment steps along the way. Second, the lack of high quality evidence focusing on the efficacy of diagnostic tests on long-term patient important outcome measures. This is in part related to lack of regulation requiring this type of evidence before these tests are available on the market. Third, many decision makers including guideline developers, policy makers and healthcare providers are not as aware of appropriate research designs and methodological limitations for evaluating diagnostic tests in comparison to treatment based investigations. Fourth, lack of clear guidance as to which outcome measures are most relevant in specific clinical situations. Fifth, even amongst well-designed studies, results are often not very generalizable as the spectrum of participants in diagnostic studies varied much more than that in treatment trials. Sixth, technology change rapidly and new diagnostic tests evolve quickly. Finally, most of the new tests offer real advantages
in addition to improved accuracy like being less invasive but they can be very costly which require careful assessment of all trade-offs.

We believe the lack of direct and high quality evidence about HCTDS should not lead to more lenient standards when assessing it. Also, this should not be used as an excuse for suboptimal decisions and recommendations about use of tests and strategies. All these reasons highlight the urgent need for clear guidance on when to make a confident decision for or against the use of a specific test and when we require additional evidence to support this type of decision.

In this chapter I provide a summary of the products that were informed by the work done in this thesis including detailed recommendations about steps to follow when making decisions about HCTDS. I also provide a summary of future direction for this research program.
PRODUCTS INFORMED BY THIS THESIS

1. **Recommendations and a framework that will facilitate making coverage, and other, decisions about HCTDS**

Based on the findings of this thesis we submitted a full report to the German Insurance Fund Agency outlining our recommendations of the steps to follow when making coverage decisions about HCTDS. We believe our recommendations and framework will facilitate making coverage, and other, decisions about HCTDS. This framework will support making judgments about when more research is required (evidence development) and when a test should or could be introduced into practice. The judgments that are involved should always be made transparent but their detail may differ depending on the available evidence. Assumptions about the evidence should also be transparent. These recommendations are:

**Recommendation 1: determine if the test is a “new” test, for the sake of recommendations and coverage decisions**

Based on our findings in chapter 1-6, a “new” test, for the sake of recommendations and coverage decisions, fulfills one of the following conditions:

- A new diagnostic test is developed
- A new technology for an existing test is being used
- An existing test is being used with a new cut-off or threshold
- An existing test is being used for a new purpose (diagnosing a different condition)
- An existing test is being used for a new application (screening or surveillance,
risk assessment and classification, diagnosis (ruling in), ruling out diagnosis, treatment triage, treatment monitoring, staging and determining prognosis).

- An existing test is being used for a new role (Tests are also used in different roles in the management pathways: triage, add-on, replacement, parallel testing and replacement of a reference test)
- An existing test is being used for the same purpose in a new population
- An existing test is being used with high quality evidence about its efficacy on patient

**Recommendation 2: Develop a care pathway that considers all factors relevant for decision-making about the test(s) and strategy(ies).**

Whether coverage decision, regulatory approval or health care recommendations, evaluation of a “new” test should focus on its integration in the management strategy, i.e. its intended use. We suggest using decision trees and examples as starting point to outlining the potential care pathways.
Recommendation 3: determine if moderate or high quality evidence for HCTDS with direct evaluation of patient important outcomes is available

- To make these decisions we suggest using the following framework

Step 1:

Study designs I

Are there studies that directly focus on: mortality, morbidity, symptoms, and/or quality of life?

Step 2:

Apply GRADE approach as for treatment or other intervention

No

Yes


Study designs II

Look for diagnostic test accuracy studies

And then make inferences from other evidence

**Recommendation 4: when relying on test accuracy evidence consider the critical information for making decisions about HCTDS**

