Using randomised controlled trials to evaluate the clinical effectiveness of diagnostic tests:

How useful are test-treatment RCTs?

By

Lavinia Ferrante di Ruffano

A thesis submitted to the
University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

Department of Public Health, Epidemiology and Biostatistics
School of Health and Population Sciences
College of Medicine and Dentistry
University of Birmingham
November 2012

Abstract

Background: Decisions on which tests to use should be informed by evidence that they do more good than harm. Test-treatment RCTs are recommended as the ‘gold–standard’ approach, but have attracted criticism that question whether they are fit for purpose. Confronting this question, the thesis investigates four key challenges by finding and analysing all identifiable test-treatment RCTs (2004–2007).

Methods: Capture–recapture analysis estimated the total population of trials; descriptive analysis characterised the diagnostic questions evaluated by RCT; reviews of reporting and methodological quality investigated how informative and valid trials are; analytic induction was used to develop a theoretical framework linking tests to health outcomes, from which a tool was designed.

Results: Published trials were poor quality, and found to be highly complex studies that will be challenging to evaluate reliably: interventions are difficult to capture and translate into protocols; several methods traditionally used to eliminate bias are more difficult to implement; test-treatment strategies impact on patient health in numerous and highly complicated ways.

Conclusion: Test-treatment trials have the potential to be very useful instruments, and though highly challenging they could be both reliable and informative. However, it must be acknowledged that trials will not be suited to all comparisons.

~ For Tony and Tom ~

"To strive, to seek, to find, and not to yield."

Alfred Lord Tennyson, Ulysses, 1842

Acknowledgements

The research contained within this thesis represents something of an academic journey that has taken me across disciplines, faculties and cities. There is no doubt that I would not have succeeded in beginning this journey were it not for Tony Waldron, supervisor and mentor in my previous life as an archaeologist. It is thanks to Tony that I developed a fascination for diagnosis and epidemiology, in the living as well as the dead. It is to Tony that I dedicate this thesis, in gratitude to the years of encouragement, fascinating discussion and of course the use of his bone lab.

I also dedicate this thesis to my darling husband Tom, who graces me every day with unending love, encouragement, and hot dinners. Though he completed his thesis some years ago, I still hope to be able to repay him the months of motivation and housework with many years of married happiness (and possibly some housework too)!

I would very much like to thank Jon Deeks and Chris Hyde, my supervisors, for taking a chance and hiring an archaeologist, and most of all for their guidance, inspiration and everything they have taught me in the last five years.

I am also indebted to several other colleagues for their advice and assistance in performing systematic searches (Anne Fry–Smith, Sue Bayliss and Anne Eisinga), for agreeing to second–screen and duplicate data extraction (Clare Davenport and Jac Dinnes), and for their valuable comments on draft chapters (Richard Riley) or co–authored papers related to my thesis (Patrick Bossuyt and Kirsten McCaffery).

Finally, to my family and friends: bless you for keeping me happy and sane, especially mum, Pat and Graham, Sophie, Lou, Lucy and Ruth.

Table of Contents

List of Tables

CHAPTER 1.

CHAPTER 2.

CHAPTER 3.

CHAPTER 4.

CHAPTER 5.

CHAPTER 6.

CHAPTER 7.

CHAPTER 8.

List of Figures

CHAPTER 1.

CHAPTER 2.

CHAPTER 3.

CHAPTER 4.

CHAPTER 5.

CHAPTER 6.

CHAPTER 7.

CHAPTER 8.

List of Boxes

CHAPTER 5.

CHAPTER 6.

CHAPTER 7.

Dissemination of research

During the course of completing this thesis, elements of its findings have been disseminated to the wider research community. Whilst the author is the primary contributor to the research, analysis and authorship of these pieces, thesis supervisors and other colleagues have also contributed for the purposes of publication.

Preliminary results to Chapters 5 and 6 were presented as oral and poster presentations at two international conferences[1–2].

Chapter 3 contributed to a peer–reviewed article[3] and two conference poster presentations[4–5].

Chapter 6 was presented as an oral publication[6] and two poster presentations, the latter winning the Thomas Chalmers prize at the 2010 Cochrane Colloquium with a consequent invitation for publication in the Cochrane Methodology Newsletter[7], as well as an invitation for a short oral presentation for best research show–case at the 2011 EVIDENCE conference[8].

More recently sections of chapters 7 and 8 have been published as a peer–reviewed article[9].

List of publications

1. Ferrante di Ruffano L, Dinnes J, Hyde C, Deeks J. A review of the use of randomised trials to assess the impact of diagnostic tests on patient outcomes [abstract]. Oral presentation at the *Methods for Evaluating Medical Tests* Symposium; 2008 July 25-26; University of Birmingham, UK.
2. Ferrante di Ruffano L, Dinnes JJ, Hyde CH, Deeks JJ. Assessing the effects of diagnostic tests on patient outcomes: How reliable, informative and practical are randomised controlled trials [abstract]? Poster presentation at the *Cochrane Collaboration Colloquium*; 2008 Oct 3-7; Freiburg, Germany. Cochrane Database of Systematic Reviews, Supplement 2008.

3. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks J. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epi 2012;65:282–287.

4. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks J. Randomised trials that evaluate tests are rare and difficult to find: Confirmation from a capture-recapture study [abstract]. Poster presentation at the *Methods for Evaluating Medical Tests and Biomarkers* Symposium; 2010 July 1-2; University of Birmingham, UK.

5. Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks J. Randomised trials that evaluate tests are rare and difficult to find: Confirmation from a capture-recapture study [abstract]. Poster presentation at the *Joint Cochrane and Campbell Collaboration Colloquia*; 2010 Oct 18-22; Keystone, Colorado, USA. Cochrane Database of Systematic Reviews, Supplement 2010; Suppl CD000002:129.

6. Ferrante di Ruffano L, Hyde C, Deeks J. How do tests impact on patient health? An explanatory framework [abstract]. Oral presentation at the *Methods for Evaluating Medical Tests and Biomarkers* Symposium; 2010 July 1-2; University of Birmingham, UK.

7. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. How do tests impact on patient health? Development of an explanatory framework. Cochrane Methods Newsletter. Cochrane Database of Systematic Reviews 2011;Suppl 1**:** 1–40.

8. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. How do tests change patient health? A framework for evaluating evidence [abstract]. Poster presentation at EVIDENCE 2011; 2011 October 24–26; London, UK.

9. Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests – A framework for designing and evaluating trials. BMJ 2012:344:e686.

1

Introduction

Evidence-based diagnosis

Will introducing a new diagnostic test benefit patients? In the era of evidence–based medicine, decisions on which tests to use should be informed by rigorous evidence that the selected interventions do patients more good than harm[1]. In contrast to the wealth of rigorous research into treatment effectiveness produced during the last three decades, comparatively little progress has been made toward ensuring that decisions on which diagnostic tests to use is similarly based on evidence of clinical effectiveness.

This is perhaps surprising given that diagnoses, informed by the results of diagnostic tests, determine which patients receive treatment. In a world where an increasing number of tests are becoming available, and financial constraints on healthcare budgets are unlikely to ease, the ability to consult high–quality evidence that demonstrates which tests will be most beneficial to the health of patients offers a key resource to both clinicians and policy makers. This is particularly critical when one considers the accumulating evidence that diagnostic test use is increasing at a very fast rate, across all quarters of the clinical spectrum[2–5]. And yet for many of these tests there is no evidence that their use leads to any improvement in the health of patients, or in cost–effectiveness[6–9]. Without this information we risk using tests that are 'inferior' to others, with grave consequences for both patients and resource–use. In the short–term fewer patients may receive prompt and appropriate treatment, with implications for longer–term differentials in patient health, including early death and severe morbidity. Inappropriate testing can also promulgate unnecessary further testing, increasing both the direct harms and costs of healthcare for no appreciable gain in patient health[5,10–11].

Recognising the importance of these issues, evidence–based diagnosis has experienced a significant investment in research activity over the last 25 years, albeit focussed on developing and promoting methods for evaluating diagnostic accuracy[12–13]. However better accuracy may not benefit patients unless it leads to changes in diagnoses and patient management[14]. In order to fully evaluate the consequences of testing, it is necessary to measure the impact that competing tests have on both intermediate processes and downstream patient health.

The test-treatment randomised controlled trial (RCT) is widely heralded as the 'gold–standard' study design for achieving this[14–18]. These designs randomise patients between testing strategies, follow participants up through their subsequent management, and measure outcomes only after treatment has been received[19]. Yet the complexity of performing these evaluations has roused concerns amongst the methodological community that these RCTs invite unique methodological challenges that may threaten how useful they really are[6,8,20–22]. However for now these concerns remain hypothetical since to date very little research has been conducted to verify their presence or extent in published test-treatment trials.

RCTs have been widely used to assess the effectiveness of screening programmes[23], however their application to diagnostic settings does not yet appear to be so common[24]. There are indications, however, that calls for these studies appear to be increasing: in 2010 the UK's National Institute of Health and Clinical Excellence (NICE) launched the Diagnostics Assessment Programme, which aims to provide evidence–based guidelines on diagnostic tests and to date has published 3 guidelines with a further nine in development[25]; only one year earlier the National Institute for Health Research Health Technology Assessment programme (NIHR HTA) released a commissioned call for research evaluating the impact of diagnostic tests on patient health and clinical management. Over the last 10 years numerous test-treatment RCTs have been funded by UK grant–awarding bodies, including the RATPAC trial of point–of–care testing for suspected acute myocardial infarction patients[26], the MRC–CUBE trial of *h.pylori* testing in dyspeptic patients[27] and trials evaluating the benefits of X–ray in lower back pain sufferers[28–29]. As the focus of diagnostic research broadens and the requirements for evidence of clinical effectiveness increase, it becomes ever more essential to investigate whether the RCT can in fact be relied upon as the gold–standard method to produce this evidence–base.

This thesis aims to investigate how useful the RCT is for assessing the patient health impact of diagnostic tests. The rest of this chapter provides a background to the area of diagnostic test research, defines the scope of this thesis, and sets out its aims and objectives. The first part positions the thesis within the analytic infrastructure of diagnostic research, introducing the reader to the phases of test evaluation and the specifics of test-treatment RCT designs. The second part reviews what is currently known regarding the methodological and reporting quality of RCTs in general. The third part looks at four of the key challenges that have been levelled at the test-treatment RCT, and summarises the findings of existing research into these issues. The final section sets out the aims of the thesis in detail, providing an orientation of how subsequent chapters evaluate each of these goals.

1.1 Evidence−based diagnosis

1.1.1 Defining diagnostic tests

Medical testing forms the basis of decision-making for clinical intervention. It describes the process of information-gathering to determine the presence, nature and future course of disease in patients, from which the most appropriate course of management is planned. This definition is necessarily expansive, since tests are used for a wide variety of purposes.

Screening tests are used for the early detection of disease in individuals who have yet to manifest any symptoms or signs of illness, and so are often performed in large cohorts of the asymptomatic population as blanket screening programmes[23]. In the UK examples include mammography screening for breast cancer in females[30] and blood spot screening for a range of potentially serious congenital conditions in newborns[31].

Unlike screening tests that are initiated by the healthcare provider, diagnostic tests are used in patients presenting with suspicious signs and symptoms to determine the likely cause of their problem. For example, individuals who arrive at their GP with complaints of dysuria and urinary urgency may be given a ‘dipstick test’ to test for the presence of micro-organisms in their urine, and so determine whether a bacterial infection of the urinary tract is the cause of their discomfort[32]. In secondary care, patients referred to orthopaedic consultants because of knee pain may be given an imaging scan, such as magnetic resonance imaging (MRI), to establish whether their symptoms are caused by any internal derangement of the knee[33]. Any process that is used to formulate a diagnostic decision in symptomatic patients may be considered a ‘diagnostic test’. This definition encompasses the directly identifiable ‘diagnostic technologies’ (as defined by the NICE technology appraisal process), such as radiography, serology or electrocardiography, but more broadly also defines any sort of information used to confirm or rule out the presence of disease, including physical examination and patient history.

Tests are also used to classify the severity (or stage) of known disease, for example tertiary care patients with an existing diagnosis of non–small–cell lung carcinoma may be given computed tomography (CT) scans, MRI scans, and positron emission tomography (PET) scans to determine the location and size of tumours[34]. Diagnostic staging is closely related to prognostic testing, where the purpose of testing is to predict the course of disease or the risk of adverse events in the future[35], or even the likely response to treatment[36]. Indeed in many situations tests may be used for both diagnostic and prognostic purposes, as with the cancer staging example above.

Lastly, tests can be used to monitor disease states, such as the serial testing used in surveillance for recurrence of bladder cancer[37], or to titrate treatment, as in the daily measurement of blood glucose levels to adjust insulin doses in diabetic individuals[38].

1.1.2 Scope of the thesis

This thesis will examine the use of RCTs to evaluate the impacts of testing on patient health, with a focus on diagnostic tests. Trials of monitoring or prognostic tests present the investigator with a different range of study design issues. Monitoring situations require study designs to take repeated testing into account, and may often not need to evaluate subsequent treatment[39]. Evaluations of prognostic tests aim to compare the accuracy of predictions for the risk of future events, and may also not need to evaluate subsequent treatment[40].

Trials of screening tests also fall outside the scope of this thesis, since the role of the RCT is already established to be well–suited to evaluating the health consequences of screening[23]. RCTs are the only design that can evaluate the health risks of overdetection (treating individuals whose preclinical disease would not have progressed), the most important measure of patient harm resulting from a screening programme[41]. Indeed many screening policies are based on RCT evidence and the initiation of new programmes must show that the benefits of screening "*outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatment)*" [42]. Arguably, screening trials must also consider several design issues that are unlikely to affect diagnostic RCTs, such as overdetection bias and lead–time bias (individuals in whom disease is detected through screening appear to survive longer when in fact there is no difference)[43].

Conversely, the use of RCTs to evaluate the effectiveness of diagnostic tests has received little attention and is not yet supported by a rigorous methodological understanding. The following section places test-treatment RCTs within the broader context of diagnostic test evaluation, and describes the theoretical design of these studies.

1.1.3 Phases of test evaluation

Over the last three decades methodologists have proposed that the ideal evaluation of diagnostic tests requires several phases of assessment in order to arrive at decisive conclusions regarding their clinical effectiveness. These research frameworks were developed as a diagnostic alternative to the now commonly accepted phased evaluation of pharmaceuticals. They therefore sought to develop a standardised approach for assessing diagnostic tests from their technological conception through to evaluating their clinical effectiveness in routine medical practice[14,44–56].

Thirty–one research frameworks were published between 1978 and 2007, proposing 19 different evaluative pathways that should guide the assessment of new tests. These models are comprehensively reviewed by Lijmer and colleagues[57], however what is of consequence here is that these frameworks identified between four and seven necessary phases of assessment. Those with four stages tend to draw a direct parallel with the pharmaceutical evaluative pathway[55–56]. However the most commonly recognised framework is that published in the seminal paper by Dennis Fryback and John Thornbury[14]. They set out six phases of diagnostic efficacy, each aiming to evaluate a different aspect of a test's performance, which are arranged hierarchically (Figure 1.1). This structure reflects the principle, common to most permutations of the framework, that in order to be effective at any given level of assessment, a test must have demonstrated its efficacy during the preceding phase of evaluation.

Fryback and Thornbury argued that the process of evaluation should commence by assessing a test's 'technical efficacy', namely the properties of a test that reflect its ability to produce classificatory information reproducibly (precision). Once proven to be a precise

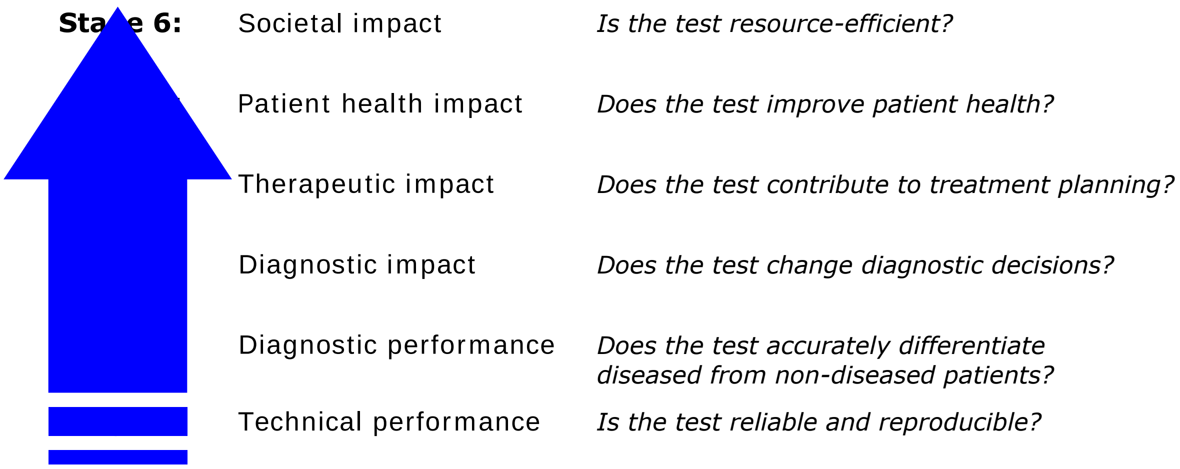

Stage 6:
Societal impact
Is the test resource-efficient?
Patient health impact
Does the test improve patient health?
Therapeutic impact
Does the test contribute to treatment planning?
Diagnostic impact
Does the test change diagnostic decisions?
Diagnostic performance
Does the test accurately differentiate diseased from non-diseased patients?
Technical performance
Is the test reliable and reproducible?

Figure 1.1: Phases of diagnostic test evaluation. Hierarchy of study designs needed to demonstrate the full effectiveness of introducing tests into clinical practice[14]. Evaluations of test accuracy are located toward the beginning of this hierarchical framework. Clinical effectiveness studies that measure health impact and cost-effectiveness comprise the final two stages of assessment.

instrument, the test's ability to correctly identify or exclude disease could be determined in studies that evaluate its 'diagnostic accuracy efficacy', now commonly referred to as diagnostic test accuracy studies[14,44–46]. Further up the hierarchy is the evaluation of diagnostic yield or 'diagnostic thinking efficacy', defined by Fryback and Thornbury[14] as measuring the extent to which diagnostic information succeeds in changing the diagnostic decisions that clinicians make. Since these changes will not necessarily lead to differences in treatment planning, tests should subsequently be evaluated with regard to the impact they have on 'therapeutic efficacy'[14,44–46]. At the summit of most of these frameworks lays the evaluation of patient outcomes, variously referred to as 'clinical outcome efficacy' or 'patient outcome efficacy', whereby the benefits and harms to patients' health as a result of testing are determined.

1.1.4 Diagnostic accuracy

Accuracy is defined as a test's ability to differentiate between diseased and non–diseased individuals[58]. Evaluating the accuracy of diagnostic tests is argued as key to determining whether they are likely to be clinically effective, since tests that are less accurate cannot hope to lead to better diagnoses, which would enable more appropriate treatments to be selected, and so improve patient health[14,59].

Three concepts are central to these studies. First, the test's performance must be measured in the detection of a single disease, referred to as the 'target condition'[60]. Second, for the purposes of subsequent analysis the target condition must be able to be dichotomised into present or absent[61]. And third, individuals classified into these two groups by the new test, or 'index test', must be compared against the 'true' situation; that is the investigator must establish, independently of the index test, which individuals are really diseased and those that are free from disease. This third requirement is approximated by using the best available diagnostic information, or 'reference standard' (generally a composite of tests), to determine the true presence or absence of disease[62].

Since several diagnostic tests already exist to detect many target conditions, the ideal approach is to evaluate the *comparative* accuracy of the index test. Paired designs involve giving study patients the index test, the existing comparator test, and the reference standard test[63]; an alternative is to randomise patients to receive the index or the comparator test, after which all patients also receive the reference standard[61]. The performance of each test relative to the reference standard can then be measured by cross–correlating the diagnoses produced by these tests. The most common statistics are the test's sensitivity, calculated as the proportion of truly diseased patients whom the index test identifies correctly, and the test's specificity, calculated as the proportion of truly disease–free individuals whom the index test identifies correctly[64].

During the last 25 years, methods for assessing and interpreting test accuracy studies have dominated research into diagnostic test evaluation[65]. This has led to significant advances in our understanding of methodological issues. For example, research has highlighted the decisive role that variability in the clinical context plays in test performance.

Estimates of test performance are shown to be affected by the demographic composition of patient groups, disease prevalence and severity, how the test is carried out and how its results are interpreted[66]. Target conditions may be detected more reliably in populations with higher disease prevalence or more severe presentations of disease, for example[67]. Variations in how the presence of disease is defined can alter the performance characteristics of tests, such as the selection of different threshold values for interpreting biochemical results or different operational criteria in the interpretation of images[68]. Numerous other factors, including the care setting, role of the new test in the existing pathway, prior investigations and practitioner experience are also known to influence diagnostic accuracy[60,63].

Methodological research has also defined the extent to which inadequate study design can bias the results of accuracy studies[69–70]. These reviews analyse large groups of published test accuracy meta–analyses by comparing studies with inadequate methods to those without the same shortcomings to determine whether effect sizes differ according to methodological quality. Results have revealed that inadequate methods and reporting are often associated with larger estimates of diagnostic accuracy, an indication of bias. For example Lijmer and colleagues found indications of bias in studies using different reference standards according to the patient's index test result (differential verification bias), performing the reference standard with knowledge of the index test result (review bias), and in studies that failed to report either the criteria used to arrive at a diagnosis or the population under study[70].

Other notable advances have been made in the field of evidence synthesis, including the development of a widely–used quality appraisal tool to assess the applicability and risk of bias in primary studies[71], the publication of a checklist to improve reporting standards[72], the design of search filters to ascertain primary studies[73], and the development of methods in the meta–analysis of data to manage the considerable heterogeneity of study populations[74].

Despite these considerable advances, there is a growing recognition that evidence of a test's diagnostic accuracy, no matter how rigorously obtained, is not in itself sufficient to recommend that the test is disseminated into clinical practice[6,44,46]. Rather, the discipline's approach to evaluating tests must broaden to also consider the impact these technologies have on downstream patient health[12,16,44,75–78].

1.1.5 Patient outcome effectiveness

Studies that assess the relative benefits of testing in patients occupy their own phase of assessment, since the hierarchical frameworks recognise that tests which appear to be efficacious in terms of their accuracy and their impact on diagnostic and therapeutic thinking may still fail to improve downstream patient health[14,17,44–46]. Two main methods have been proposed to achieve this: RCTs and decision models.

Test-treatment RCTs

The randomised controlled trial is widely proposed as the gold–standard study design for evaluating the patient health impact of diagnostic tests[15–17,19,21,44,49,51,53,79]. This position is based on the findings of extensive methodological research into treatment trials and non–randomised evaluations of treatments, which demonstrate that when conducted rigorously the RCT provides us with the most reliable tool to evaluate the comparable effectiveness of healthcare interventions[80].

When treatments are evaluated, the trial participants are randomised to receive either the new treatment or the existing treatment (or placebo), and their health response is measured after an appropriate period of follow–up. Thus not only are these designs prospective, but they ensure a direct relationship between cause (the treatment) and effect

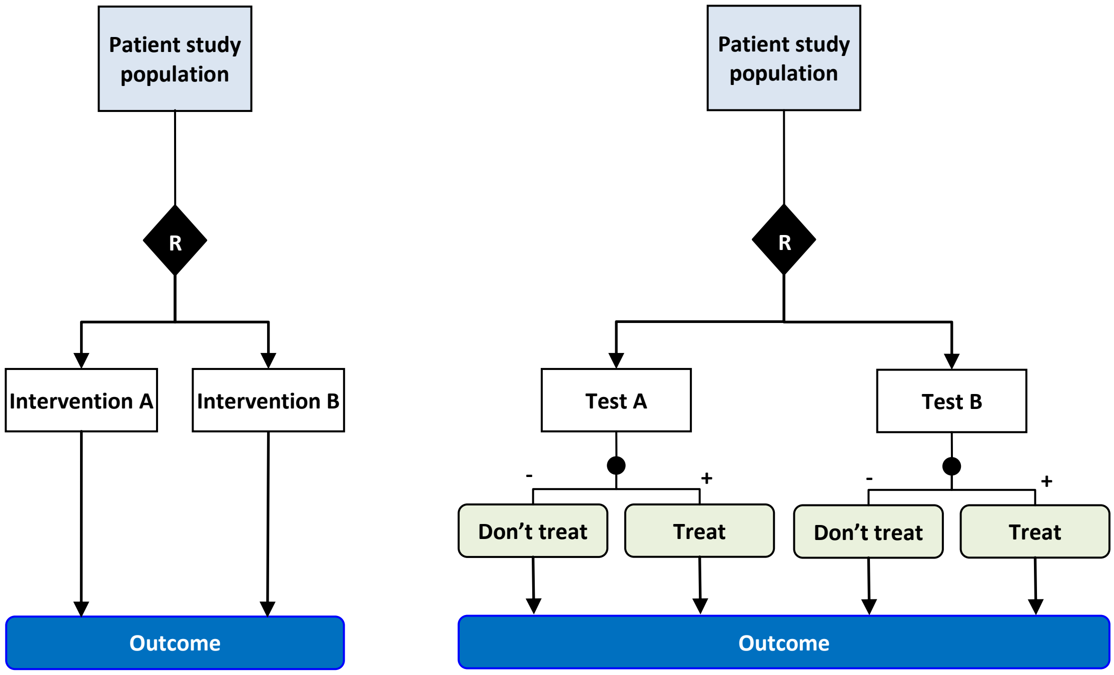


Figure 1.2: Schematic illustration of a typical single intervention RCT (left) by comparison to the multiple intervention test-treatment RCT (*right*).

(health response). Their experimental design also enables investigators to distribute patients at random[81] and to implement other controls that limit bias. In sum, the RCT is therefore the most powerful epidemiological design for concluding that observed differences in outcome are due to differences in the intervention under study, since all other things can be kept equal.

When the goal is to measure the impact a *test* has on patient health, the RCT must evaluate a different intervention. Patients are randomised to undergo either the new test or the existing test, however the downstream health response is measured after the implementation of subsequent treatment. Therefore when we seek to evaluate tests we must compare entire management pathways, called 'test-treatment' strategies, rather than single interventions (Figure 1.2). A recent example is provided by the MRC–CUBE trial which evaluated whether testing dyspeptic patients for the bacterium *Helicobacter pylori*,

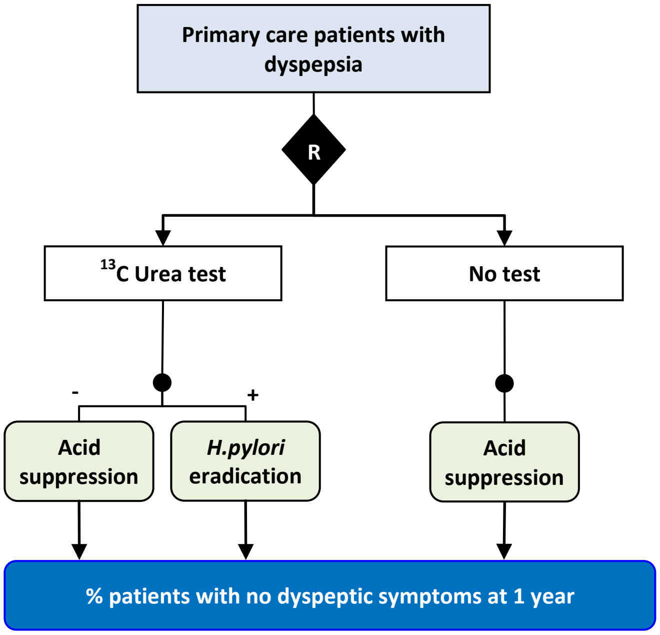

Primary care patients with dyspepsia
R
^{13}C Urea test
No test
-
+
Acid suppression
H.pylori eradication
Acid suppression
% patients with no dyspeptic symptoms at 1 year

Figure 1.3: Example of a replacement test-treatment RCT.
Patients randomised to the experimental arm receive a test for the presence of Helicobacter pylori, which is eradicated if found, while patients without bacterial infection are given proton pump inhibitors (acid suppression). Patients randomised to the control arm receive no test and are all given proton pump inhibitors (acid suppression), reflecting standard care[27].

would effectively reduce their symptoms when compared to the standard approach of giving acid suppression to all such patients[27] (Figure 1.3).

Test-treatment comparisons can take three general formats, depending on the role the new test will take within the existing strategy[22]. The MRC–CUBE trial describes a replacement comparison where the new test completely replaces the existing technique (in this case no testing), however RCTs can also measure the value of adding a new test to the existing strategy, as was compared recently in the RATPAC trial (Randomised Assessment of Treatment using Panel Assay of Cardiac markers)[26] (Figure 1.4). This NIHR HTA funded trial contrasted two strategies for diagnosing acute myocardial infarction in patients who had presented to emergency departments with acute chest pain that was suspected to have been caused by acute myocardial infarction. It evaluated whether the

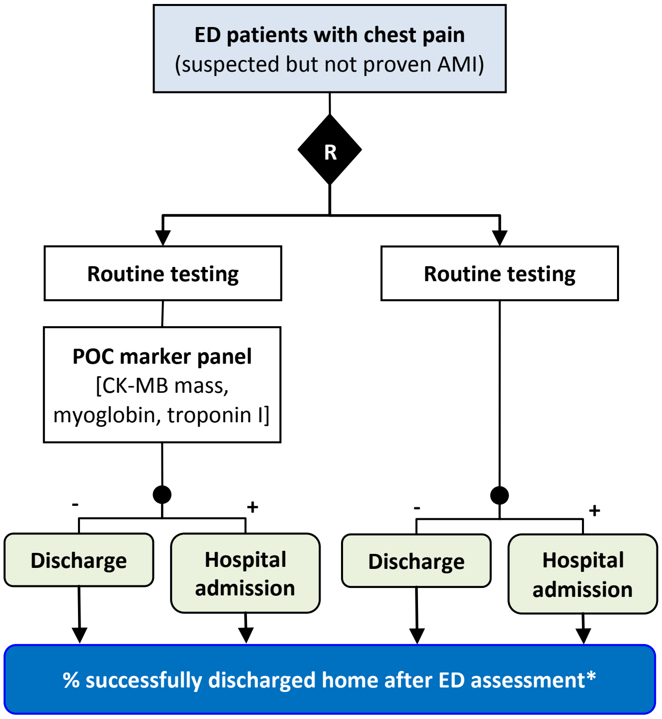


Figure 1.4: Example of an add–on test-treatment trial[26].
AMI – acute myocardial infarction; CK–MB – Creatine kinase muscle type; ED – emergency department; POC – point–of–care;
* Primary outcome defined as: discharge decision made within 4 hours of presentation AND suffered no major adverse event during the following 3 months.

addition of a new point–of–care marker panel to routinely used tests could reduce the proportion of patients suffering adverse events as a consequence of missed diagnoses, when compared to the standard measurement of cardiac biomarkers for AMI.

