Evaluation of diagnostic tests when there is no gold standard

Evaluation of diagnostic tests when there is no gold standard. A review of methods

AWS Rutjes,[1] JB Reitsma,[1] A Coomarasamy,[2]
KS Khan[2*] and PMM Bossuyt[1]

[1] Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, University of Amsterdam, The Netherlands
[2] Division of Reproductive and Child Health, University of Birmingham, UK

* Corresponding author

Executive summary

Health Technology Assessment 2007; Vol. 11: No. 50

Health Technology Assessment
NHS R&D HTA Programme
www.hta.ac.uk

Executive summary

Background

The classical diagnostic accuracy paradigm is based on studies that compare the results of the test under evaluation (index test) with the results of the reference standard, the best available method to determine the presence or absence of the condition or disease of interest. Accuracy measures express how the results of the test under evaluation agree with the outcome of the reference standard. Determining accuracy is a key step in the health technology assessment of medical tests.

Researchers evaluating the diagnostic accuracy of a test often encounter situations where the reference standard is not available in all patients, where the reference standard is imperfect or where there is no accepted reference standard. We use the term ‘no gold standard situations’ to refer to all those situations. Several solutions have been proposed in these circumstances. Most articles dealing with imperfect or absent reference standards focus on one type of solution and discuss the strengths and limitations of that approach. Few authors have compared different approaches or provided guidelines on how to proceed when faced with an imperfect reference standard.

Objectives

We systematically searched the literature for methods that have been proposed and/or applied in situations without a ‘gold’ standard, that is, a reference standard that is without error. Our project had the following aims:

1. To generate an overview and classification of methods that have been proposed to evaluate medical tests when there is no gold standard.
2. To describe the main methods discussing rationale, assumptions, strengths and weaknesses.
3. To describe and explain examples from the literature that applied one or more of the methods in our overview.
4. To provide general guidance to researchers facing research situations where there is no gold standard.

Methods

We employed multiple search strategies to obtain an overview of the different methods described in the literature, including searches of electronic databases, contacting experts for papers in personal archives, exploring databases from previous methodological projects (STARD and QUADAS) and cross-checking of reference lists of useful papers already identified.

We developed a classification for the methods identified through our review taking into account the degree to which they represented a departure away from the classical diagnostic accuracy paradigm.

For each method in our overview, we prepared a structured summary based on all or the most informative papers describing its rationale, its strengths and weaknesses, its field of application, available software and illustrative examples of the method.

Based on the findings of our review, discussions about the pros and cons of different methods in various situations within the research team and input from expert peer reviewers, we constructed a flowchart providing general guidance to researchers faced with evaluation of tests without a gold standard.

Results

From 2200 references initially checked for their usefulness, we ultimately included 189 relevant articles that were subsequently used to classify and summarise all methods into four main groups, as follows.

Impute or adjust for missing data on reference standard

In this group of methods, there is an acceptable reference standard, but for various reasons the outcome of the reference standard is not obtained in all patients. Methods in this group either impute or adjust for this missing information in the subset of patients without reference standard outcome. Researchers should be careful with these methods if (1) the pattern of missing values is not determined by the study design, but is influenced by the choice of patients and physicians, or (2) the fraction of patients verified with the reference standard is small within results of the index tests.

Correct imperfect reference standard

In this group, there is a preferred reference standard, but this standard is known to be imperfect. Solutions from this group either adjust estimates of accuracy or perform sensitivity analysis to examine the impact of this imperfect reference standard. The adjustment is based on external data (previous research) about the degree of imperfection. Correction methods can be useful if there is reliable information about the degree of imperfection of the reference standard and about the correlation of the errors between the index test and the reference standard.

Construct reference standard

These methods have in common that they combine multiple test results to construct a reference standard outcome. Groups of patients receive either different tests (differential verification and discrepant analysis) or the same set of tests, after which these results are combined by: (1) deterministic predefined rule (composite reference standard); (2) consensus procedure among experts (panel diagnosis); (3) a statistical model based on actual data (latent class analysis). The prespecified rule for target condition makes the composite reference standard method transparent and easy to use, but misclassification of patients is likely to remain. Discrepant analysis should not be considered in general, as the method is likely to produce biased results. The drawback of latent class models is that the target condition is not defined in a clinical way, so there can be lack of clarity about what the results stand for in practice. Panel diagnosis also combines multiple pieces of information, but experts may combine these items in a manner that more closely reflects their own personal concept of the target condition.