<table>
<thead>
<tr>
<th>Domains</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factors related to test accuracy</td>
<td>Test accuracy results</td>
</tr>
<tr>
<td></td>
<td>Pre-test probability, prevalence, baseline risk</td>
</tr>
<tr>
<td></td>
<td>Frequency and effect of inconclusive results</td>
</tr>
<tr>
<td>Factors related to clinical decision making</td>
<td>Alternatives in the management pathway</td>
</tr>
<tr>
<td></td>
<td>Application of the test</td>
</tr>
<tr>
<td></td>
<td>Frequency and effect of incidental findings</td>
</tr>
<tr>
<td>Overall benefits and harms</td>
<td>The link between the test results and patient important outcomes</td>
</tr>
<tr>
<td></td>
<td>Additional benefits of the test</td>
</tr>
<tr>
<td></td>
<td>Direct harms and safety considerations</td>
</tr>
<tr>
<td>Values and preferences</td>
<td>Patient and society values and preferences</td>
</tr>
<tr>
<td></td>
<td>Physician values and preferences</td>
</tr>
<tr>
<td>Resource implications</td>
<td>Direct resource requirements due to testing (unit cost and per population)</td>
</tr>
<tr>
<td></td>
<td>Indirect resource requirements (treatment cost, cost of inconclusive results)</td>
</tr>
<tr>
<td>Quality of evidence (certainty about the effect estimate)</td>
<td>Certainty about test accuracy results</td>
</tr>
<tr>
<td></td>
<td>Certainty about linked evidence</td>
</tr>
<tr>
<td></td>
<td>Certainty about other factors like resource implications and values and preferences.</td>
</tr>
<tr>
<td></td>
<td>Overall certainty about the effect of the test on patient outcome</td>
</tr>
<tr>
<td>Additional factors</td>
<td>Ethical and legal considerations</td>
</tr>
<tr>
<td></td>
<td>Feasibility, applicability and organizational considerations</td>
</tr>
</tbody>
</table>

**Recommendation 5: compile all relevant evidence (and identify gaps in this evidence) and complete an evidence to decision framework**
Should _______ be used instead of ___________ to diagnose and manage ___________? (This is a draft version– updated versions will be developed by the GRADE working group as part of the DECIDE project, www.decide-collaboration.eu)

Patients:

Diagnostic intervention:

Comparison:

Implied purpose:

Linked treatment(s):

Anticipated outcomes (prevented and caused by testing and subsequent management if applicable):

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>JUDGMENTS</th>
<th>RESEARCH EVIDENCE</th>
<th>ADDITIONAL CONSIDERATIONS/EXPLANATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>How common is the problem?</td>
<td>No  Probably  Uncertain  Probably  Yes</td>
<td>Varies</td>
<td>Provide information about the importance of the problem for the setting</td>
</tr>
<tr>
<td>Is the problem severe?</td>
<td>No  Yes</td>
<td></td>
<td>Describe if frequency and severity have an impact on considering the overall recommendation.</td>
</tr>
<tr>
<td>May skip for individual patient perspective</td>
<td>No  Probably  Uncertain  Probably  Yes</td>
<td>Varies</td>
<td></td>
</tr>
<tr>
<td>What is the diagnostic test accuracy?</td>
<td>Very inaccurate  Inaccurate  Uncertain  Accurate  Very accurate</td>
<td></td>
<td>Describe the diagnostic test accuracy (i.e. sensitivity and specificity) and if it is sufficient to continue developing a recommendation.</td>
</tr>
<tr>
<td>What is the overall confidence in the diagnostic test accuracy information?</td>
<td>Very low  Low  Moderate  High</td>
<td></td>
<td>Also see full diagnostic test accuracy evidence profile</td>
</tr>
</tbody>
</table>
### BENEFITS AND HARS

#### How important are these outcomes?

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Relative importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Outcome]</td>
<td>-</td>
</tr>
</tbody>
</table>

#### Overall, compared to the alternative, are the anticipated benefits large?

| Outcome | - |

#### Overall, compared to the alternative, are the anticipated harms small?

| Outcome | - |

#### What is the balance of the benefits and harms/burden?

- □ Benefits outweigh harms/burden
- □ Benefits slightly outweigh harms/burden
- □ Benefits and harms/burden are balanced
- □ Harms/ burden slightly outweigh benefits
- □ Harms/ burden outweigh benefits