Alternatively the new test can be inserted earlier in the management pathway, and used to select which patients will go on to receive the existing technology. This triage comparison was performed in the RELAPSE trial, which evaluated the benefit of triaging patients with a clinically suspected recurrence of throat cancer[82]. Since the standard strategy is to proceed directly to invasive inspection by laryngoscopy, the new strategy sought to spare patients unlikely to have a recurrence by first investigating with a contrast–enhanced imaging modality (fluorodeoxyglucose enhanced positron emission scanning) (Figure 1.5).

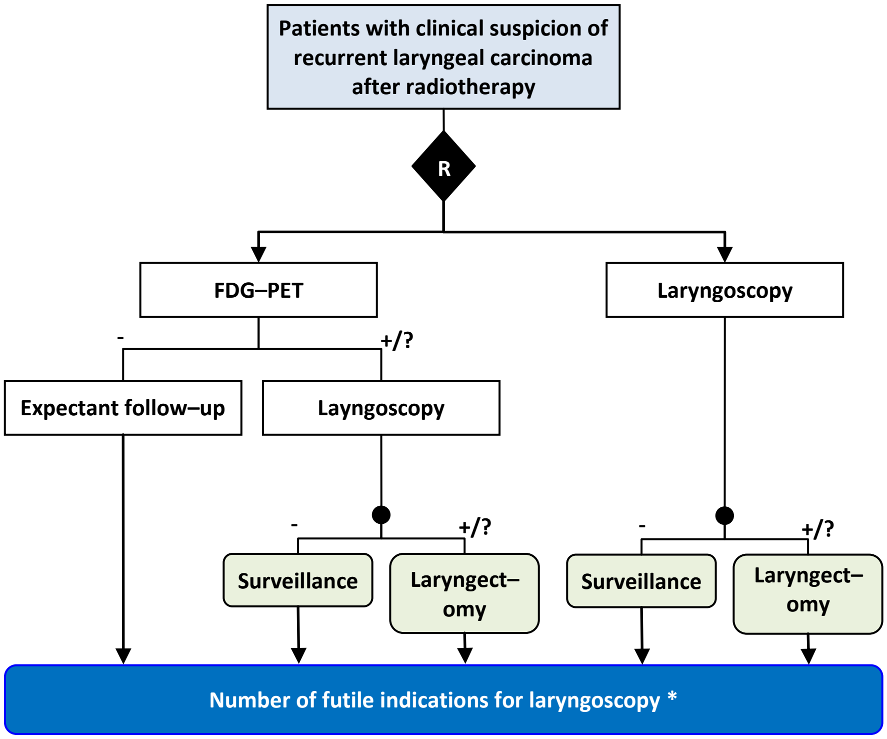

Patients with clinical suspicion of recurrent laryngeal carcinoma after radiotherapy
R
FDG–PET
Laryngoscopy
-
+/?
Expectant follow–up
Layngoscopy
-
+/?
Surveillance
Laryngect–omy
-
+/?
Surveillance
Laryngect–omy
Number of futile indications for laryngoscopy *

Figure 1.5: Example of a triage test-treatment RCT.
Patients with symptoms of recurrent cancer currently undergo laryngoscopy to investigate, followed by partial or total removal of the larynx if confirmed. In this trial patients randomised to the experimental arm receive a fluorodeoxyglucose enhanced positron emission (FDG–PET) scan, only proceeding to laryngoscopy if the results are positive or indeterminate[82].
* defined as negative laryngoscopies with no recurrence was diagnosed within 6 months follow–up.

In all types of test-treatment RCT the final measure of effectiveness is made after all tests and treatments have been administered, thus these downstream outcomes measure the impact of all these processes. However test-treatment interventions are more than the sum of multiple interventions: tests and treatments are connected by several phases of clinical decision–making in which test results must be used to formulate diagnostic decisions, which in turn inform the treatments that must be administered to each resulting diagnostic group. For the MRC–CUBE trial, the chosen measure of downstream patient health (recovery of symptoms) was assessed at 1–year follow–up, thus it evaluated the effectiveness of the tests (^{13}C Urea breath test for study arm and no test for controls), the treatments (acid suppression and bacterial eradication regime) but also of the clinical decisions that occurred between them (does the patient have a bacterial infection? should they be prescribed empirical acid suppression or the eradication regime?).

Test-treatment strategies are therefore far more complicated interventions to evaluate by comparison to pharmaceutical interventions. Not only do test-treatment strategies comprise multiple healthcare components, but also multiple episodes of decision-making. Indeed, test-treatment strategies can be described as "*interventions with several interacting components*", and so appear to satisfy the criteria for 'complex interventions' as defined by the MRC in their guidance document for developing and evaluating complex interventions[83].

Decision–models

The impact that testing strategies have on downstream patient outcomes can also be estimated indirectly using decision analysis. These models are constructed using existing clinical data and extrapolate the link between a test's accuracy and downstream health outcomes[84]. This is accomplished by setting out each test-treatment strategy along a decision tree which expresses the sequence of decisions and events that occur as a result of testing[85]. Decisions include the diagnoses that may be given, while events describe potential differences in the health status of patients such as whether or not they have the target condition. To illustrate, Howard and colleagues used modeling to compare two diagnostic strategies for managing patients with suspected common bile duct stones[86]. Diagnostic endoscopic retrograde cholangiopancreatography (ERCP) is the current gold–standard for detecting stones in the bile duct, however it is also a highly invasive procedure risking serious morbidity. This is followed by therapeutic ERCP in patients found to have stones. By comparison, the new strategy would initially examine patients using magnetic resonance cholangiopancreatography (MRCP), a non–invasive imaging technique, after which patients in whom stones were still suspected (‘test–positives’) would undergo the standard management strategy.

Figure 1.6 illustrates a simplified decision–tree that was used to compare ERCP–led and MRCP–led management. For every test-treatment strategy that is compared, each possible alternative sequence of decisions and events occupies a different ‘branch’ of the tree. Namely, each branch describes a variation in the management patients receive as a result of being allocated a particular diagnosis, and undergoing a treatment dictated by that diagnosis. The difference in ultimate patient outcomes is estimated by comparing the proportion of simulated patients in each tree who experience the desired health outcome after having progressed along a particular branch. The lower branch describes patients undergoing ERCP who could receive either a positive or negative diagnosis. Those who test positive and truly have stones (true–positives) will proceed to treatment, removal during therapeutic ERCP, as will false–positive patients. Patients who initially test negative and are truly free of stones (true–negatives) will ultimately receive different diagnoses and so won’t proceed to therapeutic ERCP, while false–negative patients will be erroneously discharged and re–present, after which the test–treat process is repeated through to completion. In the upper branch, patients who undergo MRCP and receive a positive indication for the presence of stones will proceed to diagnostic ERCP, regardless of whether this diagnosis is correct, after which all further management is the same as the standard care branch.

These models must therefore estimate how many patients in each strategy will travel along each possible branch and experience the downstream effects of doing so. These parameters are quantified by probabilities that must be retrieved from several existing primary studies[6,87–88]. For example, Howard and colleagues obtained probability estimates of a patient having bile duct stones from a database listing epidemiological studies of

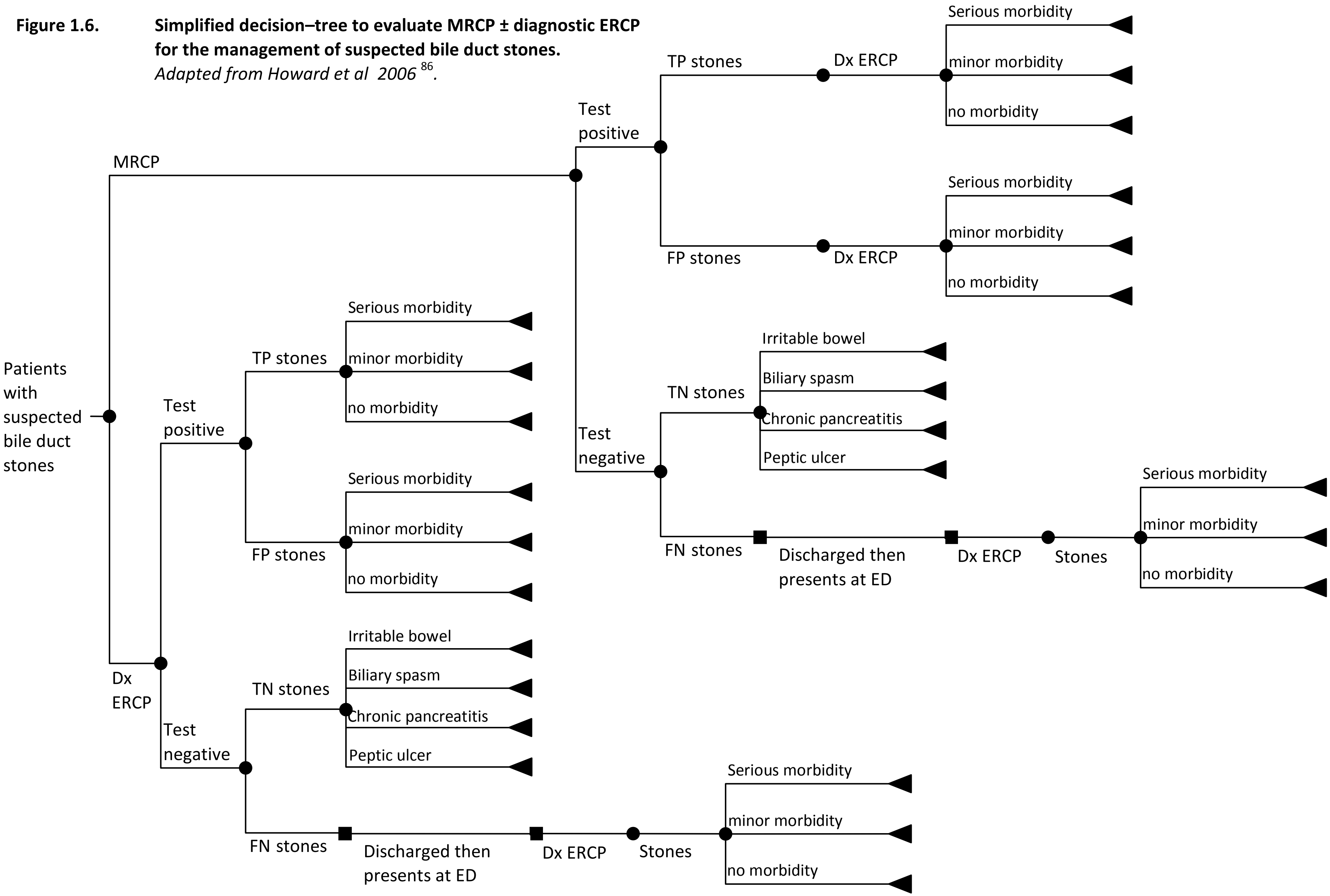


Figure 1.6. **Simplified decision–tree to evaluate MRCP ± diagnostic ERCP for the management of suspected bile duct stones.**
Adapted from Howard et al 2006[86].

disease prevalence (the Australian Bureau of Statistics); the probability of receiving a positive or negative test result was gained from their systematic review of diagnostic test accuracy comparing MRCP with ERCP[86]; and the probability of experiencing adverse outcomes, including complications and death, were extracted from various sources including previous models that used mortality registers and extensive observational surveys[89–90]. The probability of responding to treatment can also be secured from trials evaluating treatment effectiveness[87].

The issue of what evidence is needed to perform an appropriate decision–analysis has been addressed by several groups of researchers[21,84,91–92]. Attention is drawn in particular to the meticulous and thought–provoking research published by Lord and colleagues[19,21,93], who delineate what evidence of treatment efficacy should be sought and how it should be linked to evidence of a test's performance. They note that model design should be informed by appraising the trade–offs that occur within a given comparison of two tests; these trade–offs occur as a result of the various ways in which a new test is expected to improve patient health, most commonly as a consequence of superior diagnostic accuracy or through the direct harms and benefits of undergoing the test[21]. A common example is the trade–off that occurs with triage tests. CT–pulmonary angiography, for instance, can accurately diagnose suspected pulmonary embolism (PE), however a disadvantage is that it exposes the patient to radiation. D–dimer, a protein biomarker measured to detect clotting in blood samples, is highly sensitive and may be able to rule PE out safely in patients with a low clinical probability of disease, thus avoiding the more risky and costly CT[94].

The use of models to evaluate the clinical effectiveness of tests certainly offer several advantages over the RCT, as a result of which this approach is accepted by the major health technology assessment agencies, including NICE's Diagnostic Assessment Program here in the UK[18], the Agency for Healthcare Research and Quality (AHRQ) in the

USA[16], and the Medical Services Advisory Committee (MSAC) in Australia[95]. Because models are constructed using existing data, they are relatively quick to perform and at lower cost than an RCT[6,88]. Various researchers note that models can be also used to indicate the need for an RCT[8,84], to perform cost–effectiveness analyses[88] or simply to contemplate what the health consequences of different diagnostic strategies might be[87].

Of course, the validity of decision models is limited by the availability and quality of existing evidence, by the need to rely on assumptions that all patients will be treated according to the protocol, but also by the need to extrapolate the results of several studies. This inevitably must assume that the estimates are transferable, however this may not always be a valid assumption; since diagnostic accuracy varies according to several factors, including the case–mix of study populations, the role of the test in the new pathway and how the test is carried out[60], actual test performance may vary from that reported in the primary studies from which estimates are retrieved[87].

Perhaps the greatest limitation, however, is that models only provide *indirect* evidence of the effects that test-treatment strategies have on patient health. The only rigorous method for acquiring direct evidence is to perform RCT evaluations of test-treatment interventions. Yet because of their complexity, this approach has attracted some criticism regarding the feasibility with which rigorous test-treatment RCTs can be conducted. Before examining these criticisms, it is first appropriate to review existing knowledge regarding the methodological shortcomings of RCTs in general.

1.2 Bias and poor reporting in RCTs

The randomised controlled trial is argued to be the gold–standard design with which to evaluate the impact healthcare interventions exert on patient health[17]. This is due to the investigator’s ability to implement several methodological techniques that can prevent any distortions introducing bias, thus enabling the objective and reproducible empiricism that is required of scientific experiments. In practice, however, it may be difficult or undesirable to execute the necessary methods, resulting in estimates of effectiveness that may be biased and hence unreliable. The following section provides a review of the evidence regarding the biases that can occur as a result of implementing inadequate methodological safeguards in trials of treatment interventions.

1.2.1 Bias in RCTs

The goal of the RCT is to evaluate whether a new treatment succeeds in safely improving the health of patients by comparison to no treatment or existing treatment. In order to conclude that observed differences are caused only by the different treatments, trials must adhere to the fundamental principle that all other aspects of the comparison should be kept equal[96].

Biased RCTs are theorised to stem largely from four methodological misdemeanours that violate this principle: creating mismatched study groups (selection bias), treating study groups differently besides the interventions being compared (performance bias), measuring study groups differently (detection bias) or analysing mismatched study groups (attrition bias). Empirical analysis of trial design supports the role that methodological safeguards play in limiting these risks of bias by demonstrating that trials using inadequate methods tend to produce significantly different results to adequately–performed trials. The most rigorous evidence for these associations is produced by comparing the quality of trials included in subject–specific meta–analyses. Nine such 'meta–epidemiological' reviews have been published[97–105], five of which have been synthesised into two meta–meta reviews[106–107]. A third meta–meta review, the largest of all, was published recently and combines seven meta–epidemiological studies[108] (Table 1.1).

Study	Included meta-analyses	Quality components examined	MAs	RCTs
Schulz et al 1995[97]	Cochrane Pregnancy and Childbirth Group meta-analyses with ≥5 trials containing ≥25 events in the control group, and ≥1 trials with and ≥1 trials without adequate allocation concealment	Random sequence generation, allocation concealment, blinding, reporting of exclusions.	33	250
Moher et al 1998[98]	Random sample from authors' database of meta-analyses, selected from 3 areas: digestive diseases, circulatory diseases, mental health. Random sample from the Cochrane Database of Systematic Reviews, one on stroke and two on pregnancy and childbirth.	Random sequence generation, allocation concealment, blinding, reporting of exclusions.	11	127
Jüni et al 2000[99]	Handsearch of 8 journals 1993–1997	Allocation concealment, blinding.	133	NR
Kjaergard et al 2001[100,110]	Cochrane Library, Medline or PubMed with at least one trial with ≥1000 patients	Random sequence generation, allocation concealment, blinding, description of dropouts and withdrawals.	14	190
Balk et al 2002[101]	Author selected cardiovascular medicine meta-analyses, and MEDLINE + Cochrane Database of Systematic for infection, paediatrics, surgery including ≥6 meta-analyses	Random sequence generation, allocation concealment, blinding, reporting intent-to-treat analysis, reporting power calculation.	26	276
Egger et al 2003[102]	Cochrane Database of Systematic Reviews that had performed comprehensive literature searches.	Allocation concealment, blinding.	122	1175
Contopoulos–Ioannidis et al 2005[103]	Cochrane Mental Health Library all meta–analyses with at least one “large” randomised trial (sample size >800) and at least one “smaller” trial.	Random sequence generation, allocation concealment, blinding.	16	133
Siersma et al 2007[104,111]	Cochrane Library random selection, each with ≥5 trials with ≥1 inadequate and ≥1 adequate allocation concealment	Random sequence generation, allocation concealment, blinding, reporting intent-to-treat analysis, reporting power calculation.	48	523
Nuesch et al 2009[105]	Cochrane Library, Medline, Embase, CINAHL, all trials comparing therapeutic interventions in hip or knee osteoarthritis	Exclusions to primary analysis.	14	167

Table 1.1: Overview of meta–epidemiological studies investigating the association between RCT methodological quality components and treatment effects (continued overleaf).

NR–Not Reported; MAs–Meta–analyses; RCT–Randomised controlled trial

Study	Included meta-analyses	Quality components examined	MAs	RCTs
Jüni et al 2001[106]	Schulz 1995a, Moher 1998, Juni 2000, Kjaergard 2001	Random sequence generation, allocation concealment, blinding.	NR	NR
Wood et al 2008[107]	Schulz 1995a, Kjaergard 2001, Egger 2003	Allocation concealment, blinding.	146	1346
Savović et al 2012[108]	Schulz 1995a, Kjaergard 2001, Balk 2002, Egger 2003, Siersma 2007, Contopoulos–Ioannidis 2005, Pildal 2007	Random sequence generation, allocation concealment, blinding	234	1973

Table 1.1: (continued) Overview of meta–epidemiological studies investigating the association between RCT methodological quality components and treatment effects.
NR–Not Reported; MAs–Meta–analyses; RCT–Randomised controlled trial

These reviews begin by categorising all trials included in the original meta–analyses according to the adequacy with which each methodological safeguard has been performed. The effect estimates are then pooled for each category (generally 'adequate' and 'inadequate/unclear'), and a ratio of these pooled estimates is calculated for each meta–analysis. The weighted averages of the resulting relative odds ratios (RORs) are subsequently examined using a random–effects meta–meta–analysis in order to examine the association between quality components and treatment effects[109]. The main findings of these reviews are summarised below.

1.2.2 Randomisation, allocation concealment and selection bias

Randomisation of eligible participants ensures that, on average, study groups are comparable in their composition of participants, specifically regarding particular prognostic subgroups of patients that may be predisposed to experience poorer or better downstream outcomes[81]. This is achieved using two methodological safeguards. First, an allocation sequence based on random number generation is designed to eliminate the predictability of the next participant's group assignment, as well as to ensure that prognostic factors are distributed at random between study groups. Second, this schedule is concealed from recruiting physicians in order to prevent foreknowledge of which intervention the next eligible participant would receive, and thus expose the allocation process to intentional or subconscious subversion[112–113]. Published accounts of the subversion of allocation schedules[114] attest to clinicians' determination to provide what they perceive as being the best care to their patients, so adequate methods of concealing which interventions might be allocated to the next eligible patient are necessary to enforce clinical equipoise.

Evidence that inadequate generation of randomisation schedules causes bias was scarce until recently, most reviews failing to confirm that this feature is independently associated with larger effect sizes[97,98,101,103,110]. In their now seminal review of 33 meta–analyses published by the Pregnancy and Childbirth Group of the Cochrane Collaboration, Schulz and co-workers demonstrated that inadequate randomisation was only associated with larger treatment effects when limiting the comparison to adequately concealed trials[97], raising the possibility that adequacy of sequence generation is not sufficient to prevent bias if schedules are not subsequently concealed from recruiting care–providers. The recent publication of the large meta–meta review that synthesised over 230 meta–analyses appears, however, to confirm somewhat definitively that inadequate or unclear methods are associated with a clear exaggeration in treatment effect of 11% on average[108].

Most reviews found that trials using inadequate (including unclear) methods of concealment tend to have larger effect sizes of between 17%[107] and 30%[106], highlighting that allocation concealment is critical to the prevention of selection bias. Two meta–reviews found no such difference[101,104], though their results may have been influenced by using less rigorous quality appraisal criteria to define 'inadequate' methods. The criteria published by Siersma and colleagues are incomplete[104], precluding a firm interpretation of the review's results, however those published by Balk and colleagues[101] have certainly been criticised as inconsistent with standardised definitions of methodological adequacy[115]. Nonetheless, when Balk's review was incorporated into a meta–meta–review alongside four other reviews[97–100], a 25% exaggeration of treatment effects due to inadequate methods was still apparent (ROR 0.75, 95%CI: 0.63-0.89)[116]. This finding was confirmed by Savović and colleagues, though the exaggeration was found to be at a much smaller 7% (ROR 0.93, 95%CI: 0.89–0.99)[108].

Schulz and colleagues also demonstrated that unclear concealment remained associated with larger effect sizes even after accounting for possible biases introduced by inadequate methods of randomisation, exclusions after randomisation/missing data, or not blinding participants[97]. What is more, inadequate or unclear allocation concealment has also been associated with a greater likelihood of finding statistically significant treatment effects[117], as well as greater heterogeneity in treatment effects between trials of similar topics[97,100,108,111]. This indicates that selection bias is unpredictable in its impact and can either overemphasise or underestimate the true effect of interventions.

1.2.3 Blinding: performance and ascertainment bias

'Blinding' (or 'masking') is conducted to satisfy the second chief tenet of experimental study designs: to eliminate, as far as is possible, any contamination of the intervention's effect due to pre-existing beliefs regarding the intervention's effectiveness. Much as allocation concealment is used to prevent such beliefs from influencing the composition of study groups, blinding is conducted to ensure that – other than the treatments under study – the provision of care, response to care, and measurement of this response, are all conducted equitably across study groups. Thus practical procedures must be established to warrant that physicians and other care–providers remain unaware of which interventions participants have been assigned to, so that prior expectations of effectiveness do not encourage a disparity in other care that is administered. This performance bias is also avoided by blinding patients to knowledge of the intervention they are receiving, so as not to unduly influence their response to treatment, while also facilitating the proper blinding of treating staff. Those measuring endpoints must be blinded so as not to support systematic differences in how outcomes are assessed, so foiling the potential for ascertainment or detection bias.

Despite the clear rationale for the risks imposed by these two types of bias, the available empirical evidence is inconsistent in demonstrating that all open (i.e. un–blinded) trials produce more biased results than blinded trials. The strongest evidence is derived from the three meta–meta–reviews which found that absence of double–blinding was associated with treatment effects on average 7%[107], 12%[106] and 13%[108] larger than those of trials using double–blinding. These indications of bias appear to be restricted to the results of subjective outcomes[107–108]. Specifically, subjectively measured treatment effects were found to be between on average 22%[108] and 25%[107] larger in open compared to blinded trials, whilst no association was found for the similar comparison of objective outcomes. Savović and colleagues also found that both between–trial and between–meta-analysis heterogeneity were considerably higher when the analysis was restricted to trials with subjective outcomes[108].

On the other hand, two meta–epidemiological studies, again those by Siersma and Balk, failed to show any difference in effect size as a consequence of double–blinding[101,104]. This lack of consensus is likely to reflect a multitude of factors, not least the variation in how reviews judged blinding to have occurred. No review examined the *adequacy* of blinding methods but instead all used reporting of ‘double–blinding’ as a proxy for methodological sufficiency. Trial reports of the methods used to implement blinding remain very poor[118–119], and so this approach may have underestimated the impact that lack of blinding has on effect size if some trials that reported blinding did not in fact implement this safeguard adequately. Moreover, use of the term ‘double–blinding’ has been shown to denote a broad variety of precisely who should be blinded[120]. This would suggest that trials classified by these reviews as ‘blind’ are likely to vary in the degree to which they remain at risk of bias, and whether this is performance bias and/or detection bias, both of which could have confounded the association between reports of blinding and effect size.

Other causes of inconsistency in findings may be due to the varying degree to which reviews controlled for other aspects of trial quality. At least part of the impact of blinding can be explained by the impact of allocation concealment for example, since trials using inadequate concealment are also more likely not to implement blinding[107]. When this confounding was controlled for, by limiting the analysis to trials with adequate allocation concealment, blinding ceased to cause any overall difference in effect size, though maintained an association with overestimated effects once subjective outcomes only were analysed, albeit with a very wide confidence interval (ROR=0.80, 95%CI: 0.49–1.31)[107]. Again no differences were observed for objective outcomes.

This provides two important indications for optimal trial design: first that blinding adjudicators for the assessment of objective outcomes may be superfluous to the prevention of ascertainment bias since both approaches provide equivalent treatment effects. Second that since outcome assessors can substantially influence the evaluation of endpoints that accommodate an element of subjectivity, for example symptom frequency, blinding these adjudicators may be particularly critical to a trial’s validity[107,121].

1.2.4 Loss, exclusions and attrition bias

The fourth threat to the internal validity of RCTs can occur if the groups are no longer comparable at the time of analysis. Loss to follow–up as a result of participants becoming unavailable during the study period, and the exclusion of randomised participants from a trial for a variety of reasons, can theoretically both cause a systematic shift in the distribution of prognostic factors that had so carefully been eliminated by proper randomisation procedures. Individuals who are lost or excluded from trials are unlikely to constitute a random sample of those initially recruited, and instead may well differ from those that remain with regard to their treatment response or other prognostic characteristics[122]. Participants may be in too poor health to continue with the demands of study participation, while those that are withdrawn after randomisation are more likely to be participants with the poorest response to allocated treatment. The selective elimination of randomised individuals from the calculation of treatment effects is thus liable to compromise the validity of results by introducing the risk of attrition bias[123]. Moreover, the impact of attrition bias will be heightened if the reasons for losses and exclusions differ between study groups as a direct result of the interventions being received, or if study groups experience different rates of attrition[124]. For these reasons, the ideal calculation of treatment effects should include all participants in the final analysis as randomised, regardless of the interventions they actually receive or whether they complete trial follow–up[123]. This approach is referred to as 'intention–to–treat' analysis.

Three recent studies attest to the impact that attrition can have on trial results[105,125–126]. A meta–epidemiological analysis that included 14 meta–analyses evaluating treatments for osteoarthritis found that trials tended to demonstrate greater benefits of the experimental intervention if they suffered attrition than those providing complete analyses[105]. Similarly, two cohort studies that examined within–trial differences in effect size by directly comparing intention–to–treat analyses with per–protocol analyses (74 RCTs[126] and 133 RCTs[125]) both found that calculations excluding participants tended to over exaggerate the benefit of experimental treatments than estimates derived from intention–to–treat calculations[125–126]. All three studies found attrition to be associated with both overestimates and underestimates of intervention effect, concluding that attrition is unpredictable in its impact on effect magnitude.

Conversely, several studies fail to find that attrition impacts on trial results. A small meta–regression analysis of a convenience sample of 10 RCTs concluded that both attrition and differential attrition can occur at random, thus not causing attrition bias[124]. Using individual patient data for each included trial, the authors compared the level of baseline imbalance in all randomised participants to the degree of imbalance in all patients included in each primary analysis[124]. They found that attrition did not result in baseline imbalances, while the level of attrition was not correlated with the observed direction of effect. Similarly, five meta–epidemiological analyses failed to find any association between attrition bias and effect size[97–98,100–101,104]. The methods used by each were again highly variable, and all relied on different surrogate measures of methodological quality. One review used the adequacy of reporting the number and reasons for dropouts and withdrawals, regardless of the actual attrition observed[100]; another used reports of an intention–to–treat analysis regardless of whether an ITT was actually conducted[104], while two reviews used reports of exclusions to indicate poor quality[97,101]. As with the evidence for blinding, these results are highly likely to be confounded by poor reporting quality. In an appraisal of 110 randomly–selected RCTs conducted within the field of obstetrics and gynaecology, Schulz and colleagues found that trials reporting exclusions to the primary analysis exhibited inferior study quality to those with no apparent exclusions, thus suggesting that reporting perhaps did not reflect true conduct in the latter subgroup[127].

1.2.5 Sample size and type II error

Recruiting an insufficient number of participants into a trial cannot bias trial results as such, though it may distort the interpretation of results by increasing the magnitude of random error. In order to conclude whether an observed difference in outcome rates is meaningful, investigators must test its statistical significance. Yet deductions reached through hypothesis testing can succumb to two errors: false–positive conclusions and false–negative conclusions. Type I errors occur when a trial falsely concludes a difference between treatments when in reality there is none. Particularly critical to clinical trials are the risk of type II errors, in which real differences between interventions are missed due to insufficient sample sizes. The probability of avoiding false conclusions of no effect is denoted by the concept of power (1– the probability rate of type II error, or β), which is inversely related to sample size; namely the probability of arriving at false–negative conclusions decreases as the number of study subjects increases[128]. Trials with small sample sizes are demonstrably at risk of having erroneously concluded an absence of treatment effect when in reality the probability of achieving false–negative conclusions was high[129]. At least one–third of published trials may underestimate the sample size required to eliminate type II error by more than 10%, and approximately 6% of trials underestimate the required sample size by over 50%[130], suggesting that the risk of false–negative conclusions in published trials is commonplace.

Nevertheless, it is challenging to quantify the impact that inadequate approaches to managing the risk of type II error has on trial findings. Larger trials (>800 participants) have been shown to produce more conservative effect estimates with increased precision than small trials included within the same meta–analysis[103]. However this difference may be in part explained by publication bias, since smaller trials with non–significant results are less likely to be published than large trials[110]. A similar study also found small trials to have exaggerated effect sizes compared to large trials (>1000 participants), though only when comparison was limited to trials with inadequate sequence generation (ROR=0.46, 95%CI: 0.25–0.83), inadequate allocation concealment (ROR=0.49, 95%CI: 0.27–0.86) and a lack of double–blinding (ROR= 0.52, 95%CI: 0.28–0.96)[100]. That no similar difference was observed between trials employing rigorous methodological safeguards indicates small sample size may be a marker of poor design quality in this cohort.