Validate index test results

The diagnostic test accuracy paradigm is abandoned in this group and index test results are related to relevant other clinical characteristics. An important category is relating index test results with future clinical events, such as the number of events in those tested negative for the index test results. Test results can also be used in a randomised study to see whether the test can predict who will benefit more from one intervention than the other. Because the classical accuracy paradigm is not employed, measures other than accuracy measures are calculated, including event rates, relative risks and other correlation statistics.

Conclusions

The majority of methods try to impute, adjust or construct a reference standard in an effort to obtain the familiar diagnostic accuracy statistics such as pairs of sensitivity and specificity or likelihood ratios. In situations that deviate only marginally from the classical diagnostic accuracy paradigm, for example where there are few missing values on an otherwise acceptable reference standard or where the magnitude and type of imperfection in a reference standard is well documented, these are valuable methods. However, in situations where an acceptable reference standard does not exist, holding on to the accuracy paradigm is less fruitful. In these situations, applying the concept of clinical test validation can provide a significant methodological advance. Validating a test means that scientists and practitioners examine, using a number of different methods, whether the results of an index test are meaningful in practice. Validation will always be a gradual process. It will involve the scientific and clinical community defining a threshold, a point in the validation process, whereby the information gathered would be considered sufficient to allow clinical use of the test with confidence.

Recommendations for further research

All methods summarised in this report need further development. Some methods, such as the construction of a reference standard using panel consensus methods and validation of tests outwith the accuracy paradigm, are particularly promising but are lacking in methodological research. These methods deserve particular attention in future research.

Publication

Rutjes AWS, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PMM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technol Assess* 2007;**11**(50).

NIHR Health Technology Assessment Programme

The Health Technology Assessment (HTA) programme, now part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the costs, effectiveness and broader impact of health technologies for those who use, manage and provide care in the NHS. 'Health technologies' are broadly defined to include all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care, rather than settings of care.

The research findings from the HTA Programme directly influence decision-making bodies such as the National Institute for Health and Clinical Excellence (NICE) and the National Screening Committee (NSC). HTA findings also help to improve the quality of clinical practice in the NHS indirectly in that they form a key component of the 'National Knowledge Service'.

The HTA Programme is needs-led in that it fills gaps in the evidence needed by the NHS. There are three routes to the start of projects.

First is the commissioned route. Suggestions for research are actively sought from people working in the NHS, the public and consumer groups and professional bodies such as royal colleges and NHS trusts. These suggestions are carefully prioritised by panels of independent experts (including NHS service users). The HTA Programme then commissions the research by competitive tender.

Secondly, the HTA Programme provides grants for clinical trials for researchers who identify research questions. These are assessed for importance to patients and the NHS, and scientific rigour.

Thirdly, through its Technology Assessment Report (TAR) call-off contract, the HTA Programme commissions bespoke reports, principally for NICE, but also for other policy-makers. TARs bring together evidence on the value of specific technologies.

Some HTA research projects, including TARs, may take only months, others need several years. They can cost from as little as £40,000 to over £1 million, and may involve synthesising existing evidence, undertaking a trial, or other research collecting new data to answer a research problem.

The final reports from HTA projects are peer-reviewed by a number of independent expert referees before publication in the widely read monograph series *Health Technology Assessment*.

Criteria for inclusion in the HTA monograph series

Reports are published in the HTA monograph series if (1) they have resulted from work for the HTA Programme, and (2) they are of a sufficiently high scientific quality as assessed by the referees and editors.

Reviews in *Health Technology Assessment* are termed 'systematic' when the account of the search, appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

The research reported in this monograph was commissioned by the National Coordinating Centre for Research Methodology (NCCRM), and was formerly transferred to the HTA Programme in April 2007 under the newly established NIHR Methodology Panel. The HTA Programme project number is 06/90/23. The contractual start date was in October 2005. The draft report began editorial review in March 2007 and was accepted for publication in April 2007. The commissioning brief was devised by the NCCRM who specified the research question and study design. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the referees for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

The views expressed in this publication are those of the authors and not necessarily those of the HTA Programme or the Department of Health.

ISSN 1366-5278

Published by Gray Publishing, Tunbridge Wells, Kent, on behalf of NCCHTA.
Printed on acid-free paper in the UK by St Edmundsbury Press Ltd, Bury St Edmunds, Suffolk.