#### The relative importance or values of the main outcomes of interest (pick 5):**

<table>
<thead>
<tr>
<th>Critical Outcomes:</th>
<th>Large benefit</th>
<th>Small benefit</th>
<th>No effect</th>
<th>Small harm/ burden</th>
<th>Modest harm/ burden</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ...</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>2. ...</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>3. ...</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>4. ...</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>5. ...</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

**Also see full third layer Summary of Findings**

Describe narratively in the “Details of Judgment column” or rate the benefits and harms by considering:

- Benefits outweigh harms/burden
- Benefits slightly outweigh harms/burden
- Benefits and harms/burden are balanced
- Harms/ burden slightly outweigh benefits
- Harms/ burden outweigh benefits

**Describe narratively in the “Details of Judgment column” or rate the benefits and harms by considering:**

The focus is on patient important outcomes that are calculated on the basis of the pretest probability, diagnostic test accuracy, presumed natural history of disease, anticipated frequency of outcomes related to the disease and to the direct effects of the test, treatment efficacy and complications and reported as patient or population outcomes.

On average, how important are the outcomes to patients?
Overall, is the certainty about the link between the diagnostic test accuracy information and the linked benefits and harms?

<table>
<thead>
<tr>
<th>Very uncertain</th>
<th>Uncertain</th>
<th>Moderately certain</th>
<th>Certain</th>
<th>Very certain</th>
</tr>
</thead>
</table>

* The evidence about the link between the diagnostic test accuracy and benefits and stems mainly from very low quality evidence for treatment effects and natural progression/history data.

Also see full third layer Summary of Findings

What is the overall confidence in the estimates of effect for benefits and harms?

<table>
<thead>
<tr>
<th>Very low</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
</table>

This certainty is high if there is moderate or high quality evidence indicating that treatment has clear consequences for patient important outcomes. What are the underlying values and preferences for the outcomes associated with the test and the problem is our confidence in these values and preferences?

Is there similarity about how much people value the main outcomes?

<table>
<thead>
<tr>
<th>Similar</th>
<th>Probably similar</th>
<th>Uncertain</th>
<th>Probably not similar</th>
<th>Not similar</th>
</tr>
</thead>
</table>

Source of variability if any:

Are the resources required small? (may skip for individual patient perspective)

<table>
<thead>
<tr>
<th>No</th>
<th>Probably not</th>
<th>Uncertain</th>
<th>Probably yes</th>
<th>Yes</th>
</tr>
</thead>
</table>

What are the costs per resource unit? Opportunity cost: Is this intervention and its effects worth withdrawing or not allocating resources from other interventions? Differences across settings: Is there lots of variability in resource requirements across settings?
<table>
<thead>
<tr>
<th>Is the incremental cost (or resource use) small relative to the benefits?</th>
<th>No</th>
<th>Probably</th>
<th>Uncertain</th>
<th>Probably</th>
<th>Yes</th>
<th>Varies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the cost (including out of pocket) worth the benefits?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What happens to health inequities?</th>
<th>Increased</th>
<th>Probably</th>
<th>Uncertain</th>
<th>Probably</th>
<th>Reduced</th>
<th>Varies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>increased</td>
<td>reduced</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would the implementation of the intervention reduce inequities?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Is the option acceptable to key stakeholders?</th>
<th>No</th>
<th>Probably</th>
<th>Uncertain</th>
<th>Probably</th>
<th>Yes</th>
<th>Varies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCEPTABILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Is the option feasible to implement?*</th>
<th>No</th>
<th>Probably</th>
<th>Uncertain</th>
<th>Probably</th>
<th>Yes</th>
<th>Varies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FEASIBILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Is this intervention generally available? Can it be implemented?</th>
<th>None</th>
</tr>
</thead>
</table>
## Recommendation