Indeed, not all small trials are at risk of type II error, but only those that are underpowered to find the desired clinically important treatment effect. Consequently *a priori* specification of a target sample size, in which acceptable α and β error risks and estimated event rates are defined in advance of trial recruitment, provides a better proxy indicator of trial quality than generic sample size. By documenting power calculations trialists demonstrate they have considered not only the key issue of type II error risk, but have also reflected on the minimum clinically meaningful difference that should be evaluated[121]. Only two meta–epidemiological studies have investigated the impact of reporting power calculations on effect sizes, and rather unsurprisingly neither found any difference in magnitude[101,104]. Whilst reports of power calculations may indicate that efforts to minimise type II error have been made, the risk will only be reduced in trials that recruit their target sample size.

1.2.6 Reporting quality

Although randomised controlled trials have the potential to provide the most reliable assessments of healthcare effectiveness, they are often complex experiments requiring exacting methods to achieve their goal. As summarised above, those failing to achieve high methodological standards risk basing their conclusions on biased findings. What is more, the considerable proliferation of published trials[131–133] requires users and commissioners of healthcare to identify the most rigorous evidence in order to select the most effective interventions for use in clinical practice. This can only be achieved with full and clear reporting of studies[121].

Trial reports are recognised to be necessary though imperfect proxies for actual design quality and conduct[134]. Unambiguous reporting of trial methods allows adequate appraisal of methodological quality, which in turn facilitates the synthesis of evidence by enabling the most rigorous studies to be identified. Problems ensue with incomplete reporting, since readers cannot be definite that absence of description equates with the absence of method, or with inadequate method[129]. This is demonstrated clearly in the above overview of bias in RCTs, where several conclusions regarding the relationship between methodological quality and treatment effect size were marred by poor reporting.

Clear reporting is also indispensable when seeking to translate beneficial interventions into clinical practice. Comprehensive accounting of the interventions given to participants in a trial is the only approach to guarantee that clinicians can identify and replicate desirable treatments safely[135]. The poor description of interventions has been cited as a major barrier to the implementation of research findings into practice[136], while inadequate description could even cause interventions to be carried out incorrectly, potentially to the detriment of resource–use and patient health[137].

Transparent accounting of trial design, conduct and analysis has been a central consideration of the evidence–based movement for almost two decades[138]. Faced with empirical evidence that poor reporting of trials was preventing assessments of their methodological quality[139], a large group of investigators developed the CONsolidated Standards Of Reporting Trials (CONSORT) in the early 1990s[139–140]. The resulting checklist was first published in 1996, and itemises which aspects of trial design and conduct should be reported in order to ensure that studies are presented in such a way that they can be clearly and independently interpreted[138]. These standards have subsequently been revised twice in line with improvements to the evidence–base[141–142], and are now endorsed by over 50% of the core medical journals[143]. Since the original guidelines targeted parallel treatment RCTs, the most common design, several extensions have also been developed to standardise and improve the reporting of different study designs (cluster[144], non–inferiority/equivalence[145] and pragmatic[146] RCTs) interventions types (non–pharmaceutical[147], herbal medicine[148] and acupuncture[149] interventions), and data types (harms[150] and abstracts[151]). Although cross–sectional surveys of published trials attest to a poor general level of reporting quality[152], the indications are that the situation is improving as a result of the CONSORT drive to improve standards of RCT research[118,153].

1.2.7 Summary

This overview of the methodological basis of RCTs has outlined the threats to a trial's validity and interpretability that result from using inappropriate approaches to trial design and conduct. Nonetheless, many investigators have succeeded in performing rigorous trials, that implement the necessary controls to provide reliable and useful evidence for the effectiveness of treatments[154].

Using RCT designs to evaluate different types of interventions is not always guaranteed to be as successful, however. There is a growing literature that suggests complex interventions of various clinical disciplines may be less well suited to evaluation by RCT due to the feasibility of implementing the strict methodological controls required to produce unbiased evidence. For example, researchers cite the ethical and practical difficulties of blinding patients and care–providers to different types of surgery[155], or to competing critical care services[156], or the impossibility of comparing psychiatric treatments in a controlled environment when they are so variable and tailored to individual patients[157]. Although test-treatment interventions seem not to have been described previously as complex interventions, several researchers have expressed similar concerns that the RCT may also not be feasible when comparing test-treatment interventions. The following section reviews these criticisms.

1.3 Challenges to the usefulness of test-treatment RCTs

As discussed earlier, the ideal approach to evaluating diagnostic tests heralds the RCT as the theoretical ideal for establishing patient outcome effectiveness. Yet this notion appears to have been transposed somewhat automatically from the treatment effectiveness paradigm. With the predominant focus of diagnostic research on test accuracy, very few studies have been conducted that address the suitability of RCT designs for evaluating the effectiveness of diagnostic tests.

Diagnostic interventions differ considerably from treatment interventions, since in order to evaluate tests trials must assess them as components of management strategies that also incorporate decision–making and treatment. The resulting complexity of these interventions has led some researchers to question the feasibility of conducting high–quality test-treatment RCTs in view of the unique challenges that these attributes may pose. This section summarises four particular challenges that are often cited to limit the usefulness of test-treatment RCTs.

1.3.1 Availability of trial evidence

The ability to rely on evidence from trials assumes that such evidence exists. Yet researchers and guideline developers alike suggest that test-treatment RCTs are rare[16,18–19,21,51,95]. Clearly the absence of top-ranking evidence will mean that reviews will have difficulty in providing the detailed guidance that is needed on which tests will improve downstream patient outcomes.

Trial findings are notably absent in conferences[9], while the explosion of prominent research into test accuracy methods, and accompanying increase in primary test accuracy studies, has perhaps encouraged patient outcome effectiveness research to focus on methods for linking accuracy findings to trial evidence of treatment effectiveness, thus bypassing the need to consider the more time–consuming and laborious approach required to get direct evidence through RCTs. Indeed many claims that RCTs are rare can be traced to works that expound the advantages of decision modelling to overcome the practical and methodological difficulties entailed in carrying out comparative effectiveness trials of tests[6,8,21,84,87–88]. While the substantial challenges involved in designing, conducting and interpreting these studies lend credence to this assumption of scarcity, there is some evidence that these claims may be justified. An early overview seeking to characterise the sorts of evidence available for the performance of MRI failed to find any RCTs of clinical or patient outcome impact from a systematically–derived cohort of 285 articles published in 1981–1987[48]. Taking a random sample of imaging evaluations from two radiology journals published in 1988–1989, Taylor and colleagues[52] found that only 16% (24/146) assessed clinical impact. Though the authors did not report how many of these where randomised trials, the implication must be that RCTs formed a small proportion – if any of the 24 articles retrieved. What is more the authors did not mention patients at any point in their analysis, hinting that patient outcome RCTs were probably absent from their cohort of studies.

More recently, the dearth of trials has been reported by reviewers attempting to synthesise available evidence. Of the few systematic reviews that address patient outcome effectiveness, most have failed to locate relevant test-treatment RCTs[24,158–164]. Similarly, none of the three diagnostic assessment reports published by the NICE DAP so far has identified any test-treatment trials[165–167].

Though such findings do indeed support suggestions that RCTs are rare, the evidence is limited. It is also somewhat indirect considering that no study has yet sought to verify exactly how many test-treatment RCTs exist, while there could be other explanations for their absence in systematic reviews.

It is not clear, for example, to what extent it may be due to difficulties in identifying these studies. Efforts to develop reliable search methods for reviews of test accuracy have by and large failed to produce high search sensitivities with acceptable levels of precision[168] (see however the recent work of Monica Kastner and colleagues[169]). Bibliographic indexing terms for diagnostic studies have, until very recently[170], simply not existed, a factor known to have reduced the accuracy of these searches[171]. These difficulties have been compounded by a widespread inconsistency in the application of diagnostic content terminology, as suggested by the variation in performance of search strategies according to the clinical question under evaluation[172–173].

By comparison, the synthesis of diagnostic effectiveness research is a relatively new endeavour that has yet to receive attention with regard to the development of methods for the identification of randomised test-treatment trials. Here, study ascertainment may present an even greater challenge. As described above, diagnostic tests are evaluated alongside treatment pathways for which a considerable volume of therapeutic evidence may well exist that is not *per se* relevant to the diagnostic question being evaluated. In view of the many thousands of trials published every year[133], finding test-treatment trials that are not indexed as such is likely to prove very difficult.

1.3.2 Internal validity of test-treatment RCTs

As described above, threats to the internal validity of RCTs have been well delineated, and though the use of inadequate methodological safeguards is shown to cause systematic deviations of treatment effects, these can be avoided through judicial planning. Although RCTs have therefore maintained their archetypal status as the most reliable measurements of treatment effectiveness[17], concern has arisen that randomised comparisons of test-treatment interventions may commonly fail to provide reliable contributions to the effectiveness evidence–base[21,51,174–175]. This obstacle is hypothesised to stem from an underlying incongruity between the complex composition of test-treatment interventions on the one hand, and the feasibility of implementing the adequate methodological safeguards needed to maintain a trial's internal validity on the other.

Selection Bias

The earlier overview of bias in randomised controlled trials summarised how the distortion of treatment effects due to selection bias is mainly driven by unconcealed randomisation procedures which allow patient allocation to be influenced by clinicians' existing beliefs surrounding the effectiveness of interventions under study. As Wood and colleagues note, selection bias should be increased in situations where it is easier for clinicians to assess patients' prognoses regarding treatment response[107]. Since randomisation to tests will occur earlier in the management process than for treatment RCTs, fewer prognostic indications should exist at the time of allocation to test-treatment interventions. Moreover, as the methods of concealing allocation schedules are not related to the type of intervention under study we can therefore expect no greater impediment to a properly randomised and secure allocation procedure in test-treatment evaluations.

Performance and Ascertainment Bias

Conversely, some have argued that it may be impossible to control for performance bias[22,79,176–177]. Since tests produce data that must be interpreted in order to select between various treatment or other management options, in many circumstances it may be impossible to mask the identity of the tests themselves from clinicians[22,79]. Consequently, as patients are often involved in treatment selection, and their care–providers are aware of group allocations, then effective blinding of participants could also be hampered[44]. Indeed, cursory examination by one author of four published test-treatment trials found that none reported any form of blinding, suggesting these fears may be born out in reality[174]. These difficulties raise the possibility that the results of trials could reflect a measure of current clinician and patient expectations, rather than true differences in the effectiveness of test-treatment interventions. On the other hand, methodologists argue that it should generally be feasible to blind outcome assessors and thus control for ascertainment bias[22,44,174].

Attrition Bias

RCT evaluations of complex interventions are vulnerable to increased drop–out rates due to the multiple phases of treatment that patients are required to adhere to[178], an obstacle also reasoned to affect test-treatment evaluations[176]. Since the quality and type of information patients receive may differ according to the test to which they have been randomised, these trials may also be susceptible to differential drop–out and so are at an increased risk of serious attrition bias[176]. As described above the principle of intention–to–treat, or 'intention–to–test', helps to limit this bias in any trial design and so remains relevant to the test-treatment RCT[6]. However, Blackmore makes the case that conducting these analyses alongside sizeable reductions in patient compliance rates could deleteriously affect a trial's power to detect the desired clinically important difference and so expose findings to the risk of type II error[175].

Power

Though not a bias, the risk of falsely concluding the absence of an effect due to inadequate sample size constitutes an important threat to the validity of any trial, as detailed earlier in the chapter. However the hazards of insufficient power may present a greater challenge to the design and conduct of test-treatment trials, that are conceived to require far larger study populations than is typical for treatment trials[176,179–180]. It is hypothesised that in the majority of comparisons the potential benefits of downstream treatment will only be experienced by patients who are reclassified as a result of receiving a more accurate test, and so receive a more beneficial treatment[176–177]. Consequently the overall treatment effect will be diluted by what Pletcher and Pignone[177] refer to as the 'unreclassified fraction', namely the subgroup of patients who would receive the same diagnosis and treatment by both tests under evaluation. Since differences in the diagnostic sensitivity of comparative tests are unlikely to be considerable[6,51], the 'reclassified fraction' is expected to represent a small subset of the study population; as a result, sample sizes will need to be several orders of magnitude larger than is usual for single intervention trials, for whom the treatment effect could be experienced by all randomised participants[176,179]. Trials failing to recognise this methodological peculiarity therefore risk being underpowered to detect differences in patient outcomes.

The results of a small systematic study begin to suggest that this need has yet to be met, however. Prompted by the publication of four trials finding no clinical or health benefits from the use of fetal fibronectin (FFN) testing in women with threatened preterm labour, the authors set out to investigate whether these ‘negative’ findings could be explained by problems with the design of those trials[181]. The authors found that only two of the four included trials adjusted the power calculation to account for the unreclassified fraction.

1.3.3 Utility of trial evidence

Aside from the ability to conduct evaluations that are internally valid, researchers suggest that the evidence produced by test-treatment trials may often be difficult to interpret and even trickier to use. While the clear reporting of methodological safeguards should be equally possible for these studies, the utility of trial evidence may be threatened by the increased intricacy required to describe interventions that consist of multiple healthcare components[176]. Moreover, as with other types of complex intervention, test-treatment strategies are in fact greater than the sum of their parts[182]: they not only demand adequate description of two healthcare interventions, a test and a treatment, but they should address the decision–making processes that occur between the two.

The recent CONSORT extension for trials of non–pharmaceutical interventions emphasises the enhanced requirements needed to document these therapies, which like test-treat strategies often include multiple interacting interventions[147]. The authors summarise how such surgical techniques, rehabilitation programmes and behavioural interventions have been found to suffer from poor reporting due to their complexity, and the resulting possibility for variation in how they are implemented across the healthcare services.

Similarly, diagnostic decision–making is also likely to be difficult to describe in trial protocols, and perhaps more so than non–pharmaceutical therapies since tests create different patient subgroups, each of which may receive further testing and multiple treatments. Moreover, diagnostic and therapeutic decisions are highly variable. They vary between clinicians according to skill, expertise[183] and individual attitudes to the balance of risks resulting from missing diagnoses or over–treating patients[65]. Decisions taken on a particular test result also vary within individual clinicians; not only is the interpretation of diagnostic information likely to evolve as familiarity with a particular test increases[6,44], but the same test results in the same patients are shown to produce different diagnostic interpretations according to the nature of prior information available[184]. So, even if tests and treatments are well–described, it may be considerably challenging to outline the approaches used to interpret test results and select subsequent treatments, particularly to a degree sufficient to enable these interventions to be reproduced in practice[79,174,176].

Previous examinations of test-treatment trial reports have found it difficult to deduce how treatment decisions followed from test results[76,181]; Vis and colleagues failed to find any evidence for how test–treat decisions should be made in all four included RCTs[181]. Yet without a protocol which directs how test information should lead to diagnostic decisions, and subsequently to treatment selections, it is impossible to construe the meaning of observed effects since we cannot be sure of which processes are actually being evaluated[22,79]. Moreover, if there is no pre–specified instruction on how tests results should be used, clinicians taking part in trials may not have sufficient guidance to respond to diagnostic information, and to do so consistently. This could lead test-treatment interventions to fail in demonstrating an effect due to inadequate implementation of the tests involved, rather than through lack of their effectiveness[76] .

Consequently, if the complexity of test-treatment trials means users of evidence cannot discern what is being evaluated and how test–treat strategies are being used, it will be impossible to interpret trial findings, compare them across studies or to use beneficial interventions in practice. Test-treatment trials therefore run a high risk of not being informative.

1.3.4 Full evaluation of intervention effects

A final criticism regarding the utility of trials, albeit one that is encountered more rarely in the literature, is that they may fail to fully evaluate test-treatment interventions.

The demonstration of an intervention's value is produced by outcomes, hence the choice of what is to be measured in a trial is a fundamental aspect of its design and execution[185]. Evaluating outcomes which fail to measure all important effects of an intervention adequately are likely to misrepresent its effectiveness, with potentially serious implications for the content of future healthcare policy decisions. Outcomes which are too narrowly focussed on capturing an intervention's benefits, for example, may not be sensitive to an intervention's harms and could lead a trial to conclude a more beneficial impact than may be present in reality. Essentially, a trial may completely miss a potentially beneficial or detrimental effect if the selected outcome measures have not been designed to capture it.

It is well recognised that the selection of which endpoints can offer useful and comprehensive measures of treatment effect is a complex task fraught with many difficulties[185], which must also take into account the validity of outcome measures and importance of outcomes to patients[186]. Current thought alludes to a potential increase in its intricacy when we come to consider trials of test–treatment strategies, however. The outcome measurements are further removed from the intervention of interest (the diagnostic test) by a second interventional stage (the treatment), as well as the diagnostic and treatment decision-making processes in between. Thus there is said to be an indirect relationship between cause and effect, whereby the end measurement of effect captures not only the impact of testing, but also of the subsequent decision–making and treatments[59,77,175,187]. This creates two complications for trial design.

First, tests can affect patient outcomes in several ways. Some tests can cause direct harm as a consequence of undergoing testing; CT for example exposes patients to radiation, which increases an individual's lifetime risk of developing cancer[188], while endoscopic techniques are invasive and can cause more immediate harm due to procedural complications[189]. Other tests may reassure patients as to the absence of serious disease[48,175,190]; in individuals with Chronic Daily Headache, a group of conditions in which sufferers experience frequent and long–lasting headaches[191], neuroimaging has been shown to reassure as to the absence of serious organic disease[192]. Most tests are expected to alter downstream health indirectly through improvements to treatment selection afforded by better diagnostic accuracy and more appropriate decision–making[14,193]. However, several researchers note that accurate tests do not necessarily translate into health improvements because of the intervening requirement for them to enhance decision–making[14,44–46,193]. The results of diagnostic and therapeutic impact studies support this theory, finding that more accurate tests can fail to change diagnostic decisions, while tests that change diagnoses do not always lead to improvements in treatment selection. For example, contrast–enhanced CT was shown to change diagnoses in 32% (40/125) of emergency patients with acute abdominal pain, and alter treatment plans in 25% (31/125)[194]. The authors focussed on the seemingly similar percentages of change, however when cross–tabulated only 11% (14/125) experienced a change in both diagnosis and treatment plan; 65% (26/40) of changed diagnoses failed to be followed by alterations to treatment plans, while 55% (17/31) of all changed treatment plans were not preceded by a change in diagnosis. The intricacies of how tests affect outcomes are therefore likely to increase the number of potential effects in test-treatment trials, and so the number of processes that need to be measured.

Second, as patients are randomised before testing but the treatments they receive vary according to their diagnosis, test–treatment populations will include patients with different diagnoses, receiving a variety of treatments. Deeks has highlighted how this increases the number of patient outcomes that must be taken into account for a single trial[176]. Though investigators may primarily be concerned with how well a test can detect a particular disease, restricting measurements to outcomes that evaluate the response to treatment for a single ‘target condition’ will provide incomplete evaluations of effectiveness. This is because not all randomised patients will be free of disease; some individuals with negative diagnostic indications for the target condition will have another disease, as may some of those who incorrectly receive a positive diagnosis for the same condition. Since the consequences of incorrect diagnosis and/or inappropriate treatment are likely to vary between conditions, Deeks contends that these trials need to measure health events relevant to all included patients[176].

In summary, because both the number of process effects and of patient outcomes are increased in test-treatment comparisons, identifying and measuring them all could prove very challenging. In view of the dominance of test accuracy perspectives which stipulate the focus on single target conditions, these notions are arguably unlikely to have been incorporated in existing RCTs. However, in the absence of research that appraises the appropriateness of outcome selection in RCTs, the veracity of this statement remains unknown.

1.3.5 The need for research

These four areas of potential methodological difficulty threaten the suitability of the RCT to questions of diagnostic effectiveness. It is clear that if they are confirmed, it will be impossible to produce trial results that are free from bias, and upon which we can rely to conclude whether tests have succeeded in changing patient health and clinical decision–

making for the better. Indeed such a finding would seriously call into question whether the RCT can be upheld as the archetypal study design for diagnostic health technology appraisals, as is the norm for treatments. Equally, as opinion that test-treatment RCTs are often ‘unattainable’ begins to spread[16], these potentially valuable study designs may be avoided unnecessarily if certain issues are discovered to be less challenging than feared. Indeed, some authors are more optimistic arguing that though the prospect of producing trials is difficult, the existence of completed RCTs suggests the issues entailed are surmountable[8,57]. None of these claims, whether supportive or antithetical to the utility of test-treatment RCTs, has so far been defended by an empirical appraisal of the studies themselves. Research is needed to verify the extent to which these concerns are encountered in test-treatment trials that have been completed.

1.4 Research questions & thesis overview

1.4.1 Research questions

Due to the current discord between the RCT’s traditional standing as the ‘gold–standard' for evaluating the patient outcome effectiveness of diagnostic tests and hypotheses that they may not actually be useful, the main aim of this thesis is to investigate **how useful RCTs are for evaluating the patient health impact of diagnostic tests.** This central research question will be answered through four secondary aims, each designed to evaluate a key methodological challenge that has been posited to impinge on the utility of test-treatment RCTs:

1. Are test-treatment RCTs feasible?
2. How informative are test-treatment RCTs?
3. Are test-treatment RCTs internally valid?
4. Do test-treatment RCTs fully evaluate their interventions?

1.4.2 Thesis overview

The review of methodological challenges to the utility of test-treatment trials highlighted the scarcity of evidence to either support or refute the four criticisms. In order to provide this evidence, each of the four thesis aims will be addressed by examination of published test-treatment trials. **Chapter 2** presents the methods used to systematically identify a representative cohort of published test-treatment RCTs, which will form the basis of all further analysis into the methodology. The search strategy and study selection process are described, along with the general characteristics of the final group of published test-treatment RCTs.

1.4.3 Aim 1: Are test-treatment RCTs feasible?

Recognising that there may be substantial challenges involved in conducting test-treatment RCTs, a key question is whether these studies can be completed successfully. The availability of published evidence is likely to be one marker of a study design's feasibility. Though current thought argues that test-treatment trials are rarely conducted, there are as yet no studies that explore how many trials are published, nor any that investigate what questions these 'successful' trials have been designed to answer.

Chapter 3 seeks to ascertain how rare test-treatment trials really are, in order to establish whether RCTs are currently useful for providing the evidence that is needed on which tests will improve patient outcomes. Acknowledging that these studies may be very difficult to find, and thus the search conducted in chapter 2 may have missed some relevant trials, chapter 3 presents a different search strategy and compares it to the results of the original search. The 'capture-recapture' technique, developed by ecologists, is used to estimate the number of relevant RCTs missed by both strategies. This allows the total number of test-treatment trials published during the study timeframe to be estimated.

Chapter 4 aims to characterise the diagnostic settings in which RCTs have been executed to completion. It describes the cohort of trials identified in chapter 2, provides a descriptive overview of the diagnostic questions evaluated by published test-treatment RCTs and offers an insight into their methodological approaches. This contextual understanding will also serve as a basis for generating hypotheses regarding methodological or practical limitations that are discovered in the following chapters.

1.4.4 Aim 2: How informative are test-treatment RCTs?

The complexity of test-treatment interventions is also hypothesised to inhibit the interpretation of trial findings, the appraisal of their quality, and their translation into practice. The extent to which test-treatment trials are not informative is explored in **chapter 5** by critically appraising the reporting of trial conduct. Perceived barriers to the reporting of these studies are discussed by reference to the study characteristics, described in chapter 4.

1.4.5 Aim 3: Are test-treatment RCTs internally valid?

The second challenge to the utility of trials relates to claims that these interventions may be particularly susceptible to several biases due to their complexity. To date, however, there are no substantial empirical appraisals of published studies, as have been conducted in other areas of healthcare research.

Chapter 6 aims to substantiate the extent to which these trials are predisposed to the biases that are claimed to affect them by appraising the methodological quality of trials identified in chapter 2. In the chapter’s discussion, findings are compared to similar reviews of treatment RCTs and complex interventions in order to establish the relative susceptibility of test-treatment interventions to selection bias, performance bias, ascertainment bias and attrition bias.

1.4.6 Aim 4: Do test-treatment RCTs fully evaluate their interventions?

The number of ways in which test-treatment interventions exert their impact on patient health is increased, by comparison to single treatment interventions. The final challenge therefore speculates that identifying and measuring all important effects may prove very challenging. In order to evaluate this contention, the ideal analysis would appraise the *appropriateness* of outcome selection in included trials. In order to be reliable, such a judgement would require extensive clinical expertise to be able to identify the most important outcomes for the full range of diagnostic settings included. Importantly, the appraisal would have to be adjudicated by reference to a solid theoretical understanding of how test-treatment interventions cause treatment effects. This theory was found to be lacking and in urgent need of development.

In order to address this important deficit, **Chapter 7** develops a theoretical framework that conceptualises all the ways in which tests influence health outcomes. It achieves this by synthesising existing theoretical notions, and using them to generate a preliminary explanatory model. This model is tested, refined and explained by examination of the project cohort of published test–treatment RCTs using analytic inductive methods.

Based on the author’s experience of using this framework to interpret test-treatment RCTs, **chapter 8** develops the conceptual framework into a practical tool. The tool is presented as a checklist and accompanying graphic schema, and its value to the design, interpretation and appraisal of test-treatment RCTs is illustrated by worked examples, derived from the project cohort.

1.4.7 Central research question: how useful are test-treatment RCTs for evaluating the patient health impact of diagnostic tests?

Chapter 9 summarises the main research findings from chapters 2–8, and discusses the evidence they provide to address each of the four challenges. This argument is drawn together in the final conclusion to answer the overall aim of the thesis. The implications of these conclusions for practice and research as well as general limitations of the thesis are also discussed.

2

Finding test-treatment trials

Search strategy & study selection

This chapter presents the methods used to identify a cohort of published test-treatment RCTs, which are analysed with regard to their clinical context, reporting quality and quality of methods in subsequent chapters. The search strategy and study selection process are described, along with the general characteristics of the final group of published test-treatment RCTs.

In order to examine and develop the methodology that underpins evaluations of test-treatment interventions, it was first necessary to identify a group of primary studies that could highlight the strengths and weaknesses of current research practice. Since the chief concern was that these studies should be representative of their study design, a systematic method of study ascertainment was selected.

Systematic search methods hold an eminent position in evidence-based medicine, since the reliability with which evidence synthesis is produced depends on the ability to identify and incorporate all best evidence[123]. By incorporating comprehensive searches that are objective, and hence reproducible, the systematic review minimises bias in the collection of data to the increased validity of resulting effect estimates of healthcare interventions.

The aims of this search differ slightly from those of systematic reviews designed to inform a specific clinical treatment question. So as to characterise the breadth of test-treat questions evaluated by these studies (reviewed in Chapter 4), searching was not limited by patient group, condition or test technology. Secondly, the search was not intended to be exhaustive, i.e. to identify all published test-treatment RCTs ever published, but was aimed at generating a group of trials whose analysis of methodological quality could be generalised to all test-treatment RCTs.

Throughout the thesis included trials are listed separately from other references, and citations numbers (1–108) are prefixed by the letter 'T'.

2.1 Search Methods

2.1.1 Search strategy development

A search strategy was developed to identify test-treatment RCTs that had been published in academic journals. Designed to retrieve as many relevant trials as possible, the initial strategy sought to identify studies that evaluated an aspect of a diagnostic test in an RCT design, and was developed using Ovid Medline, a database that uses articles indexed by the United States National Library of Medicine (US-NLM). Since test-treatment trials are not currently indexed as a specific study type, the search concepts were taken from the key methodological elements required of included studies, defined simply as: ‘diagnosis’, ‘randomised controlled trial’ and ‘treatment’. Search terms representing each of these concepts were identified, and three similar strategies were tested (Appendix A.1, p. 350), each variation focussing on both the different sensitivity and precision of two RCTS method filters, and the inclusion of the term ‘control$’ (where $ denotes an unlimited truncation). Their combined yields were so considerable as to be considered unmanageable (31,896 to 187,895), largely due to their inclusion of a high proportion of evidently non-experimental study designs or therapeutic evaluations.

2.1.2 Final Search Strategy

In order to obtain more precise results, searching was performed in a bibliographic database known to contain a higher proportion of relevant study designs. The Cochrane Central Register of Controlled Trials (CENTRAL) includes all reports of RCTs indexed through highly sensitive searches of both Medline and Embase, as well as handsearched material, and other additional extensive database searches contributed by the Cochrane specialised registers[195–196]. This multisource composition also offered the potential to identify articles not included in the US-NLM.

Search strategy		Hits
#1	sensitiv* or diagnose or diagnosis or diagnostic* in Clinical Trials	70,052
#2	random* in Clinical Trials	335,175
#3	"study design" next "rct" in Clinical Trials	150,275
#4	(#2 OR #3)	449,453
#5	(#1 AND #4)	50,419
#6	(#5), from 2004 to 2007	12,892

Table 2.1: Search strategy for test-treatment RCTs conducted in CENTRAL Issue 2 2009 (Wiley InterScience, searched 29 May 2009) – general diagnosis textwords across all fields limited to publication years 2004 to 2007 (12,892 records). *denotes truncation of search term.

The structure of the final strategy was also modified, through an examination of terms appearing across all fields in five test-treatment RCTs, found by the author during development of the search strategy(T59–T61,T65,T71). Terms denoting a generic phase of treatment, as opposed to individual, condition-explicit treatments, were found to be too non-specific to discriminate test-treatment from treatment only RCTs, and so were discarded (Table 2.1).

Terminology relating to the concept of ‘diagnosis’ again tended to be specific to test technologies (e.g. ‘imaging’ or ‘microbiology’) or test types (e.g. ‘computed tomography’ or ‘microbial sensitivity test’). However all five articles used a variant of the term ‘diagnosis’ in either their title, abstract or keywords, while two also referred to ‘sensitivity’ in the abstract or keywords. All other terms commonly included in diagnostic accuracy filters were notably absent, including ‘specificity’, ‘accuracy’; ‘prediction’, and ‘detection’[172–173,197]. ‘Test’ was referred to in two articles(T59,T71), however was considered too non-discriminatory to include in the strategy. Under advice from an information specialist working for the Cochrane Collaboration, and highly experienced in designing searches for diagnostic test accuracy reviews (Anne Eisinga, Diagnostic Test Accuracy Working Group, UK Cochrane Centre), two strands were added to limit CENTRAL results to randomised trials. A textword component sufficed to pick up Medline RCTs, while Embase and heandsearched RCTs were targeted by the specific study type added to these studies by Cochrane indexers[198].

CENTRAL (2007, Issue 2) was searched using this strategy without language restriction, succeeded in identifying the five relevant references that had been identified previously, and so was adopted. A timeframe of the most recent four years was selected, and in order to ensure that all relevant trials had been found by Cochrane indexers and included in CENTRAL, the final search was updated in 2009 (29 May, Issue 2).

2.2 Study Selection

Electronic search returns were imported as text files into Microsoft Access 2007. All entries were cross-checked for duplication using a title-matching query function, and removed from further review. The selection methodology complied broadly with that of a systematic review, the only exception being the absence of a necessity to identify all target studies. Titles and potentially relevant abstracts/full papers were screened by the author to determine study relevance.

2.2.1 Study eligibility

Inclusion criteria dictated the selection of study type, diagnostic setting and outcomes measured. Only English-language papers were considered.

Study Type

Randomised controlled trial evaluating two or more test-treat strategies. Eligible test-treatment RCTs randomised patients to two or more testing strategies, subsequently provided treatment based on the results of the strategies, and measured at least one downstream patient outcome.

To be considered an RCT, the study had to contain an explicit statement that study patients were randomly assigned to comparative groups, as outlined in the revised CONSORT statement[121].

Diagnostic setting

Any clinical test used to classify symptomatic disease for the purposes of treatment planning. For the purposes of analysing a single methodological question, selection was limited to what could be considered the most common purpose of testing in healthcare: the single use of a test in an individual with suspicious signs or symptoms. This encompassed tests used to rule suspected disease in or out, to otherwise further a differential diagnosis, or to determine the stage or progression of existing disease. Tests used to detect preclinical disease in asymptomatic individuals (e.g. population screening), to monitor disease using serial testing (e.g. for treatment titration or ongoing observation), or purely to establish a prognosis without assisting treatment decisions (e.g. to estimate the likelihood of a future health state) were excluded. Tests were not limited by type, and all modalities were included.

Outcomes

Measurement of at least one downstream patient health outcome. These were defined as any markers of disease, physiological or psychological status pertaining to an individual, that describe an attribute of a subject's health after the full test-treatment intervention has been implemented. Studies that only measured patient outcomes during intervention implementation, for example after the test but before treatment, were excluded.

1.	Randomised trial evaluating a test, i.e. patients randomised to one of two or more diagnostic strategies
2.	Incorporation of a treatment phase, contingent on test results
3.	Evaluation of patient outcomes after treatment
4.	Test used for diagnosis or staging
5.	Full paper in English Language

Table 2.2: Summary of study eligibility criteria

2.2.2 Screening Process

Title screen

Titles were selected for review of the study abstract if they reported either a generic diagnostic term (e.g. ‘assessment’, ‘diagnosis’, ‘examination’) or a particular diagnostic test (e.g. ‘ultrasound, ‘radiography’, ‘oximetry’). Titles were excluded if they clearly described a study that was:

- Not an RCT design
- A treatment RCT
- An accuracy study of non-RCT design

Abstract and full paper screen

Abstracts were reviewed by the author using a selection proforma (Appendix A.2, p. 352) detailing the full inclusion criteria summarised in table 2.2. Full papers were ordered for all potentially relevant entries, as well as those for whom abstracts were unavailable. Final selection was conducted on the basis of the full article, and related publications were traced for further information. This included searching for associated results of included published protocols up to the end of 2009, in order to maximise the project sample size. Test-treatment RCTs that only reported downstream patient outcomes in a related publication were included.

2.3 Search Results

The search strategy retrieved 12,892 citations (Figure 2.1). After eliminating duplicates, the author checked 12,706 unique titles and excluded 88% on the basis that they were plainly evaluations of a treatment or a non-RCT study design. Of the remaining 1,569 abstracts, 1262 were excluded for reasons summarised in table 3.3, and 307 full reports

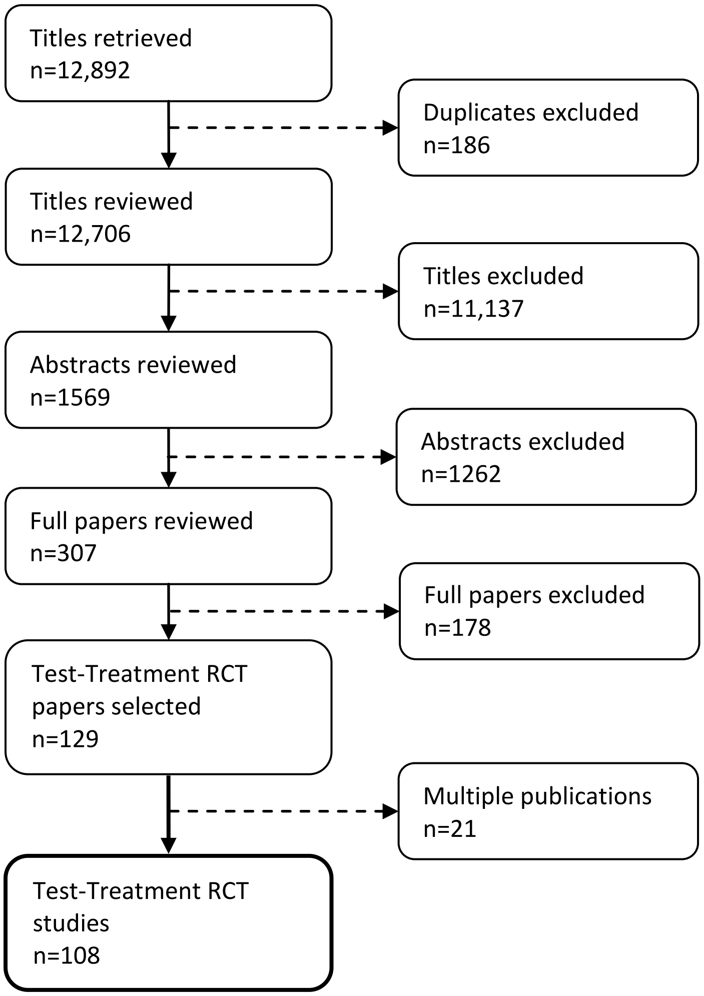


Figure 2.1: Study selection process for records retrieved by the final search strategy (CENTRAL 2009, Issue 2)

Reason for exclusion	Abstract	Full paper	Total
Non RCT evaluation of a test	599	11	**610**
Test RCT, no treatment	192	22	**214**
Test-Treat RCT, no patient outcomes	37	27	**64**
Test-Treat RCT, treatment not contingent on test results	6	3	**9**
Evaluating a different test (e.g. monitoring, screening etc)	80	66	**146**
Evaluating a treatment (any design)	261	20	**281**
Other study (e.g. case-control, cohort, behavioural interventions)	76	12	**88**
Foreign Language	8	17	**25**
No abstract	3	0	**3**
Total	**1262**	**178**	**1440**

Table 2.3: Reasons for excluding citations from the project cohort

Year	2004	2005	2006	2007	Total
Articles published per year	32	38	30	29	129
Search results	3133	3290	3225	3058	12706
Search yield	1.0%	1.2%	0.9%	0.9%	1.0%

Table 2.4: Absolute number of test-treatment RCTs published every year

were reviewed for more detailed consideration. In the final stages of selection, most exclusions were observational or descriptive evaluations of diagnostic tests (42%), such as test development or accuracy studies, while 20% were treatment evaluations that generally referred to a diagnostic test in the title (Table 2.3). In total 201 test-treatment RCTs were found, of which 64 did not evaluate patient outcomes and 9 did not provide treatment on the basis of test results, leaving 129 articles that met the predefined inclusion criteria. Common examples of excluded studies are tabulated in Appendix A.3 (p. 353).

Search precision, defined as the proportion of relevant articles in the total number of citations found, was therefore very low (1.0%, 128/12,706) with the author needing to read 100 titles to identify one relevant test-treatment trial (Table 2.4). Slightly fewer trials were published between 2006 and 2007 (n=59) than the preceding two years (n=70), though as search results were consistent across the four years this creates a marginally lower search yield for 2006–2007 (Table 2.4).

2.3.1 General characteristics of included trials

When 20 multiple publications were taken into account, the final project cohort consisted of 108 unique test-treatment RCTs evaluating at least one downstream patient outcome(T1–T108). Most trials were documented in a single results publication (92, 85%). For the remainder, 12 trials (11%) were reported across two articles (e.g. trial design protocols, economic analyses or long-term results), three trials (3%) were reported across three articles, and one (1%) was reported across five articles.

Five trials were only published as protocols, with no traceable trial results as of December 2009.

The 128 articles were published in 90 different journals, the majority in specialty journals (84, 65%). The highest number of reports appeared in *Radiology* (n=6), *Health Technology Assessment* (n=4), *Human Reproduction* (n=4), and *The New England Journal of Medicine* (n=4), while approximately two thirds (62, 69%) of journals published a single test-treatment RCT evaluating health outcomes during the four years.

2.3.2 Challenges in identifying the relevance of studies

In certain cases it was difficult to determine an article’s relevance to the project. These difficulties arise out of the project’s need to limit the scope of relevance to questions of symptomatic diagnosis and staging. In practice however the juncture between the five diagnostic settings, that is between screening, diagnosis, staging, monitoring and prognosis, can be far from distinct. Similarly, the distinction between diagnostic and therapeutic interventions can become blurred, particularly when the two are conducted as part of the same procedure. In order to ensure the reproducibility of the study selection process, a few examples of these decisions are described below:

Is the test under evaluation used for screening or diagnosis?

The most common difficulty was the need to distinguish between tests used for ‘screening’ or for ‘diagnosis’. According to the definitions set out in chapter 1 (p. 4–5), in principle screening tests differ from diagnostic tests by their intended purpose; namely the former are used to detect early disease in asymptomatic individuals. By contrast, diagnostic tests are used to establish the cause of a presenting complaint. However, since these functions occupy adjacent positions along the wide spectrum of clinical decision–making, there can be some uncertainty in discerning between the two when the clinical setting takes place at their boundary. Key to these decisions is defining whether the target population is ‘asymptomatic’, and can therefore be considered ‘preclinical’. The author’s rationale for these decisions is clarified by illustrating two test-treatment RCTs, one of which was ultimately excluded and the second included.

The first of the two examples consists of a trial evaluating two imaging regimes to detect Down’s syndrome in the unborn foetus[199]. Pregnant women at ≤13+2 weeks gestation were randomised to receive an ultrasound scan (US) at 12 weeks (experimental intervention) or at 18 weeks (standard care). The purpose of the US was to measure the degree of nuchal translucency, which is strongly associated with Down’s syndrome. Women with abnormal results would then receive invasive foetal karyotyping to confirm the likely presence of the genetic abnormality; in the experimental arm this was offered to all women, while in the controls only those of advanced maternal age (≥35 years) were put forward for further testing. Women gestating a foetus with structural anomalies, and hence confirmed to be carrying a Down’s syndrome baby, were given counselling by obstetricians, after which they could chose to terminate the pregnancy. The primary outcome was the live birth rate.

This was clearly a ‘test-treatment’ trial, in that individuals were randomised between two testing strategies and then received treatment (counselling) on the basis of test results. Several patient outcomes were measured after treatment, including the primary outcome, thus satisfying the third inclusion criterion. However, the target population could not be considered as symptomatic since women did not present with any complaints to raise the suspicion of Down’s syndrome. This was therefore deemed to be a screening trial, and excluded.

In the second example, a trial assessed the value of introducing extensive testing to detect cancer in patients with a confirmed first–episode of deep vein thrombosis (DVT) or pulmonary embolism (PE)[(T42)]. Individuals for whom their DVT/PE diagnosis was idiopathic, namely it could not be explained by known concurrent disease or history, were randomised to undergo a battery of imaging and laboratory tests (including US, CT, gastroscopy, colonoscopy, hemoccult, sputum cytology, mammography, Pap smear, prostate US, prostate specific antigen)(experimental intervention) or to receive no further testing, as standard. Since cancer is known to affect approximately 10% of individuals with idiopathic DVT, the purpose of the new strategy was to detect underlying malignancy. Patients received treatment appropriate to their diagnosis, and the primary outcome was cancer-related mortality.

As with the first example, this is clearly a test-treatment trial that measured patient outcomes after a phase of treatment. However, in this case the study population can be considered as ‘symptomatic’; since unexplained DVT may be caused by an underlying malignancy the presence of this condition was judged by the present author to act as a presenting complaint. Thus, even though the tests were described as ‘screening’ by the authors, a more detailed examination of the target population and reasons for testing revealed it to be closer to diagnosis, and so was included.

Is the test used for monitoring or diagnosis?

Monitoring situations are most clearly distinguished by the need to repeat tests[39], and so were generally easier to identify during the study selection process. Difficulties were however encountered when considering ‘continuous monitoring’ comparisons since tests were often used only once, though for an extended period of time. In these situations, decisions of whether to define studies as trials of diagnostic tests rested on how test results were used to manage patients.

To take an example, the bispectral index (BIS) is a form of brain status monitoring used whilst a patient is under general anaesthesia. It produces a summary score from various electro-encephalographic measures that indicates the patient’s level of consciousness. A trial assessing the value of adding BIS to conventional aesthetic management used it to optimise the level of anaesthesia received[200]. Since treatment was modified on multiple occasions during one patient’s operative procedure, BIS was judged to constitute treatment titration, and so the study was excluded.

Conversely, the use of cardiotocography and fetal pulse oximetry (FPO) for monitoring women in labour were considered as ‘diagnostic’ tests. One trial compared the impact of adding FPO to the standard cardiotocographic surveillance on operative delivery rates in women with non-reassuring fetal heart rates(T61). Cardiotocography is an electrophysiological test that records fetal heart rate and uterine contractions, while FPO measures the levels of oxygen in the fetus’s blood. Both are continuous monitoring tests designed to detect deteriorations in the baby’s well-being during labour[201]. During the trial, the management of labour was altered in response to any such deterioration, whereby as soon as fetal heart rates and oxygen saturation levels decreased beyond a predefined level assisted labour was initiated(T61). Since treatment decisions following these tests appear to have been made once, the trial was included as a diagnostic example.

Is the test used for diagnosis or immediate treatment?

Although *diagnosis* and *treatment* would intuitively appear to be separate entities, several instances were encountered where ‘testing’ and treatment were so closely connected that it was difficult to determine whether the ‘test’ was being used to make diagnostic decisions, or was purely being used as a therapeutic intervention. The use of medical devices during surgery serve as a clear example of the latter, for instance a trial comparing the efficacy of using ultrasound versus manual palpation to guide femoral puncture for cannulation[202]. Since ultrasound did not result in any diagnosis as such, it was judged to constitute part of the treatment process.

Trials were also excluded when tests were used for diagnostic purposes but the phase of treatment included in the trial was not contingent on results from the test. This is illustrated in a trial that randomised patients to laparoscopy (experimental intervention) or laparotomy (control) for staging and treating a known episode of uterine cancer[203]. Though the examination of resected material did eventually provide further information on cancer stage, the immediate aim of the procedures was not to select the subsequent treatment, but to remove tumours as part of the treatment process.

On the other hand, the author included similar comparisons when the test results were judged to directly inform the treatment conducted during the trail. Thus one of the selected test-treatment trials compared the use of routine axillary lymph node dissection to detect metastatic spread in women with a known diagnosis of invasive breast cancer, to a novel strategy which first triaged patients using the less–invasive sentinel lymph node biopsy[(T24)]. The results of these tests informed the choice between mastectomy or wide local excision; since these were also conducted as part of the trial, the study was included.

2.4 Discussion

The systematic search of CENTRAL led to the identification of 108 test-treatment RCTs evaluating a patient outcome after a treatment. These results demonstrate that test-treatment RCTs are carried out, although they appear to be present in small numbers. When examining included numbers by year of publication, the findings suggest that the observed trend of steep year-on-year increases in the publication of any RCT[131,133] may not hold true for test-treatment trials. Although the search was not designed to be comprehensive, the method was systematic and consistent for each of the four years so any actual increase in numbers published should be apparent. A preliminary inference must therefore be that comparative diagnostic effectiveness studies were rare between 2004 and 2007, and are probably unlikely to increase substantially in numbers in the short-term. This would favour existing opinion that randomised evaluations of test-treatment strategies are in fact rare[24].

Search precision was very low (1.0%), with the author needing to read 99 citations in order to identify one relevant study. Since searching was restricted to a database containing only clinical trials, the small numbers of relevant studies could also reflect the difficulties in identifying primary diagnostic research, largely due to the absence of diagnostic indexing terms in bibliographic databases. As a renowned problem in the synthesis of diagnostic research[173], the low search precision was anticipated and attempts to minimise this problem instituted by introducing methodological terms to maximise the search's sensitivity. An unexpected finding was the absence of terminology commonly encountered in diagnostic accuracy studies, which may make test-treatment RCTs more difficult to identify. Although these were eliminated from the final search strategy, the remaining methodological terms may have failed to locate relevant trials that did not include these methodological terms in their titles/abstracts or as index terms. If missed articles differ systematically from those ascertained by the project search, this could have resulted in a cohort that is unrepresentative of the true range of test-treatment trials.

A second potential explanation for the small number of included trials is that the author failed to identify relevant studies that were retrieved by the search. While missed numbers are likely to be low, the range of topics encountered was extensive and it may be the case that a screener with clinical expertise could have identified a higher number of test-treatment RCTs. If so, the search results presented here may have underestimated the true number of test-treatment RCTs published during the study timeframe.

In order to validate the rarity of test-treatment RCTs, the next chapter presents an independent verification of search methods which seeks to estimate the total number of these studies, and also determine the extent to which the author has overlooked relevant trials.

This chapter has been published:

Ferrante di Ruffano L, Davenport C, Eisinga A, Hyde C, Deeks J. A capture-recapture analysis demonstrated that randomized controlled trials evaluating the impact of diagnostic tests on patient outcomes are rare. J Clin Epidemiol 2012;65:282–287.

3

Estimating the number of test-treatment RCTs:

A capture-recapture analysis & inter–observer reliability study

Recognising that the methods used to find published test-treatment RCTs may have missed relevant studies, this chapter aims to definitively confirm the total number of trials published during the study timeframe. Using the 'capture-recapture' technique developed by ecologists, a second search of CENTRAL is compared to original search results to estimate the number of relevant RCTs missed by both strategies. Search results are also screened by a second reviewer to determine the extent to which the author underestimated the true number of trials picked up by both strategies.

RCT evidence of diagnostic clinical effectiveness cannot be considered useful if it is rarely available. The last chapter identified 108 test-treatment RCTs measuring patient outcomes that were published between 2004 and 2007, suggesting that these studies are very rare. However, in order to minimise the considerable number of search hits, whilst maintaining adequate search sensitivity, the search strategy may have failed to locate relevant trials that did not include these methodological terms in their titles/abstracts or as index terms. The search results therefore risk being incomplete, and the number of trials found may be an underestimate of the true number of relevant trials published in the study timeframe.

So as to provide a more definitive indication of the rarity of test-treatment RCTs, a 'capture-recapture' study was designed to estimate the total number of test-treatment RCTs published in the study timeframe, achieved by establishing the completeness of the original search. To determine whether the author's retrieval of relevant trials had been accurate and could be reproduced, an inter–observer analysis was also conducted.

3.1 Methods

The capture-mark-recapture method was developed by ecologists to estimate the size of animal populations that may be difficult to enumerate directly and completely. In its simplest form, two phases of population census are compared: within a predetermined study area, the target species are ‘captured’, counted, ‘marked’, and released. In the second phase, this process is repeated allowing the ‘recapture’ of some individuals and new capture of others. The number of discrepant captures in each census (i.e. those caught by only one survey) are multiplied and divided by the number of recaptures (i.e. those caught by both surveys) to provide an estimate of the number missed by both searches[204].

The method was first applied in epidemiology as a tool to adjust disease prevalence estimates for the likely incompleteness of the multiple registers from which the rates are derived[205–206]. The technique has since been advocated by evidence-based reviewers to establish the completeness of literature searches, also notoriously difficult to measure directly, where they tend to be referred to as ‘ascertainment intersection’ [171,207–208], as well as to develop stopping rules for systematic reviews[16] and to assess publication bias[210].

The present study uses the method to estimate the total population size of a specific interventional design: the test-treatment RCT. This is achieved by using two alternative search strategies, noting the overlap of relevant hits between the searches and then estimating the number missed by both.

3.1.1 Rationale for the second search

In order to apply the capture–recapture technique, a second search needed to be performed within the same bibliographic database. The original search strategy used methodological terms to identify test-treatment RCTs, not limited by test type; henceforth it is referred to as the ‘all–test–types’ (ATT) search. In order to identify trials indexed in CENTRAL that had eluded the ATT search, it was decided that the second search should

Imaging Test (IT) search		Hits
#1	MeSH descriptor Image Interpretation, Computer-Assisted explode all trees	4,967
#2	MeSH descriptor Magnetic Resonance Imaging explode all trees	3,624
#3	MeSH descriptor Positron-Emission Tomography explode all trees	520
#4	MeSH descriptor Endoscopy explode all trees	10,763
#5	MeSH descriptor Ultrasonography explode all trees	5,860
#6	(comput* near3 tomogra*):ti,ab,kw	5,730
#7	(positron emission near3 tomogra*):ti,ab,kw	1,227
#8	pet:ti,ab	1,320
#9	CT:ti,ab or CTs:ti,ab	4,315
#10	CAT:ti,ab near3 (imag* or scan*):ti,ab,kw	16
#11	SPECT:ti,ab,kw	790
#12	MRI:ti,ab,kw or fMRI:ti,ab,kw or WBMRI:ti,ab,kw or DWI:ti,ab,kw	2,640
#13	NMRI:ti,ab,kw	2
#14	(MR or NMR):ti,ab,kw near3 tomogra*:ti,ab,kw	30
#15	(MR or NMR):ti,ab,kw near3 imag*:ti,ab,kw	638
#16	magnetic resonance:ti,ab,kw	5,043
#17	diffusion weighted:ti,ab,kw	142
#18	T2-weighted:ti,ab,kw	305
#19	echoplanar:ti,ab,kw	23
#20	(ultrasound or ultrasonogra* or ultrasonic*):ti,ab,kw	12,033
#21	(echocardiogra* or echoencephalogra* or endosonogra*):ti,ab,kw	5,025
#22	(elastogra* or elastomet* or sonoelastic* or viscoelastic*):ti,ab,kw	288
#23	(elasticity near3 imag*):ti,ab,kw	29
#24	acoustic radiation force:ti,ab,kw	1
#25	(endoscop* or angioscop* or arthoscop* or bronchoscop* or cholangiopancreatogra* or colonoscop* or colposcop* or culdoscop* or cystoscop*):ti,ab,kw	11,450
#26	(duodenoscop* or enteroscop* or esophagogastroduodenoscop* or oesophagogastroduodenoscop* or esophagoscop* or oesophagoscop* or fetoscop* or foetoscop* or fluoroscop* or gastroscop* or hysteroscop* or laparoscop* or laryngoscop* or mediastinoscop* or neuroendoscop* or proctoscop*):ti,ab,kw	9,066
#27	(sigmoidoscop* or thoracoscop* or ureteroscop* or videolaryngoscop* or videoendoscop* or videocapsule* or endocapsule* or pillcam or mirocam):ti,ab,kw	920
#28	(video-assisted near2 surgery):ti,ab,kw	200
#29	(#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28)	58,895
#30	(#29), from 2004 to 2007	10,427

Table 3.1: The second search: CENTRAL Issue 2 2010 (Wiley InterScience searched 23 February 2010) – MeSH and textwords for five imaging modalities limited to publication years 2004 to 2007. Asterisk (*) denotes truncation of search term

target specific diagnostic tests directly by using content–specific terms. Ideally, this would incorporate the names of all diagnostic tests. This should include those evaluated by the 108 included trials, but also the names of tests not represented in the project cohort. Since the number of different tests in existence was huge, the author decided to restrict the types of tests searched for in order to make the strategy manageable. So as to be confident that the second search would yield enough relevant papers, it was decided to focus on the most frequent field of testing as determined by the ATT search. While full details of the clinical context of included trials are presented in chapter 4, suffice it to note that diagnostic imaging was most often found to be the subject of comparison in test-treatment RCTs, of which the most common modalities were: ultrasound, endoscopy, MRI, CT and PET. Accordingly, the second search was designed to target these five imaging tests directly by using specific test names, and all known permutations thereof, as MeSH terms and text words (Table 3.1). The strategy was designed in collaboration with an information specialist experienced in the ascertainment of diagnostic studies (Anne Eisinga, UK Cochrane Centre, Diagnostic Test Accuracy Working Group). No methodological terms were used, and no further restrictions were incorporated other than the publication time–frame (2004–2007) which remained identical.

3.1.2 A tale of two search strategies: Imaging Test (IT) & ATT searches

A key prerequisite of the capture–recapture method is that all ‘individuals’ (i.e. test-treatment RCT articles) in the ‘population’ (i.e. CENTRAL) should have the same probability of capture within each search[204]; this implies that the population of articles in CENTRAL should be identical for each search. Since indexing of studies by bibliographic databases is known to lag behind true publication dates[133], querying the same database on later occasions is likely to produce higher hit–rates due to this increasing population.

All Test Types (ATT) Search		Hits
#1	sensitiv* or diagnose or diagnosis or diagnostic* in Clinical Trials	73,262
#2	random* in Clinical Trials	349,718
#3	"study design" next "rct" in Clinical Trials	154,113
#4	(#2 OR #3)	441,388
#5	(#1 AND #4)	53,074
#6	(#5), from 2004 to 2007	13,495

Table 3.2: ATT search repeated in CENTRAL Issue 2 2010 (Wiley InterScience searched 23 February 2010) – general diagnosis text words across all fields limited to publication years 2004 to 2007 (13,495 records).
Asterisk (*) denotes truncation of search term

In order to ensure the consistency of the population, both searches were conducted in issue 2 of the 2010 CENTRAL database. This involved re–running the ATT search (originally run in CENTRAL 2009, issue 2) (Table 3.2), and identifying new records by eliminating duplicates using a title-matching query in Microsoft Access 2007.

Titles and potentially relevant abstracts/full papers were screened by the author, and inclusion criteria identical to the 2009 ATT search were applied: randomisation of patients to two or more testing strategies, provision of treatment based on the results of the strategies, and measurement of at least one downstream patient outcome. Tests used for screening asymptomatic individuals, repeated-test monitoring, and estimating the probability of future events were again excluded, as were foreign language papers.

3.1.3 Measuring inter−observer reliability

In order to measure the reliability of the author's screening process, a random 10% sample of records retrieved by both searches was screened independently by a colleague (Dr. Clare Davenport, University of Birmingham), a clinician highly experienced in systematic reviews and diagnostic research. The sample was created by: allocating all search hits a number (starting from 1), generating a random series of numbers using STATA (version 11 SE), and subsequently selecting the hits whose numbers appeared in the random number list. The colleague was provided with the resulting database of relevant citations along with the study protocol for including trials. Decisions to include studies were based on the review of full papers, and any discrepancies were discussed, with final inclusion decisions resolved by consensus. Inter-observer agreement was calculated for total agreement beyond chance using the kappa statistic (**κ**), in STATA version 11 SE.

3.1.4 Estimating the total number of published test-treatment RCTs

Citations meeting the inclusion criteria were compared between the two searches using the capture-recapture technique, allowing an estimate of the number missed by both searches to be calculated from the degree of overlap in ascertainment. Relevant articles 'captured' by each search were 'marked' for inclusion, and full bibliographic details compared to identify those that had been captured by both strategies. Multiple publications relating to a single study were individually checked for their capture by either search, but treated as one trial for further analysis. If any of the multiple publications were captured by a search, the study was considered 'found' by that search.

By summarising the number of relevant studies found by each search in a 2x2 contingency table, the number missed by both searches (**x**) can be estimated, and the number of published imaging RCTs ($\mathbf{N_i}$) approximated using the Lincoln-Peterson method (Table 3.3)[204]. The formulae for these two estimates rely on 3 key assumptions regarding the probability of 'catching' an article. First, that the underlying population remains constant between searches. Second, that each article therefore has an equal probability of being found in any given search. And third that the probability of finding a particular article in one search is not influenced by whether or not it is found in the other search; that is to say that each article's probability of ascertainment by both sources is independent. These assumptions imply that the probability of ascertainment by *both* searches (a/**N,** table 3.3) must be equal to the chances of being found by search 1 ([a+c]/**N**), multiplied by the chances of being found by search 2 ([a+b]/**N**).

This gives:

$$\frac{a}{\mathbf{N}} = \frac{a+c}{\mathbf{N}} \times \frac{a+b}{\mathbf{N}}$$

Which can be simplified to:

$$\mathbf{N} = \frac{(a+c)(a+b)}{\mathbf{a}}$$

Given that: $\mathbf{N} = a+b+c+\mathbf{x}$

The two formulae can be rearranged to find **x**:

$$a+b+c+\mathbf{x} = \frac{(a+c)(a+b)}{\mathbf{a}}$$

$$a(a+b+c+\mathbf{x}) = (a+c)(a+b)$$

$$\mathbf{x} = \frac{cb}{\mathbf{a}}$$

In order to extrapolate this to estimate the number of *all* published test-treatment RCTs ($\mathbf{N_t}$), regardless of test type, $\mathbf{N_i}$ was inflated by the proportion of imaging test RCTs found in the generic ATT search.

The calculation of confidence intervals using standard errors has been shown to perform poorly with small or moderate-sized capture-recapture samples, producing symmetrical intervals that tend to underestimate both the lower limit (i.e. the limit is lower in value than the total number of cases actually observed) and higher limit[211–212]. Consequently a 'test-based' approach was used in favour of the asymptotic standard error, due to the likely skewed sampling distribution of **N** [212].

		Search 2		
		Found	Missed	
Search 1	Found	a	c	a + c
	Missed	b	**x**	b + x
		a + b	c + **x**	**N**

Estimated values	**Denotation**	**Calculation**
Unobserved cell:	x	$\boldsymbol{bc/a}$
Total population:	N	$\boldsymbol{(a+b)(a+c)/a}$
Asymptotic variance*:	Var(N)	$\boldsymbol{\frac{(a+b+1)(a+c+1)(bc)}{(a+1)^2(a+2)}}$

Table 3.3: Contingency table summarising numbers of articles ascertained by two searches, with the Lincoln-Peterson estimate for total population size (N) below[204].

Where:

a is the number of relevant studies found by both searches

b is the number of relevant studies found by search 2, but missed by search 1

c is the number of relevant studies found by search 1, but missed by search 2

* Note that the asymptotic standard error was not used to calculate confidence intervals as the study population has a skewed sampling distribution.

For the interval surrounding $\mathbf{N_i}$ the Fisher's exact test of association was selected as fewer than 80% of cells in the contingency table had values of >4 [213]. Having populated the contingency table with all observed values (cells a, b, c), the value of **x** was entered as 0, thus creating a completed table for which Fisher's probability of independence was calculated. This was repeated for increasing values of **x** (incrementals of 1) to create a series of tables with increasing values of $\mathbf{N_i}$, and for each Fisher's probability was again calculated. Test–based 95% intervals included all values of $\mathbf{N_i}$ for which the ensuing probability of independence was adequate (>0.05)[212].