**Should heparins be used instead of no heparins for patients with advanced cancer?**

<table>
<thead>
<tr>
<th>Overall balance of consequences</th>
<th>Undesirable consequences clearly outweigh desirable consequences</th>
<th>Undesirable consequences probably outweigh desirable consequences</th>
<th>The balance between desirable and undesirable consequences is too uncertain*</th>
<th>The balance of desirable and undesirable consequences indicates they are very similar*</th>
<th>Desirable consequences probably outweigh undesirable consequences</th>
<th>Desirable consequences clearly outweigh undesirable consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Should heparins be used instead of no heparins for patients with advanced cancer?</td>
<td>We recommend against the option or for the alternative</td>
<td>We suggest not to use the option or to use the alternative</td>
<td>No recommendation</td>
<td>We suggest using the option</td>
<td>We recommend the option</td>
<td></td>
</tr>
<tr>
<td></td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

### Panel decisions

*Describe decision making process if relevant*

**Recommendation (text)**

*Formulate clear recommendation*

**Remarks and justification**

*Explain the rationale and provide important disclaimers and remarks*

**Implementation considerations**

*Describe issues relevant for implementation*

**Research priorities**

*Describe research priorities*

---

* In this situation no recommendation could be reasonable
Recommendation 6: Before coverage decisions are made the following algorithm should be used.

Diagnostic PICO (requires development of a complete diagnostic framework)

- Is there high or moderate quality of evidence? (focus on studies evaluating patient important outcomes)
  - Yes
    - Develop coverage decision using the evidence to decision framework (ETR) for coverage decision makers about diagnostic tests
      - This includes making decisions about
    - No
      - For the new test, or a test directly related to the new test, develop a quick ETR framework looking for:
        1. Diagnostic test accuracy evidence
        2. Linked evidence about what would be the consequences for true positive, true negative, false positive, false negative, cost and other relevant outcomes
  - No, the quality of diagnostic test accuracy evidence is low or very low
    - Do diagnostic test accuracy or patient important outcomes research
  - Yes, quality of diagnostic test accuracy evidence is high or moderate, make decision
    - Research on patient important outcomes
  - Is there high or moderate confidence in, or quality of, linked evidence?
    - Yes
      - Is the quality of linked evidence very low?
        - Yes new evidence is required before coverage is offered
          - Is research focusing on patient important outcomes possible?
            - No
              - Do research on linked evidence
            - Yes
              - Research on patient important outcomes
    - No
      - Do diagnostic test accuracy or patient important outcomes research

2. **Electronic Interactive Diagnostic Summary of Findings tables**

In chapter three of this thesis we presented the results of the extensive user testing that led to the development of the current formats of the diagnostic evidence tables (evidence profiles (EvP) and summary of findings tables (SoF)). These formats have been further developed into electronic interactive tables to suit the needs of different users by addressing the limitations of the current formats. We introduced a prototype of the interactive diagnostic SoF tables during the Developing and Evaluating Communication strategies to support Informed Decisions and practice based on Evidence (DECIDE) international conference held in Edinburg, Scotland in June 2014. We collected feedback and user tested the tables. The interactive diagnostic SoF tables will be further tested and during the Guideline International Network meeting in Melbourne, Australia in August 2014. The prototype can be accessed on this website [http://isodep.isofdx.epistemonikos.org](http://isodep.isofdx.epistemonikos.org).

These interactive tables allow the users to navigate through different entry points each of which presents different formats of the results of test accuracy systematic reviews. These interactive tables also allow users to expand or minimize the tables to allow as much or as little details as they choose for their specific needs. The tables also allow for including definitions of concepts and headings without crowding the table with information.
FUTURE DIRECTION

We will use an exploratory sequential mixed methods design to validate our findings and conclusions. We will survey an international cohort of guideline developers, policy makers and clinicians that have to make and have made decisions about different HCTDS to validate our findings about the factors to be considered for decision-making about HCTDS. We will also validate our recommendations and the framework suggested for making decisions about HCTDS. Additionally, we will validate the framework addressing when test accuracy results are sufficient and when they are not (from chapter 6).

The interactive diagnostic tables will be further tested by a variety of users including healthcare providers, systematic reviewers, policy makers, and guideline developers. We will evaluate the diagnostic evidence tables (both interactive and non-interactive) and their impact on understanding the results and making decisions in regards to the test(s) through a randomized trial.