As cell sizes were larger for the comparison of all test types, test-based confidence intervals for $\mathbf{N_t}$ were calculated using the χ^2 test. Values of **x** were again imputed in increasing value until χ^2 exceeded 3.84 (α=0.05, 1-tailed at one degree of freedom) and was rejected. The confidence interval represents the range of values of **x**, and hence $\mathbf{N_t}$,for which the null hypothesis is not rejected.

3.2 Results

3.2.1 Search results

The updated ATT search (2010, issue 2) retrieved 13,495 citations, of which 603 had not been identified in the 2009 search of CENTRAL. In addition to the 128 articles ascertained by the original search, 12 additional test-treatment RCT articles were identified, of which three[214–216] were subsidiary publications of trials already included in the project cohort[(T34,T48,T107)]. Five new articles reported on five test-treatment RCTs not ascertained by the original search[217–221]. All were published in 2007, and four evaluated a relevant imaging test[218–221] (Figure 3.1). In addition, the second reviewer identified three trials ascertained by the original search that had been missed by the author[222–224], two of which evaluated imaging modalities[222,224] though only one of these evaluated one of the target imaging tests[222].

Adding these to the 108 test-treatment RCTs ascertained in chapter 2, a total of 116 test-treatment RCTs were found by the ATT search. These were reported in 139 articles, giving a search yield of 1.0%. Of these, 75 trials (64.7%) evaluated an imaging modality (reported in 89 articles), including 68 trials (58.6%) that evaluated one of the five targeted imaging tests (reported in 83 articles)(Table 3.4).

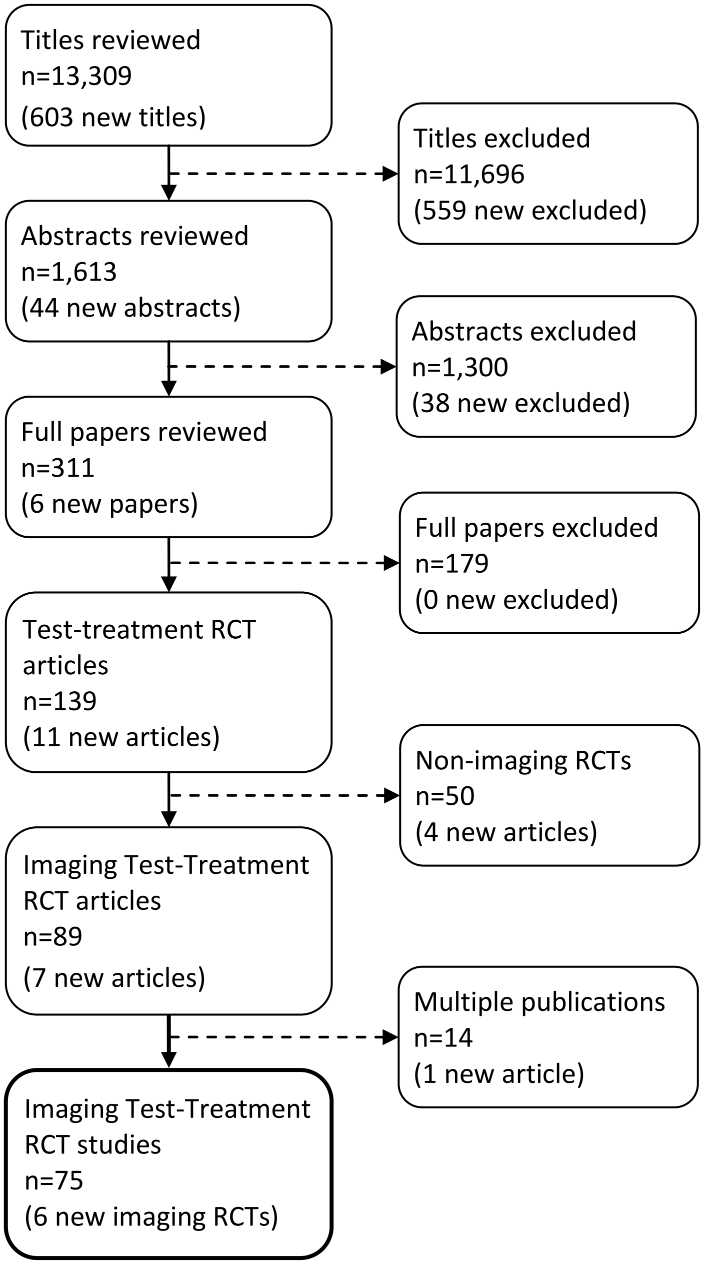


Figure 3.1: Original search (ATT) study selection process for additional records retrieved (CENTRAL 2010, Issue 2)

After removal of 34 duplicates, the IT search retrieved 10,393 unique citations; the study selection process is illustrated in Figure 3.2. A total of 97 relevant articles were identified (0.9%) reporting on 85 individual test-treatment RCTs (Table 3.4). Three RCTs (reported in 4 articles) evaluated non–imaging tests, including biochemical assays(T26,T71) and

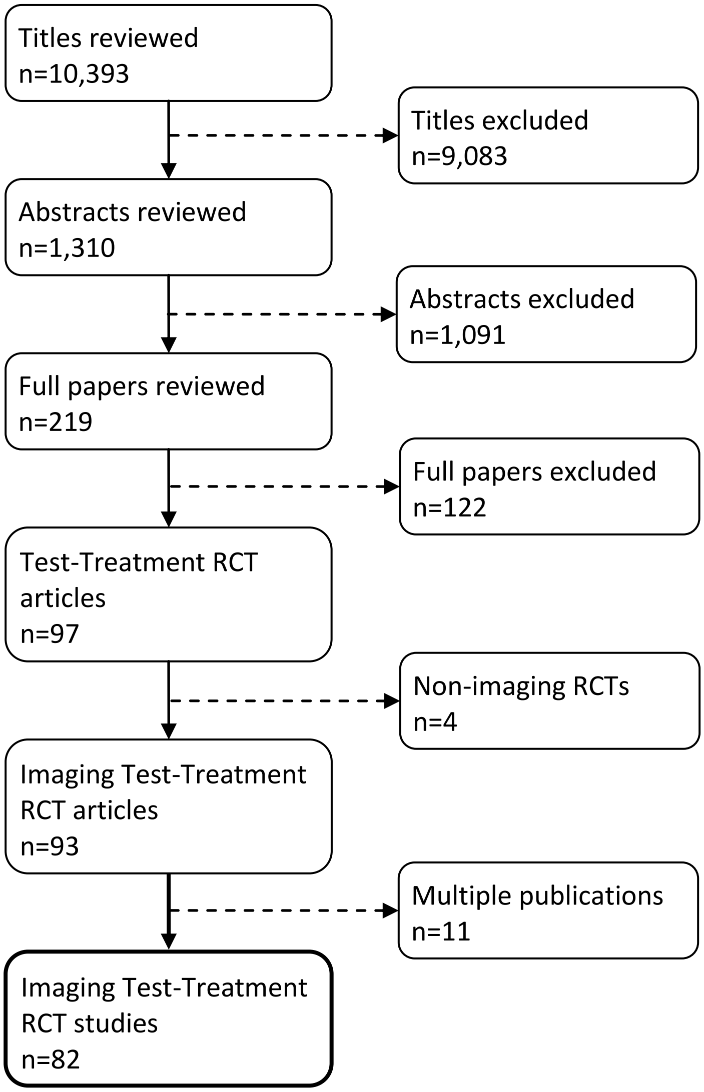


Figure 3.2: Imaging search (IT) study selection process

biopsy$^{(T24)}$. Four others (4 articles) evaluated imaging tests not targeted by the search (X–ray$^{224–225}$, bone scintigraphy$^{(T74)}$, endoscopy$^{(T1)}$). In sum, therefore, the IT strategy ascertained 78 test-treatment RCTs (reported in 89 articles) that evaluated a target imaging modality. Overall the two search strategies identified 133 distinct test-treatment RCTs evaluating any type of diagnostic test, 92 RCTs evaluating any diagnostic imaging modality, and 84 RCTs that assessed one of the five imaging tests specified in the methods (Table 3.4).

	IT Search		ATT Search		Total
Titles screened	***10,393*		*13,495*		***23,688***
Test-Treat RCT:					
articles	97	(0.9%)	139	(1.0%)	**158**
studies	85		116		**133**
Any imaging test-treat RCT:					
articles	93	(0.9%)	89	(0.7%)	**107**
studies	82		75		**92**
Target imaging test-treat RCT[†]:					
articles	89	(0.9%)	83	(0.6%)	**99**
studies	78		68		**84**

Table 3.4: Total number of relevant RCTs found by each strategy. Numbers in parentheses provide search precision, given as proportions of full articles of all titles screened.

* After eliminating 34 duplicates

[†] A subset of all imaging RCTs, these are the 'relevant' imaging trials for the capture–recapture estimate of $\mathbf{N_i}$: ultrasound, endoscopy, MRI, CT, PET.

The author found it more difficult to discern the relevance of titles retrieved by the IT search, and this is reflected in the narrow difference in proportion of abstracts retrieved from each search (IT 12.0%, ATT 12.6%; relative difference 5%) relative to the similar proportion of relevant full paper articles finally included by each strategy (IT 0.9%, ATT 1.0%; relative difference 11%). Overall search precision was very low in both searches, though when considering the identification of topic–specific imaging RCTs the IT strategy precision remained constant (0.6%) while that of the ATT strategy decreased (IT 0.9%, ATT 0.6%) with the reviewer needing to read 111 and 167 titles respectively to identify one relevant imaging test-treatment trial.

3.2.2 Estimated number of published target imaging trials

Of the total 84 topic–specific RCTs identified, 62 were captured by both searches and the IT strategy ascertained 16 trials not picked up by the general methods–term search.

		ATT search		
		Found	Missed	
IT search	Found	62	16	78
	Missed	6	**x**	6 + **x**
		68	16 + **x**	$\mathbf{N}_i$

Number of RCTs:				
Missing	**x**	$6 \times 16/62$	=	1.6
Found in total	$\mathbf{N}_i$	$(68 \times 78)/62$	=	85.6
Exact 95% CI			=	84, 89

Table 3.5: Calculation for estimating the total number of test-treatment trials evaluating an imaging test published between 2004 and 2007.

Although the ATT strategy yielded fewer relevant imaging RCTs (68 vs. 78), it did identify 6 trials that were undetected by the targeted, content-specific strategy. Table 3.5 summarises the overlap in ascertainment of included imaging trials. Since trials can only be present in whole numbers, the number missed by both searches (**x**) is calculated to be 2 (rounded up from 1.6), providing an estimate of 86 (95% CI: 84, 89) for the population of these trials published between 2004 and 2007, and indexed in CENTRAL.

3.2.3 Estimated number of all published test-treatment trials

Of the 116 trials found by the ATT search, 58.6% (n=68) evaluated a target imaging trial. Assuming this proportion is a true reflection of the total frequency of these trials published between 2004 and 2007, the total number of all test-treatment RCTs published in the same timeframe can be extrapolated by inflating the estimate for $\mathbf{N}_i$ by 58.6%. This gives 85.6/0.586, or 146, which provides an estimate of 37 trials (146/4) published per year.

		ATT Search		
		Found	Missed	
IT Search	Found	68	17	85
	Missed	48	**x**	48 + **x**
		116	17 + **x**	**N**

Number of RCTs:				
Missing	**x**	$\mathbf{48 \times 17/68}$	=	12.0
Found in total	$\mathbf{N_i}$	$\mathbf{(116 \times 85)/68}$	=	145.0
Exact 95% CI			=	137, 157

Table 3.6: Calculation for estimating the total number of test-treatment RCTs, regardless of test type, published between 2004 and 2007.

Alternatively, $\mathbf{N_t}$ can be estimated directly using capture–recapture analysis of all trials identified by both searches. Table 3.6 illustrates the resulting contingency table, demonstrating that of the total 133 RCTs 68 were ascertained by both searches, 48 were found by ATT but not by IT, while 17 ascertained by IT were not picked up in the ATT strategy. Accordingly, **x** is estimated as 12 and $\mathbf{N_t}$ as 145 (95% CI: 137, 157), which provides an estimate of 36 trials (145/4) published per year. This is a very similar figure to that achieved through extrapolation, which would seem to lend support to the estimate.

3.2.4 Inter−observer agreement

Overall the author identified 20 test-treatment RCTs in the random 10% sample (n=2591), of which 5 were missed by the second reviewer (Table 3.7). The second reviewer identified 8 additional potentially relevant studies, of which 3 [222–224] were agreed to satisfy all inclusion criteria at consensus bringing the total number of RCT articles to 23. All

2nd Screener		Author			A_{Obs}	A_{Exp}	κ	95% CI
		Include	Exclude	Total				
IT Search	Include	5	0	5	1.00	0.99	0.91	0.77, 1.00
	Exclude	1	1184	1185				
	Total	6	1184	1190				
ATT Search	Include	10	3	13	1.00	0.98	0.74	0.56, 0.93
	Exclude	4	1384	1388				
	Total	14	1387	1401				
Both Searches	Include	15	3	18	1.00	0.99	0.68	0.64, 0.93
	Exclude	5	2568	2573				
	Total	20	2571	2591				

Table 3.7: Agreement between two screeners in the identification of test-treatment RCTs. Selection was based on review of full articles.

studies missed by both reviewers were originally eliminated on the basis of title only. Considering the number of citations that were double–screened, observer variability was minimal with observed agreement in over 99% across the whole 10% sample, and agreement beyond chance calculated to be substantial overall[226].

Agreement differed according to search strategy; with only one disagreement[(T29)] concurrence for content–specific citations was near–perfect. Agreement in identifying ATT strategy trials was substantial, however more studies were missed by both reviewers possibly due to the need to identify diagnostic RCTs from a very large volume of treatment RCTs and diagnostic accuracy evaluations.

Comparing individual searches and overall ascertainment, observed agreement was higher than the resulting **κ** value. It has been demonstrated that **κ** is influenced by the distribution of observations in a 2 x 2 table, such that if cells of agreement (e.g. include – include) or disagreement (e.g. include – exclude) are asymmetrical **κ** becomes distorted[227]. In this study, very low prevalence (1.1% overall, hence skewed agreement) and skewed

Search for all RCTs indexed in CENTRAL		Hits
#1	RCT in all text	87,794
#2	"Randomized Controlled Trial" in Publication Type	
#3	(#1 or #2), from 2004 to 2007	

Table 3.8: Approximate search for the number of all RCTs indexed in CENTRAL at the time the IT and updated ATT searches were conducted (Issue 2 2010, Wiley InterScience searched 10 November 2010) – RCT text words across all fields and MeSH terms limited to publication years 2004 – 2007.

disagreement have constrained **κ**, which is evidently low for the ATT search despite comparably high levels of observed agreement.

3.3 Discussion

This simple study finds that test-treatment RCTs are rare, with only approximately 36–37 test-treatment RCTs published per year between 2004 and 2007. This is astoundingly low when compared to the approximate rate of 21,949 per year of all RCTs indexed in CENTRAL (Table 3.8).

3.3.1 Validity of the population estimates

Extrapolation was used to estimate the total number of test-treatment trials, however the validity of this method relies on imaging trials being as likely to include methodological terms and descriptors as trials of other test technologies. Though this could not be verified in the current study, it is nonetheless somewhat reassuring that the capture–recapture estimate arrived at an almost identical total population estimate. However, the estimates may not be valid if the assumptions inherent in the capture–recapture method do not hold for test-treatment RCT publications. As set out in the methods to this chapter, three key assumptions should be examined.

First, the method assumes that the underlying population is ‘closed’, namely that there are no new additions or subtractions to it between searches. This requirement was met by ensuring that both searches were run concurrently in the same version of CENTRAL.

Second, all articles should have the same probability of capture in a search. It is possible that certain subgroups of test-treatment RCTs have traits that predispose them to ascertainment relative to others. Particular clinical specialties, types of diagnostic test or even the journal or year of publication may tend to be more accurately indexed, or better reported, making them more likely to be detected than other test-treatment RCTs. For example, if imaging RCTs were more likely to contain diagnostic methodology descriptors than non–imaging RCTs, they would be more ‘catchable’; hence the proportion of imaging tests observed in the ATT search would be overestimated. This would mean that the method has produced an underestimate of the true total test-treatment RCT population. It is difficult to evaluate the impact such ‘variable catchability’[204] has had on the final population estimate, and further cross-sectional work would be needed to examine the nature of associations between indexing quality and clinical specialities, journals and test types through time.

Third, the capture-recapture method assumes that the probability of finding a particular article in the IT search should be *independent* of the chances that it is picked up by the ATT search. The approach used in this study aimed to satisfy this requirement by employing two separate ascertainment concepts, general diagnostic and methodological terms in the ATT strategy and test names in the IT strategy. The possibility remains, however, that an RCT appropriately indexed by methodological terms will also be well indexed by content-specific terms. If this were true then its chances of retrieval by the IT strategy would be *positively dependent* on the chances of retrieval by the ATT strategy, and so the estimate would underestimate the true population. Nonetheless, the searches identified articles published during a time period in which the CONSORT guidelines have been routinely applied in many journals; thus to some extent the dangers of dependency have been mitigated.

3.3.2 Reliability of the population estimates

The most likely source of error in the study data concerns the accuracy with which articles were classified as test-treatment RCTs. As argued in chapter 2, the low search precision and ensuing high 'number needed to read' could have impacted on the accuracy of article identification, since with over 23,000 records it is likely that some titles were mistakenly excluded due to decreased attentiveness during long periods of screening. Any such misclassification would cause an underestimate.

Indeed, though inter–observer agreement was strong, the second reviewer identified three studies within the 10.8% random sample (2,591/23,888) that were missed by the author. One could therefore extrapolate that 28 test-treatment RCTs ascertained by the searches were missed in total (3/0.108). Extrapolating this underestimate across the whole sample would serve to inflate the estimated number of test-treatment trials published per year by 17.5% to 42–43 RCTs. This would be confirmed more robustly with a second screen of all records, which unfortunately was not possible due to time constraints.

3.3.3 Implications for finding test-treatment RCTs

Test-treatment RCTs were very difficult to find, a discovery carrying implications for the future ascertainment of these studies. Building effective strategies for identifying test-treatment RCTs presents several challenges that will need attention if future reviews are to ascertain the literature on a single diagnostic topic comprehensively.

Ascertainment is very resource-intensive requiring many thousands of records to be checked for yields of 0.9–1%, which risks detrimentally affecting the accuracy of the screening process. Of course this study does not fully reflect that of subject-specific systematic reviews, where the search concepts and structure would be more developed and where the addition of disease–specific terms could help to increase the precision of search results.

Both searches missed relevant imaging trials, however most disturbing are the number missed by the IT strategy, approximately 9% (8/86) of relevant imaging trials, suggesting that diagnostic content-specific terms are applied inconsistently across Medline, Embase and the other sources included in CENTRAL. The ATT search was less effective at identifying targeted imaging RCTs missing approximately one–fifth (18/86) of studies, though it did capture others missed by the IT strategy emphasising the importance of including methodological terms to maximise sensitivity. Nevertheless, these results also testify to the inconsistent use of methodological and general diagnostic descriptors.

These findings imply that future searches will not be able to guarantee the ascertainment of all targeted test-treatment RCTs for a systematic review, a failure which could result in biased results and loss of precision. These issues are likely to be resolved by standardising the application of existing content-specific diagnostic terms and introducing methodological terms specific to the evaluation of diagnostic tests. A methodological indexing term for test accuracy studies was recently introduced by Embase[170]; though a commendable improvement to the previous situation, efforts to extend such terms will be required if the detection of test-treatment RCTs is to improve. As with any study design, test-treatment RCTs will stand the greatest chance of being ascertained if they are identifiable as such.

Further research is clearly needed to characterise the precision and sensitivity of test-treatment RCT search strategies, and to develop methodological and content filters that maximise sensitivity for the least losses in precision as has been done for the field of diagnostic accuracy.

3.4 Conclusions

This study confirms previous claims that RCTs evaluating the impact of diagnostic tests on patient outcomes are rare. It is estimated that 36–37 test-treatment RCTs were published every year between 2004 and 2007. Even if the methods used have produced a slight underestimate, the true figure is likely to be just a tiny fraction of the total number of RCTs that are published in medical journals every year. Based on these data, it is therefore unlikely that RCT evidence of the patient health impact of diagnostic tests will be available to inform guidelines on the use of diagnostic tests in many settings. Guidelines will therefore frequently be based on lower grade evidence, which may risk erroneous conclusions.

Despite substantial overall agreement, the independent check of accuracy in identifying these studies found that three test-treatment RCTs ascertained in the original search were missed by the author. Since only 10% this search were checked, it is possible that up to 28 relevant trials have therefore not been included in subsequent analyses. If, as seems, likely, these were missed due to unclear description, any conclusions on the quality of reporting (presented in chapter 5) may therefore be generous estimates of the current situation. These implications are considered further in chapter 5 and the general discussion and conclusions of the thesis (chapter 9).

The next chapters outline the clinical context of test-treatment RCTs included by the author (chapter 4) and investigate their reporting quality (chapter 5) and methodological quality (chapter 6) in order to address whether RCT evidence, when available, is likely to be reliable and informative.

4

Trials of Test-Treat Strategies:

Characteristics of included trials

This chapter aims to characterise the diagnostic settings in which test-treatment RCTs have been executed to completion. It does so by providing a descriptive overview of the published trials identified in chapter 2, surveying the diagnostic questions these studies have answered and presenting an insight into their methodological approaches.

Methodological research into assessments of diagnostic accuracy has shown how estimates of test performance are affected by many elements of the clinical context, including the case–mix of study populations, how tests are carried out and how their results are interpreted[66], the care setting, role of the new test in the existing pathway, prior investigations and practitioner experience[60,74]. These factors are also likely to influence the clinical effectiveness of diagnostic tests. Moreover, the need to compare test-treatment strategies in order to achieve this means that the clinical context is potentially much more variable and complicated.

The thesis has so far shown that these trials are rare (chapter 3), however in the absence of previous methodological reviews of test-treatment RCTs, very little is known regarding the sorts of diagnostic problems these designs have been used to evaluate, and the methods employed to answer them. Consequently, this chapter was designed to identify the settings, tests and effectiveness questions that RCTs have successfully been used to evaluate to completion.

Accordingly, the following report describes the key clinical and methodological characteristics of these studies so as to summarise the spectrum of interventions evaluated, their diagnostic contexts and broad methodological approaches. This characterisation provides a first indication of the range of clinical questions that have been found to underpin diagnostic comparative effectiveness research.

4.1 Methods

To discern the particulars of the diagnostic questions that published test-treatment trials have addressed, information was extracted from each trial in the project cohort to answer the following four questions (Table 4.1):

1. What diagnostic tests have been evaluated?
2. In which clinical settings are these tests used?
3. What are the tests being used for?
4. What types of trial design have been used?

In view of the ample array of diagnostic technologies used in healthcare, the first question was designed to assess the breadth of test-treat topics, and determine the nature of any disparities in how frequently the various technologies have been evaluated for patient benefit. Questions two and three sought to elucidate the context in which these diagnostic effectiveness evaluations were conducted, characterising the clinical pathways involved and outlining the diagnostic questions they address. The objective was to establish where in the healthcare process these tests were evaluated, and consequently also to examine the degree of patient group selection and how narrowly focussed the diagnostic questions were. The final question provides a methodological orientation of included studies by summarising characteristics of study method and describing variations therein.

4.1.1 Item generation

The aim at this stage was to provide a descriptive account of included test-treatment RCTs, and not to review quality of reporting or methods. In the absence of existing tools or previous studies attempting to characterise test-treat interventions, information determined to be important to the description of diagnostic accuracy studies and RCTs was extracted.

Thematic questions:		**Information extracted:**	
1	**What diagnostic tests have been evaluated?**		
		1a	Experimental test *†
		1b	Control test *†
2	**In what clinical pathway is the evaluation taking place?**		
		2a	Medical speciality
		2b	Country †
		2c	Care service *†
3	**What are the tests being used for?**		
		3a	Patient group *†
		3b	Target condition *
		3c	Prior tests
		3d	Management decision
4	**What types of trial design have been used?**		
		4a	RCT design †
		4b	Number of study groups †
		4c	Test comparison *†

Table 4.1: Items generated to characterise the diagnostic interventions in their clinical setting.
* denotes item present in STARD checklist[72]
† denotes item present in CONSORT checklist[121]

The STARD[72] and CONSORT[121] checklists were chosen as each was developed from a comprehensive systematic analysis of existing publications on the conduct and reporting of diagnostic studies and RCTs respectively, in addition to Delphi consensus by a panel of experts[138,228]. Where this information was insufficient to describe the complex interventions and clinical setting fully (noted in table 4.1), additional rubrics were identified. This process of identification was iterative; trial reports were examined several times allowing pertinent items to be added and modified during extraction.

4.1.2 Data Extraction

Data on 12 items were extracted by the author to a relational database (Microsoft Office Access 2007), according to the definitions outlined in table 4.2. Due to the widely varying

Item extracted	Description
Country	Countries contributing at least one investigative centre to the trial
Medical specialty	The medical department chiefly responsible for managing patient care
Care service	Healthcare service that testing will take place in, defined as (DH): ***Emergency Care:*** Urgent healthcare services available to those who need medical advice, diagnosis and/or treatment quickly and unexpectedly. ***Primary Care:*** General healthcare provided by GP practices, dental practices, community pharmacies and high street optometrists, that may involve onward patient referral to emergency, secondary or tertiary services. ***Secondary Care:*** Pre-arranged, non-emergency care provided by medical specialists in hospitals or clinics. Patients are usually referred from primary care professionals, such as a GP. ***Tertiary Care:*** Highly specialised consultative care, usually on referral from primary or secondary care services, provided by centres with personnel and facilities for special investigation and treatment.
Control test	The test specifically referred to as the comparator, the current care standard, or common clinical practice. In a minority of cases, the control arm was identified implicitly by the manner in which study results were discussed.
Experimental test	The test specifically referred to as the new test to be introduced under evaluation.
Comparison type	The prospective role of the experimental test in the existing diagnostic pathway[11]: ***Replacement:*** total replacement of the control test by the experimental test ***Triage:*** introducing the experimental test to select which patients receive the control test ***Add-on:*** addition of the experimental test alongside the control test
Patient group	The target study group, defined as those eligible for randomisation.
Target condition	The disease or condition to be confirmed through the present episode of testing. Related to the purpose of diagnosis, this could also constitute the identification of risk factors, stage or grade of known disease or aetiology of known condition.
Prior Tests	The tests used to manage the patient's current condition, prior to enrolment in the trial. A current condition was defined as a single pathological entity, that may have received prior treatment. Disease recurrence was included under this definition, comorbid conditions were not.
Management decision	The diagnostic and treatment decisions arrived at through testing
RCT design	Design structure and point of randomisation
Number of study groups	The number of intervention groups patients were initially randomised between.

Table 4.2: Information extracted to characterise test-treatment RCTs.

organisation of healthcare structures across the world, a meaningful comparison of care settings would be impaired using local definitions. Accordingly, care settings for each trial were extracted using the definitions given by the UK Department of Health[229], regardless of the country of origin.

4.1.3 Analysis

Once extraction was complete, data were exported to Microsoft Excel 2007. Data were analysed using an inductive method to create an integrative synthesis[230]. For each item descriptions extracted from all trials were examined together and common themes identified, from which a series of categories were generated that could characterise the range of variation observed across the included studies. This was often performed with reference to related items to ensure that the author's categorisation for each trial was consistent with the meaning of the study.

The analysis presented below is intended to be descriptive, and while it provides frequencies with which each characteristic was observed in the cohort, the review aims to furnish the reader with a more qualitative description[230] of the range of diagnostic questions addressed by test-treatment trials included in the project cohort.

4.2 Results

As reported in chapter 2, the project cohort consists of 108 individual test-treatment RCTs that evaluate the patient health impact of diagnostic tests, after a phase of treatment. The three additional trials that were missed by the author, and identified by the second reviewer in chapter 3, are not included in the project cohort. Citations of trial reports are denoted separately from other references, given as numbers preceded by the letter 'T'.

Country of Study	n	(%)
Western world*	2	(2)
Worldwide†	4	(4)
Argentina	1	(1)
Australia	6	(6)
Belgium	1	(1)
Brazil	1	(1)
Canada‡	9	(8)
China	3	(3)
Czech Republic	1	(1)
Denmark	3	(3)
France	4	(4)
Germany	1	(1)
India	1	(1)
Iran	1	(1)
Israel	1	(1)
Italy	4	(4)
Japan	2	(2)
Indonesia	1	(1)
Netherlands	16	(15)
Poland	1	(1)
Portugal§	1	(1)
Spain§**	5	(5)
Sweden	2	(2)
Switzerland	4	(4)
UK**	20	(19)
USA‡	18	(17)
Total	**108**	

Table 4.3: **Location of clinical centres under study.**
* North America and Europe only
† Western world, Asia (*India, Pakistan*), Australasia, Africa (*Egypt, S. Africa*), South America (*Mexico, Venezuela*)
‡ 3 studies carried out in Canada and USA
§ 1 study carried out in Portugal and Spain
** 1 study carried out in Spain and UK

4.2.1 Clinical setting

Studies were conducted in 30 countries located across the globe (Table 4.3). Most (97, 90%) took place in a single country, while trials evaluating international test-treat practice included centres from between 2 and 11 countries. Trials were more commonly conducted

Clinical Specialty	Care Setting						
	Emergency	Primary	Secondary	Tertiary	Multiple	**Total**	**(%)**
Cardiovascular Medicine	9	1	17	5	4	36	(33)
Obstetrics and Gynaecology	0	0	8	10	0	18	(17)
Gastroenterology	0	5	9	0	0	14	(13)
Orthopaedics	1	2	6	1	0	10	(9)
Oncology	0	0	0	6	0	6	(6)
Multiple*	0	1	1	1	1	4	(4)
Neurology	1	0	1	0	1	3	(3)
Respiratory	0	1	1	1	0	3	(3)
Embryology	0	0	0	2	0	2	(2)
Ophthalmology	0	0	0	2	0	2	(2)
Otolaryngology	0	0	0	2	0	2	(2)
Psychiatry	0	1	1	0	0	2	(2)
Urology	0	1	0	1	0	2	(2)
Emergency Medicine	1	0	0	0	0	1	(1)
General Medicine	0	0	0	0	1	1	(1)
Geriatrics	0	0	1	0	0	1	(1)
Infectious diseases	0	0	1	0	0	1	(1)
Endocrinology	0	0	0	0	0	0	(0)
Neonatology	0	0	0	0	0	0	(0)
Total	**12**	**12**	**46**	**31**	**7**	**108**	

Table 4.4: Range of clinical departments and care settings the test-treatment trials were conducted in.

***** 1 study covered 7 specialities (including Endocrinology and Neonatology), and the remaining three studies covered two specialities

in North America and Europe, the UK being the most common single country of origin. Consistent with this distribution and the study selection criteria, the majority of trials were conducted in English speaking countries (56, 52%).

Test-treatment interventions were evaluated across eighteen medical specialties, though the distribution was not even (Table 4.4). Cardiovascular settings alone accounted for one in three trials. Almost 75% of studies were conducted in one of four departments: cardiovascular, obstetrics and gynaecology, gastroenterology or orthopaedics, while fewer than one in three trials were conducted in the remaining 14 specialties.

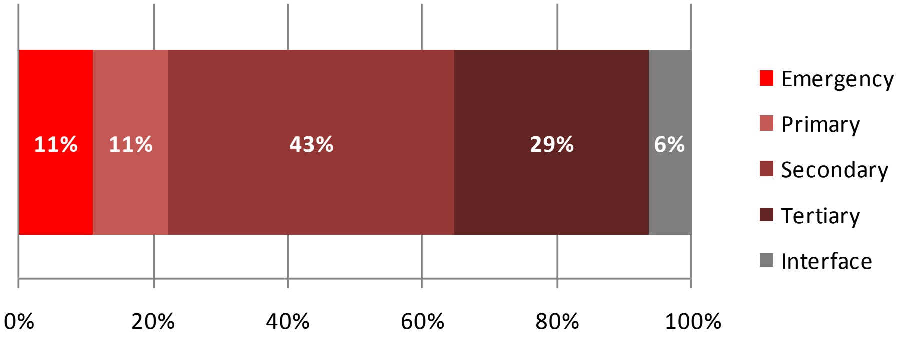


Figure 4.1: Proportion of trials conducted in each care setting

Nearly half of all studies managed patients in a secondary care setting, namely in a hospital following referral from community or emergency services (Figure 4.1), while a considerable proportion of the remainder assessed tests used in specialised tertiary clinics. Seven studies took place at the interface of two to three services, either because the treatment options varied according to the diagnosis provided by the intervention (n=5), or because management in different care settings was integral to the comparison itself (n=2). For example one trial set out to assess whether ED patients suspected to have suffered a transient ischaemic attack (TIA) would experience a reduction in disease-event rates by immediately undergoing a comprehensive testing protocol in the ED, rather than being admitted as inpatients to receive a slower and more ad-hoc process of diagnosis(T86). In this example, both the types of tests administered and the organisation of testing were the subject of evaluation.

4.2.2 Tests evaluated

Most trials (97, 90%) compared two testing strategies, though three- (n=6), four- (n=4) and five-group (n=1) comparisons were also encountered. Consequently, 224 intervention

Test Genre	Control	(%)	Experimental	(%)
Biochemical	10	(9)	26	(21)
Biopsy	2	(2)	3	(2)
Clinical assessment	15	(14)	13	(10)
Electrophysiological	7	(6)	11	(9)
Imaging (radiology)	21	(19)	41	(33)
Imaging (endoscopy)	15	(14)	11	(9)
Imaging (total)	*36*	*(33)*	*52*	*(42)*
Telemedicine	2	(2)	2	(2)
Multiple test interventions*	9	(8)	11	(9)
No Test	18	(16)	5	(4)
Not reported	11	(10)	1	(1)
Total	**†110**		**†124**	

Table 4.5a: Test types evaluated in test-treatment RCTs.
*detailed in Table 4.5b
†Note that the denominators refer to the number of interventions not the number of trials

arms were compared, 110 receiving a control test and 124 receiving an experimental test (see table 4.2 for definitions). A broad range of tests was evaluated, covering nine types of technology. Imaging tests were by far the most common subject of evaluation, with 52 (42%) experimental interventions assessed in a total of 50 (49%) different test-treatment RCTs (Table 4.5a). Endoscopies comprised a fifth of these studies, and were carried out for a variety of investigations including the upper or lower gastrointestinal tract(T13,T30,T77,T88,T105), joint space(T35), bladder(T1,T90), uterus(T104), stomach(T9,T95) and abdominal cavity(T27). The majority however evaluated new radiological techniques involving at most minimally-invasive procedures to produce images of anatomical structures including X-ray(T6,T14,T20,T28,T40,T73,T99,T108), magnetic resonance imaging (MRI) (T4,T5,T10,T19,T47,T70,T92,T103,T107), computed tomography (CT)(T12,T68,T82,T85,T91), positron emission tomography (PET)(T52,T84,T96), single positron emission computed tomography (SPECT) (T32,T83,T92) or ultrasound (US)(T8,T11,T15,T22,T38,T58,69,T72,T76,T78,T89,T92,T102).

Four studies evaluated the effectiveness of not giving any test to the patient group (‘No test’); three of these assessed the value of eliminating further treatment(T18,T45,T63) (see

Multiple test interventions	Control	(%)	Experimental	(%)
Biochemical + Clinical assessment	1	(10)	4	(40)
Biochemical + Electrophysiology	2	(20)	0	(0)
Biochemical + Imaging (radiology)	1	(10)	2	(20)
Biochemical + Imaging (endoscopy + radiology)	0	(0)	1	(10)
Biopsy + Imaging (radiology)	1	(10)	0	(0)
Clinical assessment, Electrophysiology, Imaging (radiology)	2	(20)	0	(0)
No Test + Imaging (radiology)	1	(10)	0	(0)
Unit of care	1	(10)	4	(40)
Total	***9**		***10**	

Table 4.5b: Composition of multiple test interventions.

*Note that the denominators refer to the number of interventions not the number of studies

below) and one the benefits of treating all patients without prior testing(T71). Approximately one in five trials evaluated biochemical assays, designed to measure a variety of biological substances including blood serum protein levels (n=6), amino acids (n=2), or microbiological cultures (n=7). Trials did not solely focus on assessing new or technologically advanced tests, and 13 comparisons investigated a new role for clinical assessment. Signs and symptoms were for example used in the implementation of 'watch-and-wait' policies, in order to restrict the use of more technologically advanced tests to patients who fail to respond to treatment(T27,T33,T36,T65,T71). Tests measuring electrophysiological responses accounted for approximately 10% of experimental interventions(T2,T39,T41,T54,T56,T61,T67,T78,T80,T101), all but two(T39,T54) of which entailed measurement of cardiovascular properties. Eleven trials evaluated protocols containing multiple new tests (Table 4.5b), including four that assessed the impact of a new diagnostic unit for the provision of specialised diagnosis and management in an earlier care setting(T16,T29,T48,T53).

Non/minimally-invasive imaging modalities were almost twice as likely to be evaluated as experimental interventions than control interventions (33% vs. 19%, Figure 4.2), and the

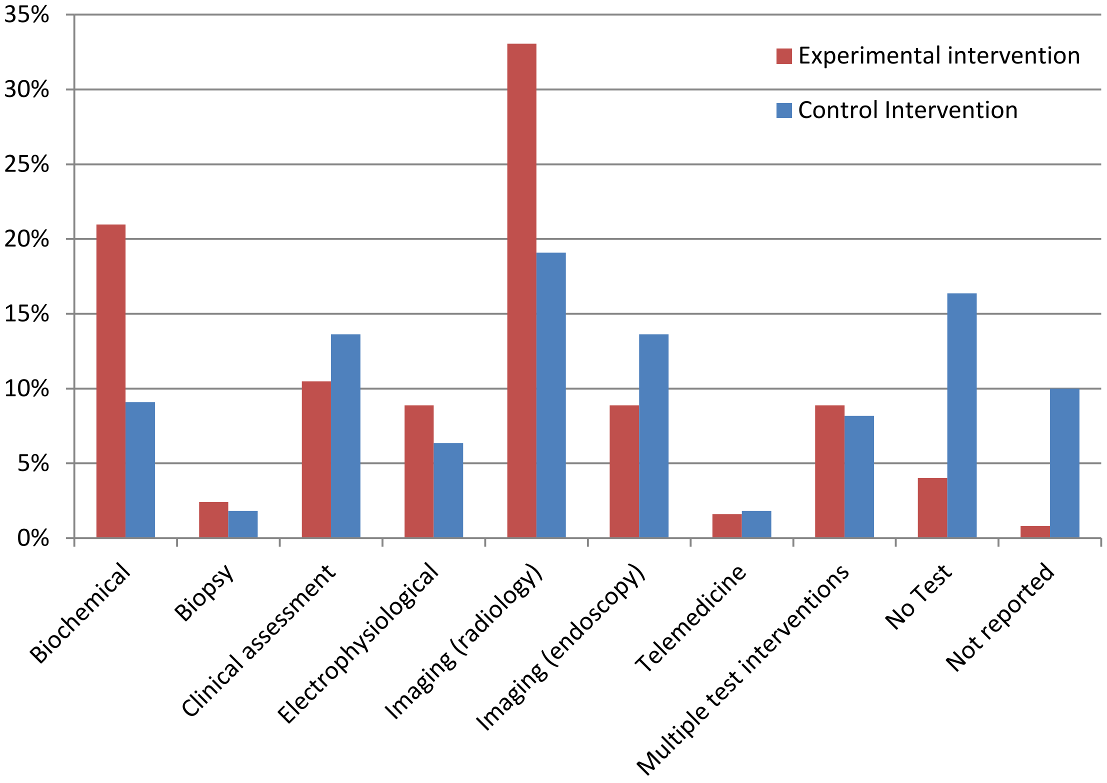


Figure 4.2: Distribution of experimental and comparator intervention test types.

same was also true of biochemical assays (21% vs. 9%), which may reflect a general increase in the use of these two technologies through time. Conversely, ‘no test’ strategies were more commonly used as comparators (16% vs. 4%), signalling a trend towards the introduction of new tests into healthcare management.

Studies were also less likely to report which tests formed the control intervention, 11 trials (10%) stating only that ‘standard care’ was used(T10,T16,T32,T43,T45,T56,T69,T72,T86,T97,T100) compared to one trial failing to report which tests were used as part of the experimental intervention(T86).

Experimental Test Genre	Triage	(%)	Replacement	(%)	Additional	(%)	Total	(%)
Biochemical	2	(10)	14	(19)	10	(34)	**26**	(21)
Biopsy	1	(5)	2	(3)	0	(0)	**3**	(2)
Clinical assessment	5	(24)	7	(9)	1	(3)	**13**	(10)
Electrophysiological	0	(0)	6	(8)	5	(17)	**11**	(9)
Imaging (radiology)	8	(38)	24	(32)	9	(31)	**41**	(33)
Imaging (endoscopy)	0	(0)	8	(11)	3	(10)	**11**	(9)
Telemedicine	0	(0)	2	(3)	0	(0)	**2**	(2)
Multiple test interventions	5	(24)	5	(7)	1	(3)	**11**	(9)
No Test	0	(0)	5	(7)	0	(0)	**5**	(4)
Not reported	0	(0)	1	(1)	0	(0)	**1**	(1)
Total	**21**		**74**		**29**		**124**	

Table 4.6: Experimental test types and how they substitute the existing diagnostic pathway.

4.2.3 Comparisons made

Control interventions generally constituted current practice (101, 94%), although this was unclear in five studies. Two trials did not compare existing practice to new, one evaluating two competing variations of a standard care protocol(T88), while another directly compared the health impact of two recent technological developments(T19).

Two thirds of studies evaluated a diagnostic strategy designed to replace the existing approach (74, 60%), while 23% (29/124) assessed the impact of adding tests and a further 17% (21/124) the addition of triage tests (Table 4.6). Seven distinct types of replacement comparison were observed, reflecting variations in the number, order and nature of the experimental test, relative to the control test, in an existing pathway. Examples of each are illustrated in table 4.7. Approximately half (47%) were typical examples of a new test replacing the current clinical standard, studies also evaluated the benefit of using existing tests in a different order (‘Strategy replacement’, 4%) or administered in different healthcare settings (‘Delivery change’, 3%); of evaluating technologically improved existing tests (‘Test updated’, 5%); of introducing a test where previously there had been no patient selection for subsequent treatment (‘Testing introduced’, 24%), or conversely eliminating

Replacement comparisons	Description	Example
1. Standard replacement (n = 35)	The direct replacement of an existing test by a new test	Is CTA in place of DSA better for detecting vascular blockages in patients with known PAD(T12)
2. Strategy replacement (n = 3)	Replacement of the order of existing tests	Should MRI be given immediately to all patients followed by an orthopaedic consultant, or should all patients receive an orthopaedic consultation first?(T107)
3. Testing introduced (n = 18)	New episode of diagnosis introduced into pathway	Should patients with idiopathic VTE be screened for malignancies using a battery of tests?(T42)
4. Testing eliminated (n = 5)	Existing episode of diagnosis eliminated from pathway	Is it safe to discharge patients with suspected VTE on the basis of prior tests, rather than perform further US?(T63)
5. Triage eliminated (n = 7)	Selective application of old test replaced by routine use in all	Should CA be given to all ACS patients immediately, instead of on the development of a worsening condition during admission?(T99)
6. Test updated (n = 4)	Technological development of the existing test	Would patients with a bacterial infection benefit if physicians received microbiological identification and susceptibility results more quickly?(T23)
7. Delivery change (n = 2)	Change in how existing test is administered	Will there be any change to patient health if upper GI endoscopy/FS is nurse-led rather than doctor-led?(T77)

Table 4.7: Examples of replacement comparisons.

CTA – Computed tomographic angiography; DSA - Digital subtraction angiography; PAD – Peripheral arterial disease; US – Ultrasonography; CA – Coronary angiography; ACS – Acute coronary syndrome; GI – Gastrointestinal; FS – Flexible sigmoidoscopy

an existing episode of diagnosis from the pathway ('Testing eliminated', 7%); or lastly of eliminating pre-selection for the use of an existing test ('Triage eliminated', 9%).

In contrast, add-on and triage comparisons were relatively uniform. Experimental tests were added to the same part of the existing test-treat pathway in order to evaluate the benefit of extra information to diagnostic decision-making. For example focussed

Extent of prior testing category		n	(%)
Unselected:	No prior testing – first point of healthcare contact for current complaint	5	(5)
Limited selection:	Limited prior testing – one or two prior tests constituting a basic assessment of signs and symptoms	42	(39)
Narrow selection:	Moderate prior testing – several preceding tests resulting in a suspected differential diagnosis	17	(16)
Considerable selection:	Extensive testing already conducted, generally indicating patient has received a diagnosis and treatment plan prior to entering the study	43	(40)

Table 4.8: Apparent selection of patient populations taking into account the extent of prior testing received.

ultrasound was provided to patients with suspected torso trauma in addition to the standard battery of tests in order to reduce complications of delayed treatment(T72), while transvaginal US and fetal fibronectin measurements were added to the standard speculum examination to assess the presence of preterm labour in high risk patients(T94).

Triage tests were employed either to restrict the use of an existing invasive test, such as using MRI to determine the need for arthroscopy in knee injury patients(T47), or to otherwise reduce diagnostic resource use to a subset of patients, for example restricting the use of CT in suspected appendicitis patients to those with continuing worsening of symptoms(T60).

4.2.4 Extent of prior testing

The degree of patient selection in trial populations was observed to fall into four broad categories, reflecting the extent of previous testing (Table 4.8). The most common reason for conducting diagnostic clinical effectiveness trials appeared to lay in discerning the impact of further testing, either in already highly selected populations (43, 40%), for example evaluating blue light cystoscopy to more precisely locate known bladder cancer tumours for resection(T1), or in patient groups who had been referred from primary care for

Diagnostic purpose of intervention	Total	(%)
Various	1	(1)
Determine risk of disease progression/future event	4	(4)
Rule out disease	13	(12)
Narrow the differential diagnosis	16	(15)
Screen for concomitant disease	21	(19)
Determine extent of disease/disease characteristics	23	(21)
Confirm a suspected diagnosis	30	(28)
Total	**108**	**(100)**

Table 4.9: Purpose of diagnostic investigations in included trials.

further examination (42, 39%), such as elucidating a differential diagnosis in patients with dyspepsia(T44).

Only five studies (5%) were interested in evaluating tests at the beginning of a pathway, for example diagnosing urinary tract infections at first presentation of symptoms(T34), while the remaining 17 studies (16%) evaluated mixed patient groups with some existing diagnostic information, for instance to determine the cause of undifferentiated chest pain in ED patients with indeterminate ECG(T48).

4.2.5 Purpose of test-treat interventions

Testing was found to serve six diagnostic functions (Table 4.9). Diagnostic interventions most often sought to confirm a diagnosis suspected on the basis of prior tests, for example to confirm the presence of clinically suspected hip dysplasia in neonates using US(T15). Tests were also commonly evaluated earlier in the management pathway, to rule out disease for example to exclude the possibility of coronary artery disease in patients presenting with acute chest pain(T82), or to contribute to a differential diagnosis in patients with more generalised symptoms or indeterminate signs, such as refining the potential cause of unexplained syncope(T53).

Nearly half the trials (48, 44%) evaluated tests in patients with known disease, most often to establish the extent or characteristics of disease, for example to stage breast cancer(T24), locate the blocked vessel in patients with peripheral artery disease(T70), or identify the organism causing infection in pneumonia inpatients(T59), but also occasionally to determine a patient's short-term *risk* of disease progression, for example the likely progression to a full stroke in TIA patients(T86). A less expected though frequent finding was the use of tests for 'opportunistic screening', that is checking for a possible concomitant condition in patients at risk due to underlying disease. For instance imaging patients with idiopathic DVT to check for related occult malignancies(T42), or screening embryos for chromosomal abnormalities in women of advanced maternal age scheduled for IVF(T98). While the purpose of testing in this latter category of trials was observed to be closer to 'screening' proper, upon closer scrutiny the participants were already patients within the healthcare system, could be described as symptomatic and were therefore not 'preclinical'. Details of these inclusion decisions were described in chapter 2 (section 2.3.2, p.58–62).

Once test results were produced, ensuing diagnoses were used to determine the nature of subsequent treatment (55, 51%), the need for further investigation (23, 21%), to aid in the implementation of a predetermined treatment (19, 18%), or to identify a subgroup of patients in whom a particular treatment may be most suitable (10, 9%). While each of these types of management decision was used alongside most diagnostic purposes, there was a clear tendency for specific management decisions to follow each diagnostic decision. Interventions designed to confirm, rule out or screen for possible diagnoses were most often used to select the most appropriate treatment, those allowing a differential diagnosis to be reduced were most often used to direct towards appropriate further investigation, while tests used to establish the extent of disease were most often used to aid in treatment planning (Table 4.10).

Diagnostic Purpose	Treatment Purpose									
	Rx planning	(%)	Need for Rx	(%)	Further Dx	(%)	Suitability for Rx	(%)	Total	(%)
Risk of disease	0	(0)	1	(2)	2	(9)	1	(1)	**4**	**(4)**
Rule out disease	0	(0)	8	(15)	3	(13)	2	(20)	**13**	**(12)**
Narrow DDx	0	(0)	6	(11)	10	(43)	0	(0)	**16**	**(15)**
Opportunistic screening	3	(16)	14	(25)	1	(4)	3	(30)	**21**	**(19)**
Extent disease	13	(68)	4	(7)	3	(13)	3	(30)	**23**	**(21)**
Confirmatory	3	(16)	22	(40)	4	(17)	1	(10)	**30**	**(28)**
Total	**19**	(100)	**55**	(100)	**23**	(100)	**10**	(100)	***108**	**(100)**

Table 4.10: Cross-tabulation of diagnostic and management decisions.
* One study(T57) used a variety of tests covering all test and treatment purposes
Dx – Diagnosis; DDx – Differential diagnosis; Rx - Treatment

4.2.6 General characteristics of trial design

In virtually all trials (101, 94%) the unit of allocation was individuals, while the remainder randomised healthcare centres(T64,T66,T71,T97), days on which patients presented(T43,48) or families(T51). The latter evaluated the use of a genetic test for familial hypercholesterolaemia in both patients with suspected disease, and their relatives.

In almost all studies groups were evaluated in a simple parallel fashion (105, 97%). Three trials employed a 2x2 factorial design to evaluate combinations of test strategies and treatment strategies(T18,T62), or of test-treat strategies and an educational intervention(T97).

In 99 trials (92%) eligible patients were randomised at the time when a clinical decision to perform a test would be made in practice. Eleven studies correspond to Lijmer and Bossuyt's[22] definition of 'randomised disclosure' designs, whereby randomisation is essentially delayed to coincide with the point of releasing test results. In the trial evaluating the addition of TV-US and fFN mentioned above for example, patients received all tests and were randomised to treatment proceeding on the basis of all three tests (the experimental intervention) or the control test only(T94) (Figure 4.3a). Figure 4.3b illustrates a

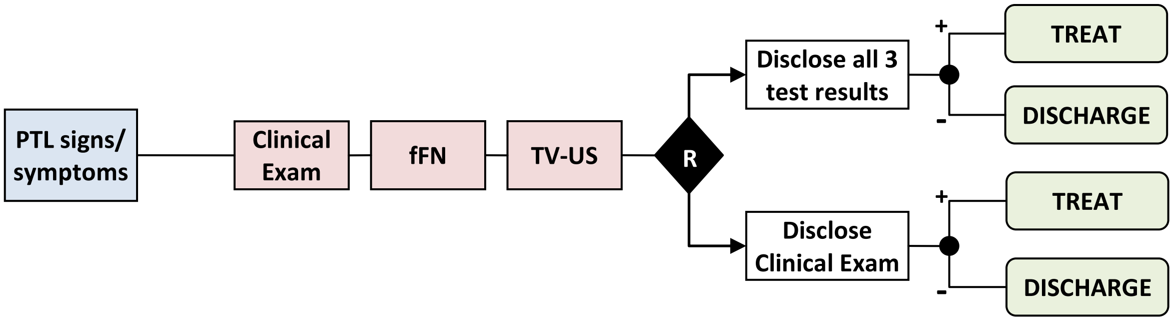


Figure 4.3a: **Example of a randomised disclosure design.** Women with suspected signs of preterm labour (PTL) receive the standard speculum examination, and two experimental tests: fetal fibronectin measurement (fFN) and transvaginal ultrasound (TV-US). Randomisation of test results follows, with patient management proceeding on the basis of all three tests (the experimental intervention) or the control test only[(T94)]

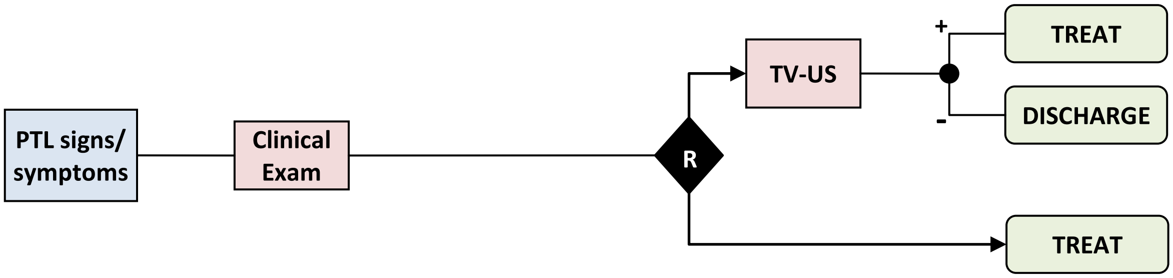


Figure 4.3b: **Example of a standard parallel design.** Women with suspected signs of preterm labour (PTL) receive the standard speculum examination, and are randomised to receive a transvaginal ultrasound (TV-US) to guide treatment decisions, or to all receive treatment for PTL[(T102)].

trial evaluating a similar topic using the standard point of randomisation. Though in both trials patients are treated according to tests given in one arm only, the randomised disclosure comparison allows investigators to establish the prevalence of target disease in both randomised groups, and to identify and follow patients with discrepant test results. Of course not all test-treat strategies are amenable to this design. The eleven examples evaluated biochemical assays[(T21,T26,T43,T49,T65,T94,T100)] and/or radiological imaging modalities(T4,T8,T69,T94,T103), none of which are invasive or carried out by the treating physician. No other variations in the style of randomisation were found.

4.3 Discussion

RCTs that evaluate the patient health impacts of test-treatment interventions are carried out in a wide range of diagnostic settings. Numerous types of tests were conducted to inform an extensive spectrum of questions, in diverse patient groups, and across almost all medical departments. Though small in number, the participation of 30 different countries attests to a widespread recognition of the need to evaluate the clinical effectiveness of diagnostic test practice. Although a key objective of the search was to locate test-treatment RCTs regardless of study topic, the ensuing project cohort contains a very heterogeneous collection. No two trials evaluated the same test-treat interventions in comparable populations.

4.3.1 Imaging tests and cardiovascular medicine are the most frequent diagnostic settings

There was a clear predominance for certain study topics, in particular evaluations of imaging modalities and evaluations of tests used in cardiovascular medicine, however the extent to which this reflects the ‘feasibility’ of performing test-treatment RCTs is unclear. This could in part be a manifestation of the well–established research careers that characterise these two disciplines, indeed much of the methodological development of diagnostic test research originates from the field of radiology[62,187,231]. This finding may also reflect the types of tests being developed and trends in disease; imaging technologies are known to be amongst the most proliferous types of test[232], whilst the need to diagnose and treat cardiovascular disease has been an increasing concern for several decades[233]. Nonetheless, biochemical assays have also proliferated[234], though this review suggests they seem not to have reached evaluations of clinical effectiveness as frequently as imaging tests, and indeed this supports concerns raised within the clinical chemistry community[13,235].

4.3.2 Test-treatment RCTs are highly diverse

These trials emphasise the complexity of factors taken into consideration when deciding on diagnostic practice. Other aspects of test-treat management were evaluated alongside test performance; comparing the organisation of test-treatment delivery was a recurring theme, seen in trials assessing 'units' of care that sought to standardise complex diagnostic protocols, or providing specialised diagnostic services earlier in patient care pathways. These interventions are likely to pose particular methodological problems due to the complexity of test strategies being evaluated, and the opacity of less standardised control strategies that have highly variable protocols.

Moreover, the situation is often more complicated than evaluating the introduction of a test to detect a single target condition. Though concerned only with 'diagnosis', tests can be ordered to address six broad types of diagnostic dilemma, and to aid various management decisions. Extracting a 'target condition' from each study was not always possible. It applied most clearly to confirmatory and opportunistic screening trials, which are typified by the targeted search for a specific condition in a more homogenous patient group. Conversely, tests were often used to sort patients with limited prior test results into multiple management pathways or to guide treatment in patients with known disease. Derived from test accuracy research, this terminology simply does not apply to all test-treatment situations, exemplifying the dynamic context of diagnostic decision-making that is being evaluated in clinical trials compared to the more artificial settings in which diagnostic accuracy must be measured.

4.3.3 Test-treatment RCTs take place in specialised settings

RCTs are expensive and resource-intense enterprises, and it is therefore not surprising that their application to diagnosis appears focussed on evaluating more technologically complicated tests. Since primary care services are usually limited in the availability of these tests, the predominance of secondary and tertiary care settings is likely to reflect this tendency – and indeed, most primary care test-treatment comparisons (14/18, 78%) evaluated clinical assessment and/or biochemical assays, both of which are easily accessible technologies.

This carries implications for the composition of patient groups that tend to be evaluated in these RCTs. The further along the referral process a test takes place, the more highly selected we can expect patient groups to be. This means that test-treatment RCTs are more likely to address highly selected patient groups, and indeed this is born out by the finding that almost half the included trials sought to further categorise patients with existing disease. A potential consequence could be that there will be very little difference in the diagnostic performance of tests in these groups, suggesting sample sizes would need to be very large in order to capture true differences in treatment effects.

4.3.4 RCT study design

The preponderance of replacement comparisons could be consistent with the principle that RCTs should be conducted toward the end of the evaluative trail, following prior determination of test accuracy, treatment efficacy or other attributes of the test-treat strategy. When tests are first introduced, they may be assessed as add-ons to existing tests in order to limit adverse events, whilst as experience of a test’s utility grows physicians may start using it on its own. If this is the case, the presence of so many replacement comparisons could signify that RCTs are being used appropriately as late study designs. Alternatively, it could be that total replacement comparisons are more strongly related to the perceived need to conduct health outcome RCTs, while add-on or triage comparisons are considered to be safer and are consequently not evaluated beyond test accuracy.

The diversity found in the composition of test and treatment components draws attention to particular combinations that might be best suited to evaluation through the RCT design. To this regard, endoscopy interventions are of particular interest. Since testing and treatment are often conducted as part of the same procedure, the impact of each component is inseparable from the other, and consequently, the only way to evaluate the performance of diagnostic endoscopy empirically will be using an RCT study design. The types of trial design available will also be limited when evaluating settings in which test and treatment components are so closely situated, since it is not feasible for all subjects to receive both tests and be randomised to treatment on the basis of test results. This is also true when evaluating tests that are highly invasive, such as large-sample biopsy techniques.

The use of 'randomised disclosure' RCT designs, in an albeit limited number of trials, demonstrates that trialists do already experiment with the structure of test-treatment trial designs. Since none discussed methodological issues or described their RCTs as 'randomised disclosure' designs, their use may represent an intuitive step based on the practicalities of organising the trial, rather than being informed by methodological theory.

Lijmer and Bossuyt[22] discuss a second, potentially very powerful trial design that could apply to test-treatment comparisons, 'discordant test result' RCTs. In these designs, eligible patients are given all comparative tests though only patients in whom test results do not agree (e.g. test-positive according to the existing test and test-negative according to the new test, or vice-versa) are randomised to receive management according to the results of each diagnostic strategy being compared. Discordant test result RCTs are not included in the project cohort, since they were excluded by the thesis selection criteria (outlined in chapter 2, p. 55), and so a separate review will be needed to ascertain and examine them. Several examples have been identified retrospectively amongst excluded studies, confirming that these are indeed carried out. In one such study women with symptoms of a lower urinary tract infection (dysuria), but with negative urine dipstick test results, were randomised to receive three days of antibiotics (i.e. treatment according to symptoms) or to placebo (treatment according to the dipstick test result). Women responded very favourably to antibiotic treatment in the absence of a positive biomarker, experiencing significantly reduced rates of dysuria and increased speed in the resolution of symptoms[236], and hence the trial demonstrated that although the urine dipstick accurately predicts the absence of a detectable infecting organism, it does not predict patients' response to treatment. In eliminating patients from randomisation who would experience no change in management as a result of receiving the experimental test, these designs enable follow-up to focus solely on the patient group who could show a difference in treatment effect, thus increasing the power of the study in comparison to standard pre-test randomised designs. As patients are randomised at the point of *treatment*, discrepant design RCTs are even more elusive to the reviewer since they ostensibly resemble treatment RCTs, and thus risk being more difficult to ascertain when laying in large bibliographic databases that contain many thousands of treatment trials. For future methodological studies to investigate them systematically, it will be essential to introduce diagnostic indexing terms.

4.3.5 Study limitations

During data extraction, difficulties were encountered in trying to discern the composition of test-treatment interventions, as well as the aim of diagnosis, in a large proportion of studies. Often, these two key pieces of information were not explicitly reported but instead concealed within disparate sections of publications, or not reported at all. Consequently, summarising the structure of these test-treatment trials has required careful threading together of often implicit information within the trial reports, supplemented by the examination of external sources detailing current clinical practice.

Since the classifications generated to characterise this group of trials is inherently dependent on published descriptions, these details may not reflect the breadth of questions addressed in all test-treatment RCTs. Chapter 3 concluded that between three and 28 trials may have been overlooked by the author during the study selection process, thus there may well be other aspects of important variation in study settings that are not represented in the project cohort. It is possible, therefore, that the considerable variation observed by this analysis has underestimated the true state of affairs.

The categorisations imposed to enable the synthesis of these studies required the simplification of complex clinical settings. Although the author has consulted clinical diagnosticians during this process, it is possible that some more complex medical comparisons have been incorrectly classified. In addition, due to the time required to analyse such studies in depth neither the extractions nor classifications were double–checked, leading to the possibility that the author's interpretation may differ from that of other non–clinical researchers. In the absence of existing tools to describe test-treatment RCTs, efforts to ensure the relevance and consistency of data extraction entailed using pre–defined items from existing, validated checklists that are considered important to the description of test accuracy studies and general RCTs, as well as ensuring that the retrieval of this information was systematic. Nonetheless, analyses using the inductive approach will always in part reflect the analyst's experience and perspective, since descriptive categories are generated by the analyst's interpretation of the literature. As a result the classifications illustrated in this chapter present a fusion of the information presented in trial reports and the author's perspective, which has been informed by considerable discussion of many included trials with clinicians and methodological experts.

Since the process has been highly subjective, it would be interesting in future to determine whether the resulting classifications are found to be useful to the description of test-treatment trials.

4.4 Conclusion

Examination of this systematically–derived cohort of published trials demonstrates that test-treatment RCTs have answered a wide variety of diagnostic questions. While there was a clear predominance for certain study topics, including cardiovascular department, imaging tests, replacement comparisons and confirmatory purpose of testing, overall the cohort is very heterogeneous. This is due to the complexity of any given test-treatment strategy, which must take into account factors that influence the individual elements of these interventions, diagnosis, decision-making and treatment provision, but also the delivery of such care. These were all identified as important sources of variation in the clinical context. Since variability can limit the applicability of study results, and the extent to which different studies may be synthesised for meta-analysis, an important task will be to determine what effects these sources of variation can have on the results of test-treatment RCTs.

Several study features, key to understanding how and why interventions are being evaluated, appear to be unique to test-treatment RCTs. Examples include the diagnostic purpose and therapeutic aims of tests, important additions to the 'target condition' which is recommended to be reported in diagnostic accuracy studies[72]. Explicit descriptions of which tests and treatments make up the interventions being compared is also essential, as poor reporting impedes users of evidence from understanding these trials and from making reliable and informative syntheses in future. These traits would indicate that further guidance is needed to ensure that these unique studies are reported more comprehensively, and perhaps would benefit from an extension of CONSORT guidelines to test-treatment interventions.

In conclusion, the range of diagnostic questions observed in this cohort of test-treatment RCTs suggests that, despite their current rarity, it is feasible to conduct these trials across many clinical disciplines. But can test-treatment RCTs be conducted well? This fundamental question is addressed in the following two chapters, which appraise the reporting quality (chapter 5) and methodological quality (chapter 6) of test-treatment RCTs. These reviews also explore whether particular attributes of the study settings highlighted in this chapter create obstacles to the production of informative and valid test-treatment RCTs.

5

Trials of Test-Treat Strategies:

The reporting quality of published trials

This chapter builds on the findings of chapter 4 to evaluate the second challenge levelled at test-treatment RCTs: that these trials risk producing evidence that is difficult to interpret and to use. This is achieved by evaluating the reporting quality of trials identified in the project cohort, in chapter 2.

Inadequate documentation of RCTs inhibits the interpretation of results, as well as the ability to translate interventions into practice[121,136]. The complexity of test-treatment interventions is hypothesised to create particular difficulties in the ability to produce full and informative accounts of their conduct. Since diagnostic decision–making is highly variable, trials must pre–specify how test results lead to diagnoses and treatment plans in order to be sure of how interventions are creating the observed effects[22,76,79]. However, researchers claim it could be challenging to document multiple interventions entailing decision–making to the extent necessary to allow findings to be interpreted, to be compared across studies, and to enable beneficial interventions to be translated into practice[176].

The following study has been designed to evaluate the extent to which test-treatment RCTs produce informative reports. It aims to achieve this by systematically appraising the reporting quality of the published trials ascertained in chapter 2 and characterised in chapter 4.

5.1 Methods

Assessments of reporting quality focussed on three aspects that are fundamental to interpreting and using trial findings: the need to understand what happened, to whom, and how it was measured. Accordingly, trials were appraised regarding the reader’s ability to discern the selection and flow of participants through the trial, how participants should be managed according to the allocated test, and how their response to these interventions was measured; specifically:

1.1 Were test-treatment interventions completely identified and described?

1.2 Did trials make clear the proportion of eligible patients recruited into the study, and document whether some participants did not receive the allocated intervention, were lost to follow-up, or were not analysed?

1.3 Were primary outcomes completely defined, clearly conveying what was being measured, how and by whom?

Reporting of methodological safeguards employed by trialists to maintain a study's internal validity were also appraised, however this was conducted as part of the review of methodological quality presented in chapter 6.

5.1.1 Design of a quality assessment tool

In the absence of an existing quality assessment tool specific to test-treatment RCTs, standardised data collection and appraisal forms were designed. Items were identified from two validated, internationally accepted standards for the reporting of RCTs: the CONSORT checklist[121] and the extension of the CONSORT statement for non-pharmacologic therapy interventions[147]. Table 5.1 lists the extracted items.

This new quality appraisal tool was tested by extraction of five test-treatment trials, randomly selected from the project cohort. Minor changes were made to improve the standardisation of data collection and quality assessment.

5.1.2 Data extraction

Trials identified by the project search with at least one publication of study findings were included. All articles reporting on the same trial were examined. Important related publications not identified by the restricted timeframe of the project search, such as

Trial Documentation Objectives	Item
1.1 Does the report give a full description of all competing test-treatment interventions?	*Was the test method reported?* *Were treatments reported?* *Were diagnostic decisions reported?* *Were treatment decisions reported?* *Was an algorithm diagram provided for each intervention?*
1.2 For each group, is it clear whether some participants did not receive the allocated intervention, were lost to follow-up, or were not analysed?	*The number of eligible participants* *The number of participants randomised to each arm* *The number receiving the allocated intervention* *The number who completed management as allocated* *The number of participants included in the main analysis* *Use of a CONSORT diagram to record participant flow*
1.3 Were primary outcomes completely defined, clearly conveying what is being measured, how and by whom?	*Is a primary outcome clearly defined?* *What was the primary outcome?* *How was it measured?* *Who measured it?*

Table 5.1: Items extracted to evaluate the reporting quality of test-treatment RCTs.

original trial reports, preliminary design protocols or long-term follow-up papers, were also traced through citations and author-title searches of Medline. Data were extracted to a purpose built relational database (Microsoft Office Access 2007) and appraised for reporting quality using the methods reported below. Extraction and quality assessment were performed by the author.

5.1.3 Appraisal of reporting quality

Description of interventions

Trials were assessed for description of the interventions under evaluation. Test-treatment interventions were considered in four components: the test method, criteria used to form the diagnostic decision, criteria used to select treatments, and the treatment method. Each component was judged to have been reported if any relevant information was described or referred to by citation to another study (Box 5.1). Specifically, quality judgements were not predicated on whether adequate clinical detail was reported for a given component to be replicated in clinical practice. Although this would have been desirable, it would have

Box 5.1: Definition of the four components used to assess the description of test-treatment interventions, with examples.

Test Method: **Technique used to perform the test.** Reporting the name of the test only was considered insufficient.
e.g. *“Radiographs of the knee were obtained in the lateral and anteroposterior projection and were supplemented with patellar or tunnel views if pathologic abnormalities of the patellofemoral joint or intercondylar notch were suspected"* (T5)

Diagnostic Decision: **Description of the operational criteria used for arriving at a particular diagnosis using the test results.**
e.g. *"If the lung scan showed no abnormalities, pulmonary embolism was excluded; if there were 1 or more segmental perfusion defects that were normally ventilated, the scan was considered diagnostic for pulmonary embolism (“high-probability scan”); and if there were perfusion defects that did not meet criteria for a “high-probability scan,” the scan was considered nondiagnostic."* (T63)

Treatment Decision: **Description of how treatments were selected as a result of the diagnosis.**
e.g. *“Stones detected on EUS* [endoscopic ultrasound] *were removed endoscopically during a separate session; stones detected on ERC* [endoscopic retrograde cholangiography] *were removed immediately, during the same session. When the initial ERC or EUS failed, a second procedure was carried out."* (T95)

Treatment Method: **Description of how selected treatments were administered.** Reporting of the treatment name only was considered insufficient.
e.g. *“After ultrasound diagnosis of an anal sphincter tear… women were brought immediately to the operating room to provide appropriate lighting, instruments, and assistants and underwent a surgical exploration of the perineum by the obstetrician-in-charge under senior supervision. The anal sphincter was exposed and its integrity assessed by inspection and palpation. The ends of the sphincter were approximated end-to-end with 2–0 monofilament polyglyconate sutures (Maxon, Sherwood Davis & Geck, St. Louis, MO). Postoperatively, women received dietary advice to avoid constipation, with occasional use of stool softeners. For women allocated to the control group, the obstetrician sutured the perineum after clinical examination.”* (T11)

required considerable consultation with a wide range of clinical experts, which unfortunately was not feasible for the current project.

Following good practice recommendations by the MRC for the conduct of complex interventions[83], studies were also appraised for their use of diagrams depicting the competing care pathway algorithms. Where present, diagrams were considered complete if they reflected all four test-treat components for each trial arm, and partially reported if at least one component was not represented.

Clear accounting of participant flow

Reports were reviewed to assess whether all participants could be clearly accounted for throughout the study, with or without the aid of the flow chart recommended by CONSORT. For each study group, the number of participants evaluated for eligibility, randomised, receiving the allocated intervention, completing follow–up and included in the main analysis were extracted. For cluster-randomised trials the number of clusters randomised and analysed were also extracted.

If the figures reported for one of the five participant flow elements did not agree, for example if the number analysed did not tally with the numbers randomised and lost to follow up, then that element was considered as inadequately reported. Trials reporting all five elements were deemed fully reported.

Complete description of primary outcomes

Trials were appraised for clear reporting of a primary endpoint. Following the approach of Chan and Altman[152], when studies failed to define their main measure of effect clearly outcomes were preferentially extracted according to the variable used in a power calculation, followed by a main outcome described explicitly in primary study objectives. If none of these was provided, the primary outcome was categorised at ‘not defined’.

Endpoints were classified as patient or process outcomes. For the purposes of description, outcomes were further categorised according to the response they were designed to measure.

The quality of reporting how outcomes were measured was determined using two criteria: documentation of the method of measurement and the timing of measurement. Methods were considered as reported if a validated tool was used (for example the Short-Form 36 to assess general health), if non-validated but fully described tools were used, or if criteria to direct a rigorous assessment of outcome were provided (for example the operational definition of a target condition and test methods used to arrive at a diagnosis). Documentation was considered complete when the time at which the primary assessment should be conducted was also made explicit.

5.1.4 Analysis

Data were exported from the extraction database to Microsoft Excel 2007 for sorting and analysis. The objective of this review was to describe the frequency with which test-treatment trials were found to have reported their conduct appropriately. Consequently, this chapter presents a descriptive summary of these findings using percentages that reflect the categorical nature of the data. Comparisons between frequencies were used to enhance the description of findings, and aimed to highlight potential associations between the variations in reporting quality and aspects of the study settings, as characterised in chapter 4. The author did not intend to evaluate specific hypotheses regarding these associations, but rather to generate hypotheses for how easily the methods of trial design could be employed to conduct informative test-treatment RCTs. As a consequence testing for the statistical significance of these comparisons was not appropriate.

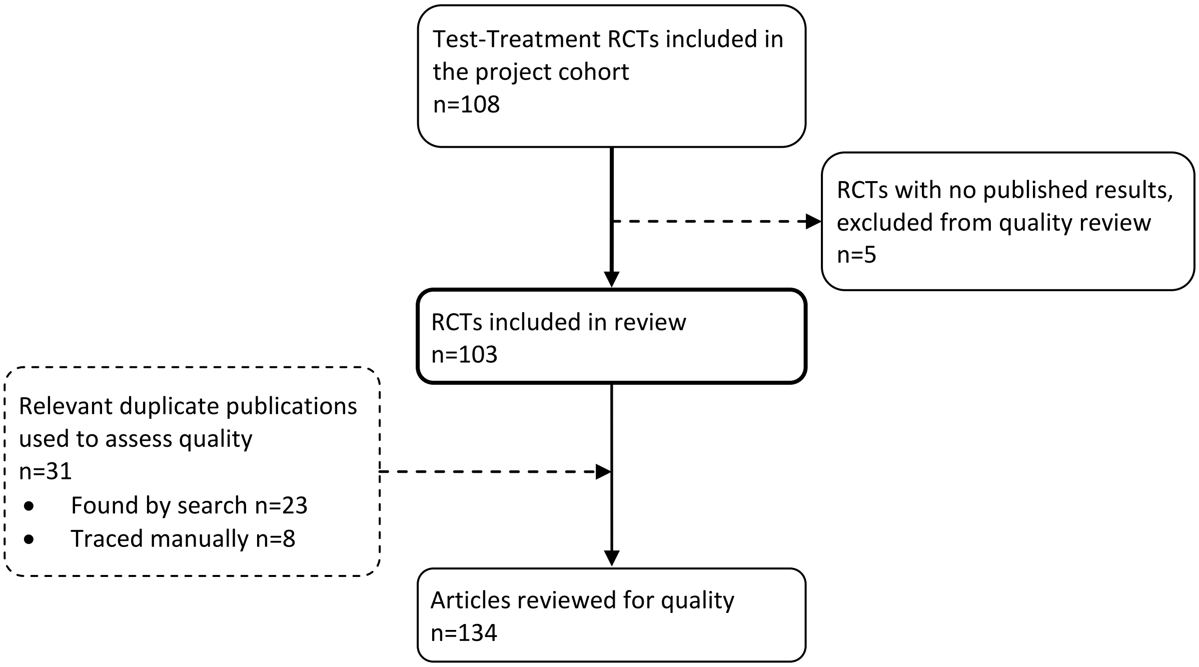

Test-Treatment RCTs included in the project cohort
n=108
RCTs with no published results, excluded from quality review
n=5
RCTs included in review
n=103
Relevant duplicate publications used to assess quality
n=31
• Found by search n=23
• Traced manually n=8
Articles reviewed for quality
n=134

Figure 5.1: Test-treatment RCTs appraised for quality of reporting and methods.

5.2 Included studies

Of the original cohort, five trials(T22,T25,T43,T80,T84) had no published results and so were not appraised for reporting or methodological quality (Figure 5.1)*. This revised cohort (N=103) included 103 RCTs that compared 105 control arms with 119 experimental arms.

Thirty–two trials had multiple publications (range 2-5), including those found by the project search and others published outside the project timeframe retrieved by the author through targeted searches of Medline. These most commonly reported additional economic analyses(T14-15,T38,T48,T49,T55,T57,T61,T68,T71,T90,T107), trial design(T14,T38,T57,T66,T96-97,T101), long-term follow-up results(T38,T44,T46,T55-T56,T90,T99), sub-group analyses (T14,T16,T18,T32,T41,T55,T73,T89), reproduction of trial results in a full health technology appraisal report(T34,T36,T47) or

[*] Citations for included test-treatment RCTs are prefixed with a 'T'

reproduction of an HTA assessment as a short journal article(T57). These additional publications are listed as references in Appendix C (p.380).

5.3 Results

Summaries of reporting quality appraisals are provided in Appendix D (p.386).

5.3.1 Documentation of test-treatment protocols

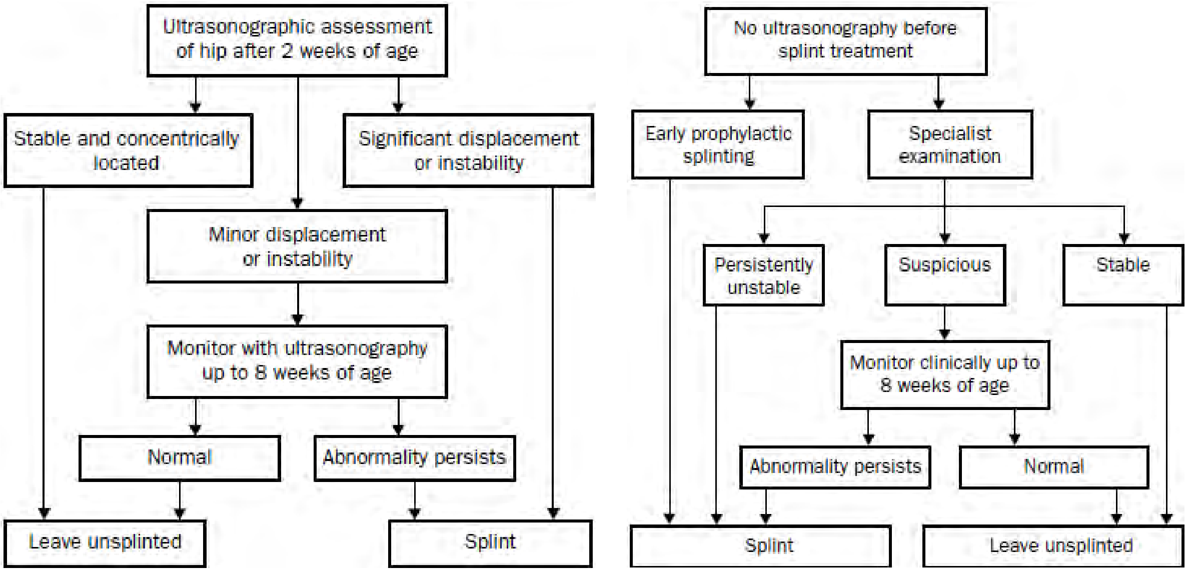
Use of a protocol diagram

Diagrams illustrating the competing test-treatment strategies were included by approximately one-fifth of trials, depicting 24/119 experimental interventions and 22/105 control interventions. These ‘care pathway’ diagrams were found to be most informative if they illustrated how patients might travel through the process according to diagnostic findings and their treatment decisions. Fewer than 10% (8 experimental, 7 control) provided these full accounts, while the majority of diagrams were considered incomplete due to the absence of decision-making protocol elements.

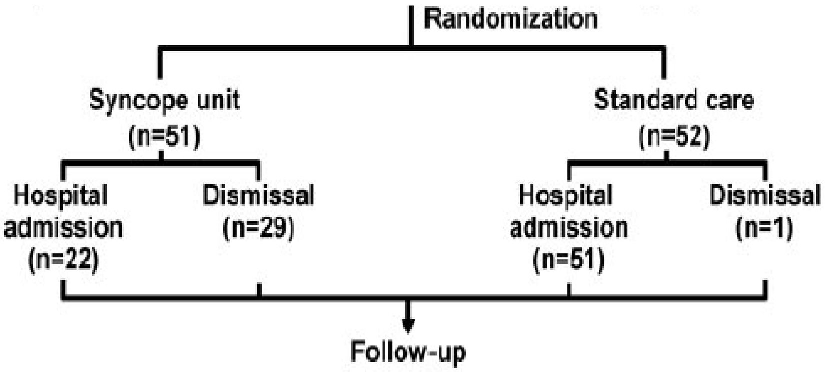
Figure 5.2a shows a direct copy of a published care pathway for a trial evaluating the benefit of adding an ultrasound assessment of the hip to the existing clinical assessment in order to confirm the presence of mechanical hip instability in neonates(T15). It was considered complete due to its clear delineation of the tests given, the main diagnostic decisions taken on the basis of test results, and which treatments these categories should lead to. A representative example of a partial diagram is provided in Figure 5.2b, again a published care pathway, but this time for a trial assessing the benefits of investigating patients in a specialist unit, rather than in the emergency department (ED) as standard, in order to establish the cause of syncope for the purpose of directing further investigation and treatment(T53). This trial was arguably making a much more complex comparison than the previous ultrasound for hip example for several reasons, the most important being that

Figure 5.2: Graphic representation of two test-treat algorithms.

A: Complete algorithm showing all test-treat decisions. In this example neonates with suspected hip instability were randomised to receive standard specialist examination (right) versus an additional ultrasound scan (left) to inform whether child should be splinted(T15). *Reproduced with permission.*


Ultrasonographic assessment of hip after 2 weeks of age
Stable and concentrically located
Significant displacement or instability
Minor displacement or instability
Monitor with ultrasonography up to 8 weeks of age
Normal
Abnormality persists
Leave unsplinted
Splint
No ultrasonography before splint treatment
Early prophylactic splinting
Specialist examination
Persistently unstable
Suspicious
Stable
Monitor clinically up to 8 weeks of age
Abnormality persists
Normal
Splint
Leave unsplinted

B: Partial algorithm showing treatment decisions but omitting diagnostic decisions. Patients with syncope of undetermined cause were randomised to a routine ED investigation versus more formalised evaluation in a syncope unit(T53). *Reproduced with permission.*



Number of missing elements (in at least one arm*)	Comparator (N=103)	(%)	Experimental (N=103)	(%)
None	6	(6)	10	(10)
1	16	(16)	20	(19)
2	23	(22)	35	(34)
3	19	(18)	26	(25)
4	40	(39)	19	(18)

Table 5.2: Contrast in the fullness of reporting control vs. experimental interventions. Table presents counts of the total number of included trials (N=103) that omitted reporting between zero and all four elements of test-treat protocols.
* Some trials evaluated multiple comparator arms and/or multiple experimental arms.

the diagnostic strategies involved multiple tests conducted in two different care settings. However, as is evident, neither the tests, diagnostic decisions or treatment decisions were illustrated in the care pathway. Understanding how patients with unexplained syncope were investigated, and the basis upon which they were referred for further management, would have to be sought from the written description.

Reporting of methods and decisions

Written descriptions of the clinical processes involved in test-treatment interventions were very poor, characterised by frequent omission of multiple intervention elements and a very low level of detail.

For experimental interventions, all four elements of methods and decision-making were at least partially described (or appropriately cited) by 10% (10/103) of trials[†], while only 6% (6/103) achieved this for the comparator intervention (Table 5.2). Trials were twice as likely to omit all description for control interventions, with experimental protocols more likely to be at least partially reported than their comparators. Only three test-treatment RCTs(T76, T98, T105) outlined four protocol elements for all study arms.

[†] *For trials with >2 study groups, the best reported arm was used.*

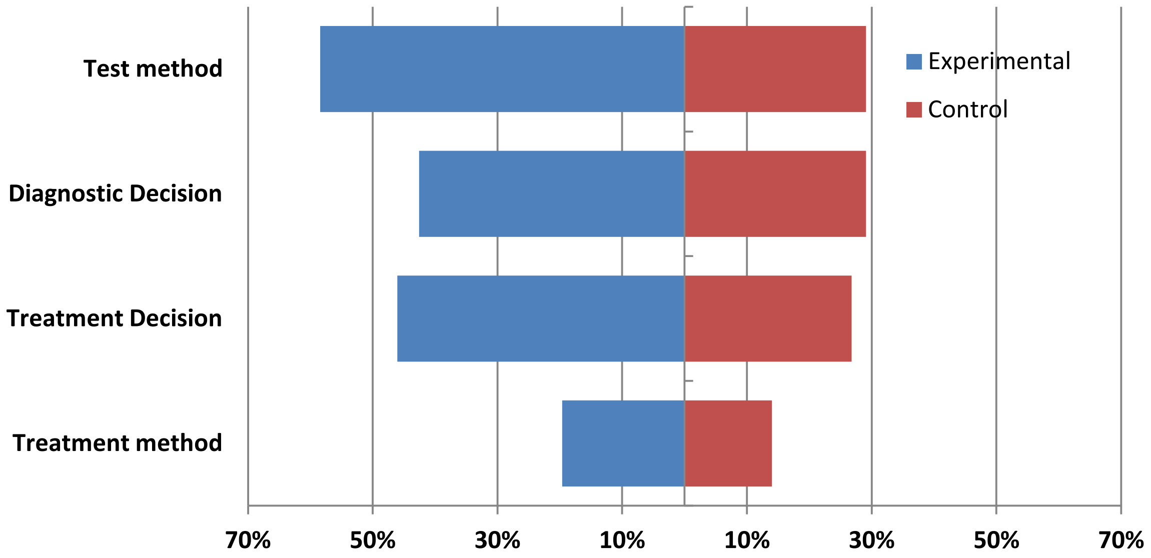
As described in chapter 4, not all test-treatment strategies being evaluated by these trials included tests and treatments; some evaluated the benefits of not testing and giving treatment to all for example. Since description of certain elements was therefore not required in these trials, the denominators for frequency calculations were reduced to represent only those trials in which description of each given element was necessary. Two experimental and 17 control strategies did not involve a diagnostic test (i.e. the comparison was to ‘no test’)(T13,T18,T28,T37–39,T42,T45,T49,T52,T58,T63–4,T76,T96,T102–5) and so the denominators for reporting of test method and diagnostic decision–making were reduced accordingly (Experimental n=101, Comparator n=86). Similarly, one experimental and 8 control strategies consisted of treating all patients(T13,T38–39,T52,T63,T76,T102,T104–5) and so did not involve any treatment decision (Experimental n=102, Comparator n=95), while one experimental and three control strategies did not give any treatment to any patients(T39,T63,T76,T104) and so did not involve any treatment methods (Experimental n=102, Comparator n=100).

Test methods were the most commonly described element, reported in 58% (59/101) of experimental and 29% (25/86) control intervention protocols (Figure 5.3).

For experimental interventions the criteria by which management decisions were made were reported by fewer than half the studies: diagnostic decisions in 43% (43/101) and treatment decisions in 46% (47/102). By comparison, only a third of trials reported these essential elements for control protocols (diagnostic decisions in 29% [25/86]; treatment decisions in 27%, [26/95]). Treatment methods were most poorly reported element, outlined for experimental strategies in only 20% (20/102) of trials, and for control strategies in only 14% (14/100) of trials.

Using the number of missing protocol elements as a proxy for quality of reporting, certain types of diagnostic tests appeared to be better described. Whether evaluated as the experimental or comparator test, biochemical, electrophysiological and clinical examination

Figure 5.3: Proportion of RCTs describing each element of the test-treat protocol according to study group. For trials with >2 study groups, elements were considered reported if described for at least one experimental intervention.



techniques all tended to have 3 – 4 elements missing, while imaging and endoscopies were more likely to be more fully reported (Tables 5.3 and 5.4). Descriptions of new radiological imaging modalities were of better quality than existing ones, as were interventions introducing multiple new tests. Studies that failed to identify comparator tests beyond being 'standard care' performed poorly in all elements.

5.3.2 Participant Flow

Full accounts of participant flow, encompassing all 5 items recommended by CONSORT, were provided by 44 (43%) trials, including 20 (19%) that provided a full flow diagram. Other than one study (cluster-randomised) that reported none of these details(T64), the

Experimental test	Number of missing elements						% missing	
	0	**1**	**2**	**3**	**4**	**Total**	**≤2 elements**	**>3 elements**
Biochemical	1	2	3	12	7	25	24%	76%
Biopsy	0	1	1	0	0	2	100%	0%
Clinical	2	1	1	4	5	13	31%	69%
Electrophysiology	0	1	2	4	3	10	30%	70%
Endoscopy	3	2	3	1	2	11	73%	27%
Imaging	2	7	14	10	6	39	59%	41%
No test	3	0	2	0	0	5	100%	0%
Telemedicine	0	0	0	0	2	2	0%	100%
Various	1	3	3	3	2	12	58%	42%
Total	**12**	**18**	**31**	**37**	**31**	**119**		

Table 5.3: Distribution of protocol reporting quality by experimental test type.
Denominator is the number of experimental interventions (N=119)

Comparator test	Number of missing elements						% missing	
	0	**1**	**2**	**3**	**4**	**Total**	**≤2 elements**	**>3 elements**
Biochemical	0	1	3	5	1	10	40%	60%
Biopsy	0	0	0	0	1	1	0%	100%
Clinical	0	1	1	4	9	15	13%	87%
Electrophysiology	0	2	0	2	3	7	29%	71%
Endoscopy	0	3	5	1	5	14	57%	43%
Imaging	1	2	6	5	7	21	43%	57%
No test	4	6	7	0	0	17	100%	0%
Standard Care	0	0	0	0	10	10	0%	100%
Telemedicine	0	0	0	0	2	2	0%	100%
Various	0	2	0	2	4	8	25%	75%
Total	**5**	**18**	**24**	**22**	**46**	**105**		

Table 5.4: Distribution of protocol reporting quality by control test type.
Denominator is the number of comparator interventions (N=105)

		Cluster–randomised trials	**Individually-randomised trials**		**Total**	**(%)**
		(n=6)	**2 arms** (n=87)	**>2 arms** (n=10)	(N=103)	
Complete	Full diagram	3	16	1	20	(19)
	Full text, no diagram	0	6	1	7	(7)
	Partial diagram & text	0	15	2	17	(17)
	Total	***3***	***37***	***4***	***44***	***(43)***
Incomplete	Partial diagram, no text	0	10	1	11	(11)
	Partial text, no diagram	0	22	3	25	(24)
	Partial diagram & text	2	18	2	22	(21)
	No diagram or text	1	0	0	1	(1)
	Total	***3***	***50***	***6***	***59***	***(57)***

Table 5.5: Use of the CONSORT participant flow diagram by test-treatment RCTs. Denominator is the number of test-treatment RCTs (N=103)

remaining 58 (57%) trials published partial information (Table 5.5). Use of a consort diagram was more frequently associated with more complete accounting of participant flow, with over half such trials considered complete (37/70) compared to just a fifth of those not employing flow diagrams (7/33).

Numbers screened for eligibility and receiving the allocated intervention were the most frequent omissions, while the number of patients included in the primary analysis were provided in all but two studies(T41,T64) (Table 5.6). Allocation was amongst the most well-reported of items; only five individually–randomised trials (5%) failed to indicate the number of patients randomised to each arm(T7,T49-T50,T82,T86). Two cluster-randomised trials

Participant flow	Eligibility	(%)	Allocation	(%)	*Rec'g Itvn	(%)	Follow-up	(%)	Analysis	(%)
Reported	**65**	**(63)**	**96**	**(93)**	**71**	**(68)**	**80**	**(78)**	**101**	**(98)**
CONSORT diagram	*51*	*(50)*	*63*	*(61)*	*47*	*(45)*	*44*	*(43)*	*42*	*(41)*
text only	*14*	*(14)*	*33*	*(32)*	*24*	*(23)*	*36*	*(35)*	*59*	*(57)*
Not reported	**38**	**(37)**	**7**	**(7)**	**34**	**(32)**	**23**	**(22)**	**2**	**(2)**

Table 5.6: Publication of participant flow numbers in test-treatment RCTs.
Denominator is the number of test-treatment RCTs (N=103)
* **Rec'g Itvn** – Receiving Intervention

also provided insufficient information in this respect; one did not report the number of clusters allocated to each intervention, although the number of participants enrolled in each arm was provided(T51), while the second trial provided the number of randomised clusters but not participants(T64).

5.3.3 Definition of primary outcomes

A total of 150 primary outcomes were reported in 97 test-treatment trials. A quarter of trials (27, 26%) failed to clearly identify a principal endpoint, however in 13 trials this was deduced using the variable reported in a power calculation(T5,T9-10,T12,T23,T36,T40,T46-47,T50,T58,T60,T65), while in another seven a variable was deduced from the study aim(T7,T28,T33,T76,T87,T105,T108). Six trials(T2,T29,T31,T59,T79,T82) defined none of these.

The majority of studies focussed on one primary outcome (79/103, 77%), while the remaining 18 trials(T4,T7,T11,T16,T18,T24,T26,T28,T30,T32,T39,T51,T53,T55,T66,T93,T100,T107) used between two(T11,T16,T26,T30,T32,T53,T55,T100,T107) and 15(T24) separate measurements (median 3, IQR: 2–4).

Primary outcomes more often measured patient health (54%) than clinical processes (39%) (Table 5.7).

Outcome Type	**Number of trials (N=97)**	**(%)**
Patient	***53**	**(54)**
Symptom rate	*13*	*(25)*
Clinical status	*9*	*(17)*
Adverse event rate	*8*	*(15)*
Function	*8*	*(15)*
Residual disease rate	*7*	*(13)*
Recurrent disease rate	*6*	*(11)*
Quality of life	*5*	*(9)*
Mortality	*4*	*(8)*
Health perception	*2*	*(4)*
Psychological morbidity	*2*	*(4)*
Absenteeism	*1*	*(2)*
Satisfaction	*1*	*(2)*
Patient outcome total	***†66***	
Process	***38**	**(39)**
Therapeutic yield	*17*	*(45)*
Timing of care	*8*	*(21)*
Cost	*7*	*(18)*
Appropriateness of treatment decision	*5*	*(13)*
Diagnostic yield	*4*	*(11)*
Process outcome total	***‡41***	
Composite	**7**	**(7)**

Table 5.7: Types of outcomes measured as primary endpoints in test-treatment RCTs

* One trial measured both patient and process endpoints(T93)

† 6 trials measured >1 type of patient outcome: Absenteeism, Function, Health perception, Psychological morbidity, Quality of life, Satisfaction, Symptom rate(T4); Mortality, Quality of life(T18); Psychological morbidity, Adverse event rate, Function(T24); Symptom rate, Function, Quality of life(T28); Disease rate, Symptom rate(T66); Quality of life, Function(T107).

‡ 3 trials each measured two different process outcomes: Diagnostic yield and Therapeutic yield(T53); Timing of care and Cost(T55,T100).

Sixty-eight patient outcomes were measured by 53 trials as primary outcomes. The most common measure was symptom frequency, for example the rate of epigastric pain or regurgitation in patients managed for dyspepsia(T13,T40,T44, T46,T65,T71) (Table 5.7). Adverse events (8, 15%) were experienced as a consequence either of diagnostic procedures, such as the rate of arm dysfunction suffered as a result of performing biopsy of the axillary nodes(T24), or treatment harms and failure, for example fecal incontinence in women treated for severe perineal tears(T11). Evaluation of clinical status (9, 17%) generally constituted a surrogate for downstream health, for instance the clinical pregnancy rate (rather than live, healthy birth rate) in women treated for primary infertility(T104). Assessments of function (8, 15%) could be either surrogate outcomes, such as the evaluation of maximum exercise endurance in patients with coronary artery disease (T20,T41,T92), or true health measures, for example physical mobility experienced by patients being treated for chronic lower back pain(T4,T28) or knee derangements(T107).

Seven trials (13%) were primarily interested in the rate of residual disease, a marker for the degree of success in diagnosing and treating disease, for example measuring the eradication rate in patients with *H.pylori* infection(T105), or the rate of venous thromboembolism in patients managed for a suspected pulmonary embolism(T63). Another six trials measured disease recurrence, defined as new episodes of the presenting condition, for example recurring episodes of varicose veins after the removal of primary obstructed veins(T58), or of bladder cancer after tumour resection(T1).

Mortality was the primary measure in only 4 trials, conducted in geriatrics(T18), oncology (T42), respiratory medicine(T62) and infectious disease(T23). Similarly, perceptual, emotional and behavioural responses to the test-treat process, including quality of life measures, were less often the primary focus of test-treatment RCTs, and tended to be measured alongside other patient outcomes.

Process measures

Thirty–eight trials examined a total of 41 process measures as the primary outcome. The quantification of management decisions was the most frequent measure of test-treat processes. Approximately half evaluated an aspect of therapeutic yield, defined here as the treatment rate, in order to examine the impact of testing on treatment decisions (Table

5.7). For example, antibiotic use was quantified in patients with potential pneumonia to establish whether point-of-care C-reactive protein tests would lower prescription rates(T97), while in two secondary care trials the proportion of women delivering by Caesarean-section was evaluated to assess the impact of tests used to determine dystocia(T54,T81). The *appropriateness* of treatment decisions was measured in another four studies(T10,T17,T32,T60), whereby the working diagnosis was confirmed using additional tests after treatment had been given. For example, patients presenting to the ED with chest pain were either hospitalised or discharged once acute cardiac ischaemia had been diagnosed or ruled out(T32).

Diagnostic yield, or the rate of diagnoses made, was also used as a primary outcome, though by only three (8%) studies(T53,T88,T93). For example, a trial comparing two types of endoscopy to investigate obscure gastrointestinal bleeding measured the number of cases in whom a definite source of bleeding could be identified(T88).

Eight studies concentrated on the timing of care, measuring either the total length of treatment, such as length of hospital stay(T50,T55,T72,T85–86,T94,T100) or the time taken to reach a diagnosis(T56). Cost, either of total management(T5,T55,T70,T100) or diagnostic procedures only(T12,T19,T83), was the primary calculation in seven (19%) studies.

Composite measures

Adverse event and treatment rates were combined into single composite measures in seven trials, most commonly in cardiovascular settings(T6,T14,T89,T96,T99,T101) where the prevalence of myocardial infarction, death and revascularisation procedures provided a summary rate. The combined frequency of procedural morbidity, length of hospitalisation and the consequences of missed diagnoses were also used in one gastrointestinal trial, to evaluate the effectiveness of two endoscopic techniques for the clearance of bile duct stones(T95).

Description of outcome measurement	**No. of primary outcomes**	**(%)**	**No. of trials**	**(%)**
Complete	74	(49)	*53	(51)
Incomplete	57	(38)	‡36	(35)
Not reported	19	(13)	†14	(14)
Total	**150**	**(100)**	**103**	**(100)**

Table 5.8: Completeness of outcome measurement reporting

* Trial considered fully reported if all primary outcomes completely described

† Trial considered not reported if no primary outcomes were described

‡ Trials with mixed reporting for multiple outcomes considered as incompletely reported.

Describing outcome measurement

Description of how outcomes were measured was considered adequate enough to attempt replication in approximately half (53, 51%) of included trials (Table 5.8). Complete reports were presented for 49% (74/150) of primary outcomes, while neither the method nor timing of measurement were described for 13% (19/150). Incomplete reports most commonly omitted the time at which outcomes should be measured (43/57, 75%). Overall, the method of measurement was not provided or poorly described for 28% of outcomes (42/150) measured in 32% (33/103) of included trials. Box 5.2 illustrates two examples of reports that were judged to be completely and incompletely reported.

5.4 Discussion

The review of reporting quality finds that test-treatment RCTs are currently poorly reported, providing often incomplete accounts of precisely what happened, to whom, when, and how this was measured. It is clear that this suboptimal detail will impinge on the user’s ability to interpret the meaning of trial results, and potentially also to use such findings to improve clinical practice. The following discussion considers the main findings with regard to how test-treatment trials compare to similar cohorts of treatment RCTs and complex

Box 5.2: Examples of adequate and inadequate reports of outcome measurement

Example of incomplete reporting of outcome measurement:
A trial sought to determine whether the addition of a SPECT Tc 99m sestamibi scan to the standard clinical evaluation strategy (not reported) would more accurately distinguish between ischaemic and non–cardiac causes of chest pain, in order to more appropriately identify the patients in need of further cardiological investigations and treatment(T32). The study measured two primary outcomes: 1. the rate of inappropriate discharge of patients with acute cardiac ischaemia (ACI), and 2. The rate of inappropriate admission of patients without ACI. The following extract describes how these outcomes were measured, and was judged as incomplete since the criteria for diagnosing ACI using the follow–up tests has not been provided or cited:

"To make the final diagnosis of the presence of ACI, biomarkers and serial ECGs were obtained. Protocol-specified follow-up stress testing with perfusion or echocardiographic imaging was also performed. For patients admitted to the hospital, this was usually accomplished during the period of hospitalization. For patients discharged from the ED, a return visit to the study site 24 to 36 hours later was made for follow-up biomarkers, ECGs, and stress testing. The confirmed final diagnosis was assigned by the principal investigator at each site based on cardiac enzyme levels, ECGs, stress test, and when available, cardiac catheterization data." (T32)

Example of complete reporting of outcome measurement:
The following excerpt is from a trial evaluating whether patients undergoing explorative investigation for clinically suspected bile duct stones might benefit from first receiving a less invasive endoscopic ultrasound, proceeding to the more invasive endoscopic retrograde cholangiography only if stones are still suspected. The primary outcome evaluated whether the new strategy could reduce the rate of negative outcomes in these patients, and the method of arriving at these measurements was judged to be comprehensive, consisting of adequate description as well as citations:

"The primary end point of the study was the proportion of patients with negative outcomes, to related to [sic] *either endoscopic procedures (complications) or false−negative diagnosis of stones...Complications of endoscopic procedures were assessed prospectively by a single investigator who was not blinded to group assignment. Severity was graded according to consensus criteria [17,18] as minimal (no need for hospitalization), mild (2±3 days of hospitalization), moderate (4±10 days of hospitalization), severe (> 10 days of hospitalization, or surgery, or intensive care unit admission), or fatal. Because all patients were hospitalized, no clear distinction between the first two categories was possible, and they were merged as minimal−to−mild. Acute pancreatitis was defined as a new or worsened abdominal pain which lasted for more than 24 hours, and was accompanied by a serum amylase level greater than three times the upper normal limit [17]. Transient abdominal pain that required medical intervention (face−to−face doctor attention and analgesic/antispasmodic drugs), but subsided within 24 hours and did not cause prolongation of hospital stay, was recorded as a separate category and graded as a minimal−to−mild complication. Bleeding was defined as clinical evidence of hemorrhage, such as melena or hematemesis, with an associated decrease of at least 2 g/dL in hemoglobin concentration, or the need for transfusion [19]. A negative outcome related to false−negative diagnosis of stones was defined as an occurrence of either of the following: (i) detection of bile duct stones during follow−up, or (ii) hospitalization possibly related to bile duct stones but without definite stone confirmation (acute pancreatitis, acute cholangitis, obstructive jaundice)."* (T95)

intervention RCTs, with particular thought given to barriers which may prevent full and clear reporting in these trials.

5.4.1 Trials often fail to provide complete accounts of participant flow

Clear and complete accounting of participant flow as recommended by CONSORT was performed in around half the examined trials, with one–fifth of all trials providing a complete flow diagram. Of the trials with incomplete reporting, the most worrying insufficiencies were the numbers screened for eligibility and the numbers receiving the allocated intervention. Without knowledge of the former trialists will be unable to guarantee that study populations are representative of the general patient population. This is particularly important when evaluating tests, since differences in the case–mix of populations are known to impact on test performance[67]. Importantly, transparency regarding the numbers of patients receiving the allocated intervention are key to discerning the extent to which observed health effects are due to the interventions being compared[121]. Therefore these findings suggest users will have difficulty in fully interpreting the meaning of results in a large proportion of test-treatment RCTs.

Reporting of participant flow varies considerably in the literature. The prevalence of providing flow diagrams ranges from 28%[118] to 86%[153] in cohorts of recent unselected trials, largely comprising single–intervention treatment RCTs, with suggestions that the more superior accounts are encountered in articles published by leading journals[118]. An earlier review, of 270 RCTs (mostly pharmacological trials) published in 5 leading journals during 1998, found 52% provided a flow diagram of varying degrees of completeness[237]. The most poorly reported aspect was the number of patients receiving the allocated intervention (27%, 73), as per the present review though apparently considerably worse than found in test-treatment RCTs. A similar review of 63 complex intervention RCTs, evaluating weight loss interventions, found that 25% had provided flow diagrams, though the authors based this appraisal on 'appropriate labelling' of diagrams and did not state whether those encountered were complete. Though only 20% of test-treatment trials provided full diagrams, in total 68% (70/103) provided them in some form. Thus, though suboptimal, the present cohort would appear to reflect a similar quality of reporting of this item when compared to RCTs in general, and possibly better than found in other complex intervention trials.

5.4.2 Trials often provide clear definition of primary measures

Clear definition of primary endpoints was achieved by three–quarters of included trials, a considerably better performance than found by reviews of both single–intervention and complex intervention RCTs. For example a review of all trials indexed in PubMed in December 2006 (n=616) found primary measures of effect to be clearly defined by 53%[118], while a previous review by the same researchers of trials indexed in PubMed in December 2000 (n=519) found a slightly lower rate of 45%[152]. Similarly 46% of trials evaluating complex interventions for the treatment of weight loss (n=63) identified a primary endpoint. Moreover, the criteria used to appraise this item follow the approach used by one of these reviews[152], suggesting the comparison to be an accurate representation of better reporting by test-treatment RCTs than other, contemporary trials. Although clearly a positive impact on the ability to interpret trial findings, poor description of the measurement of these outcomes by half the included studies somewhat hindered the ability to make full use of results. Of particular concern was the tendency to omit the timing of primary measurements, leaving the reader uninformed regarding when observed effects should become manifest in a similar patient group. Nevertheless, the superior reporting quality by comparison to treatment and complex intervention trials would suggest that there are no particular barriers to the effective reporting of outcome measurement in test-treatment RCTs.

5.4.3 Trials provide incomplete and insufficient descriptions of test-treatment interventions

Documenting the interventions used to test and treat patients proved to be by far the most poorly reported aspect of these trials: only three trials provided a full written description of all methods and decision-making in all study groups, and only one of these also provided a care pathway diagram[(T105)]. This is significantly poorer than found even in reviews of complex interventions. Description of back pain interventions were absent in 87% of examined RCTs[238] and surgical interventions in more than half[239]. Glasziou and colleagues found better reporting of interventions in drug compared to non–pharmaceutical trials[136]. Other researchers report similar findings; in a direct comparison of the methodological quality of 60 pharmaceutical and 50 complex non–pharmaceutical treatment RCTs, drug interventions were more often described in enough detail to be reproduced than non–pharmacological interventions, though the latter could still be replicated in 82% of cases[240].

Moreover, by focussing on the *frequency* of reporting interventions this study is likely to have overestimated reporting *quality*, since components were judged to have been described even if only partially outlined. While this approach was necessary, since determining the adequacy of descriptions would have required significant input from a wide range of clinical experts, it has masked the relative importance of reporting omissions whose impact on interpretability are likely to vary according the clinical setting. For example, test–treat strategies that seek to investigate presentations for which there are multiple potential diagnoses, and consequently more treatment options, may be completely un-interpretable if only partial description is given; conversely protocols attempting to confirm well-defined disease entities that can be clearly dichotomised into two simple treatment options may be more easily reconstructed.

The importance attached to full and clear accounting of any healthcare intervention was highlighted in chapter 1 (p. 31–32). Not only is it required to interpret the meaning of a trial's results, but it is crucial to clinicians' ability to identify desirable processes and replicate them safely in practice[136–137,147]. The considerable difficulties encountered in this review in trying to decipher how diagnoses were made and how patients were subsequently managed emphasises that test–treat strategies may be particularly sensitive to poor future implementation. Most important to the utility of results is the delineation of how test data should be interpreted and how the resulting diagnostic categories should be utilised to guide management, while the subsequent selection of treatments must be made explicit[22]. Yet decision–making processes were amongst the most poorly reported aspects of test-treatment trials.

Although the reproducibility of interventions was not measured directly in this review, failure to outline diagnostic procedures, treatment procedures and decision-making criteria must mitigate a reader's ability to reproduce these processes in practice. Based on the level of reporting encountered, the current study must conclude that in only three trials is this likely to have been possible[(T76,T98,T105)].

Barriers to the documentation of test–treat interventions

It is certainly true that the complexity of these interventions makes full and accurate documentation difficult. The variations in frequency of elements described according to the types of tests being evaluated suggests that some diagnostic tests, and their associated care pathways, are more difficult to outline. However, there is no specific evidence that this difficulty is solely due to the *complexity* of clinical processes and decision-making. Clinical consultations, for example, were largely poorly reported. Although these strategies were often simple, consultations are likely to be less amenable to standardisation, and therefore more difficult to translate into a prescriptive format. On the other hand, complex endoscopic techniques, often part of multistage diagnoses, were often well reported. A closer examination of how protocols were reported reveals specific conceptual barriers that could be responsible for the observed difficulties.

Firstly, a considerable focus was placed on describing experimental techniques to the considerable detriment of what they were being compared to. Whether reflecting a widely–held assumption that 'conventional investigations'(T56) are standardised and therefore need not be made plain or appropriate methods cited, comparator strategies are equally under evaluation in any RCT[156] and not reporting them poses an irrevocable impediment to the interpretability of effect estimates. Without knowing which tests and treatments are being administered to one study group, it is impossible to discern how favourable effects have benefited patients.

Secondly, trials focussed on reporting diagnostic procedures, commonly omitting to elaborate how test results should be used to arrive at diagnostic decisions, how this should inform management, and which treatments should subsequently be prescribed. This suggests that trialists may perhaps be following the more familiar reporting practices of standard pharmaceutical RCT evaluations, regarding tests as singular interventions rather than as part of a broader test–treat strategy. Tellingly, a single trial identified itself as a complex intervention RCT(T107). The common failure to recognise that decision–making forms part of test-treatment interventions may explain why such elements were so rarely described.

Lastly, trials exhibited difficulties in identifying all diagnostic tests used in an intervention. This difficulty varied according to the clinical context and study group. Greater obscurity was encountered in descriptions of interventions using batteries of tests, such as the 'accelerated diagnostic protocol' for assessing the likelihood of progression to full stroke in transient ischaemic attack patients(T86), or the trial evaluating a specialist unit of care for establishing the cause of syncope discussed above(T53). Even though these interventions did not appear amongst the most poorly reported (Tables 5.3 & 5.4), this is probably an artefact of the appraisal criteria used by the author, where protocol elements were judged as 'reported' even if some tests and decision–making processes were not provided. There was also a tendency across many studies to disregard the role of clinical examinations in diagnostic pathways, with a predilection only to recognise more technological procedures as true ‘tests’. This point is illustrated in a trial comparing the standard use of early MRI to investigate back pain with a more selective approach providing MRI only on clinical indication(T36). The authors do not acknowledge the role of the ‘clinical indication’ as a diagnostic process, yet they document clinical indication test–negative patients when discussing the reasons for patients not proceeding to MRI. As a consequence, precisely which signs and symptoms were used to indicate the need for an MRI were not reported. Not only is it difficult then to replicate this strategy, but this non–recognition has led to the failure to document important variations in how clinicians use such additional diagnostic information, even though it could well have influenced the overall effectiveness of the test-treatment policy.

These impediments to full and proper reporting of interventions may therefore reflect misplaced perceptions regarding what should be evaluated in a test-treatment trial. It could also reflect the common publication of test-treatment RCTs in specialist journals, which may presuppose a high degree of assumed knowledge amongst their readership. While many test-treatment strategies are highly familiar to specialists, in reality there is likely to be wide variation in how techniques are implemented, and how decisions are made.

These observations suggest that the difficulties in describing test-treatment interventions are surmountable. However, the common absence of decision–making criteria may reflect the inability to standardise diagnostic and therapeutic decision–making processes. Although the suitability of interventions to being standardised was not measured directly, this might explain the dearth of decision–making in at least some trial protocols. Trialists may have sought to perform pragmatic evaluations to capture important variations in practice between study sites. Indeed, several reports describe themselves as ‘pragmatic’ studies in which treating clinicians were given discretion regarding how management decisions were made(T5,T12,T19,T34,T36,T57–8,T60,T66,T68,T73,T77,T92,T97,T107). Nonetheless, several of these documented at least some of the interventions actually used during the study period(T12,T34,T57,T66,T73,T77,T92,T97,T107), suggesting that test–treatment strategies not suited to standardisation can also provide informative documentation of patient management.

5.4.4 Study limitations

The primary limitation to this review concerns the process of data extraction, which was conducted solely by the author. The lack of a second, independent reviewer means it is possible that the review's findings in part reflect certain inaccuracies made in error by the author. These effects should have been minimised by taking a systematic approach when examining each trial as well as using a standardised extraction tool, developed from existing, validated guidelines.

A second important issue concerns whether the results presented here are generalisable to all test-treatment RCTs. Efforts to obtain a representative sample of trials were made by ensuring that the search strategy (reported in chapter 2) used only methodological terms, and did not target any specific tests, diseases or clinical disciplines. It may be, however, that particular disciplines are more or less likely to use these generic methodological terms than others, in which case the project cohort will not be entirely representative of all test-treatment trials published during the search period (2004–2007). Both the capture–recapture analysis and independent check (reported in chapter 3) concluded that trials have been missed; some were missed by the search strategy whilst at least three others were overlooked by the author during screening. While it is not possible to compare the content of all missed trials to those included in the project cohort, it is unlikely that their inclusion would drastically change this review's conclusions. Both types of omission are likely to be associated with a poorer quality of reporting, and so this review may be presenting a more favourable conclusion than is realistic.

Reporting quality has improved over the last 15 years[118,153] so it is possible that trials published since 2007 offer superior accounts of their conduct. However, in the absence of reporting guidelines specific to test-treatment RCTs quality is unlikely to be substantially improved in more recent trials.

Only primary outcomes were extracted, since the aim was to identify the main methodological issues with test-treatment RCTs. Although problems in measurement and description of secondary outcomes will be typified by those encountered in primary endpoints, the range of outcome types may be different. When looking more broadly at all trial outcomes the issues cited here could vary in their relative impact, particularly if secondary aims address a greater number of psychosocial, patient health or physician-led questions.

Lastly, this review was limited to examination of three aspects of trial reporting. These were selected because they are considered key to interpreting all trials, and could be appraised objectively without recourse to detailed clinical expertise. Nonetheless there are several other aspects of test-treatment interventions that will probably require a detailed appraisal of reporting quality in the future. Chief amongst these is the documentation of how interventions were actually administered. Fidelity to the intended intervention is known to be poorly reported in all trial types[137], and has been found particularly wanting in studies of complex therapeutic or surgical interventions[239]. Likewise, care–provider skill and experience are integral to clinical decision–making and, as is recommended for the documentation of complex interventions in general[83,147], any appraisal of protocol reporting quality is arguably incomplete without also considering this issue. Although neither of these was investigated in this review, the difficulties encountered in piecing together the composition of test–treat interventions from descriptions of intended patient management were such that reporting of actual practice is likely to have been even poorer. As the scope of this thesis commanded the retrieval of test-treatment RCTs conducted in all medical disciplines, it was also outside the present realms of feasibility to perform an evaluation of the replicability of test-treatment interventions since this would have required wide consultation with clinical experts from all included disciplines. Consequently, the adequacy of protocol reporting has been judged on the basis of a minimal descriptive presence. Given that the ability to reproduce test-treatment interventions is likely to require more comprehensive reporting than the criteria used here to appraise reporting quality, it would be appropriate to conclude, if somewhat tentatively, that the replicability of published test–treat protocols is likely to be even poorer than is suggested in the present review.

5.5 Conclusions

To conclude, this review finds a clear need for improvement in the reporting of test-treatment RCTs. This cohort of 103 trials was characterised by several limitations that may hinder the interpretation and application of trial findings.

1. Incomplete accounting of participant flow is likely to hamper a full interpretation of the meaning of observed test-treatment effects;
2. Incomplete reporting of how primary outcomes are measured is likely to obstruct the replication of measurements, and therefore lead to problems when seeking to compare effects between trials;
3. Insufficient documentation of interventions precludes the interpretation of how test–treat processes are related to differences in trial outcomes, and is also likely to prevent the translation of beneficial interventions into clinical practice.

These failings are partly explained by the suboptimal quality of reporting generally found in all RCTs[118,152–153]. However, the considerably inferior quality of intervention documentation encountered in test-treatment trials, when compared to standard, non–complex intervention trials, does appear to confirm that this issue is more prominent in test-treatment trials.

Nonetheless, the review suggests that all three issues are likely to be surmountable by improved reporting. Adherence to the CONSORT guidelines will be valuable for improving reporting of participant flow and primary outcome measurement, however test-treatment interventions are shown to require more detailed attention to several components than treatment interventions, or even other types of complex interventions, if they are to provide useful information. Since the review findings would indicate that trialists may not always be aware of what needs to be reported, more specific guidelines are likely to be necessary to achieve the necessary improvement in reporting of these complex interventions.

The following chapter continues the analysis of test-treatment RCTs by appraising the methodological quality of included trials in order to determine how valid and reliable these trials are.

6

Trials of Test-Treat Strategies:

The methodological quality of published trials

This chapter presents the third analysis of the project cohort identified in chapter 2 and characterised in chapter 4. Its aim is to address the third challenge levelled at test-treatment RCTs, that they risk producing unreliable evidence due to the feasibility of implementing the required methodological controls when evaluating such complex interventions. This is achieved by appraising the methodological quality of included trials. The chapter ends with a discussion of its findings, with particular regard to whether the quality observed is comparable to that found in treatment RCTs and complex intervention RCTs.

Inadequate design and study conduct has been empirically demonstrated to damage the reliability and utility of RCTs by exposing their results to bias[106–108].

The complexity of test-treatment interventions is claimed to predispose them to particular difficulties in implementing the measures necessary to limit bias. Test results must be interpreted by clinicians, and diagnoses recounted to their patients hence it may be impossible to eliminate both performance bias and ascertainment bias[22,79,176–177]. The need for patients to progress through multiple interventions (tests and treatments) may increase the proportion who drop–out, and since the quality and information patients receive differ according to the interventions used, these trials may also be susceptible to differential drop–out, which places them at increased risk of attrition bias[176]. Sample sizes must be considerably larger in order to account for the probability that effects are only experienced in patients who receive different care as a result of their diagnoses; trials omitting this inflation risk being underpowered to detect patient health effects, while it may not be feasible to recruit the necessary number of patients in adequately powered trials[6,176]. Since diagnostic decision–making is highly variable, trials must pre–specify how test results lead to diagnoses and treatment plans in order to be sure of how interventions are creating the observed effects[22,76,79]. However, it will be challenging to report the multiple interventions used and document decision–making to the extent necessary to allow findings to be interpreted, compared across studies, and to inform how beneficial interventions should be translated into practice.

These are serious claims, though there is little existing evidence to support them. The following study has been designed to address these challenges by examining the extent to which they are encountered in completed trials. It presents a systematic appraisal of the methodological quality of test-treatment RCTs ascertained in chapter 2 in order to evaluate the extent to which these trials are susceptible to the biases and challenges to utility that are claimed to confront them. Since the cause of difficulties in conducting and reporting these trials is hypothesised to lay with the complex make–up of test-treatment interventions, the analysis ultimately aimed to determine how they perform by contrast to standard treatment RCTs.

6.1 Methods

To evaluate the reliability of test-treatment RCTs, assessments of methodological quality focussed on six key indicators of trial internal validity, including control of selection bias at recruitment, attrition bias during primary analysis, information biases arising from the differential behaviour of participants, care–providers and outcome assessors (performance bias and ascertainment bias), and the minimisation of type II error. The review therefore sought to answer the following questions:

1. Did methods of sequence generation adequately protect against selection bias?
2. Did methods of allocation concealment adequately protect against selection bias?
3. Do trials control for performance and ascertainment bias?

4. Can trials control for performance and ascertainment bias?

5. Were primary analyses conducted appropriately, to minimise the effects of selection bias and attrition bias?

6. How did studies determine sample size?

6.1.1 Design of a quality assessment tool

In the absence of an existing quality assessment tool specific to test-treatment RCTs, standardised data collection and appraisal forms were designed. Items required to answer the six review questions were identified from three validated, internationally accepted standards for the conduct and reporting of RCTs: the CONSORT checklist[121], the extension of the CONSORT statement for non-pharmacologic therapy interventions[147] and the Cochrane Collaboration's 'Risk of Bias' tool[123]. Table 6.1 lists the items extracted.

This new quality appraisal tool was tested by extraction of five test-treatment trials, randomly selected from the project cohort. Minor changes were made to improve the standardisation of both data collection and quality assessment.

6.1.2 Data Extraction

All included trials with at least one publication of results were appraised, following the approach detailed in the review of reporting quality (chapter 5). Data were extracted to a purpose built relational database (Microsoft Office Access 2007) and appraised for methodological quality using the methods reported below. Extraction and quality assessment were performed by the author.

6.1.3 Interobserver reliability

In order to ensure consistency in the application of quality criteria, a 65% convenience sample was selected prior to any extraction and independently assessed by a second reviewer with considerable experience in conducting systematic reviews. Duplicate

Trial Methods Objectives	Item
1. Did methods of sequence generation adequately protect against selection bias?	*Was the method of sequence generation reported?* *Was it adequate?*
2. Did methods of allocation concealment adequately protect against selection bias?	*Was the method of allocation concealment reported?* *Was it adequate?*
3. How often were participants, care–providers and outcome assessors blinded to test-treatment interventions?	*Were clinicians/participants/outcome assessors blinded to the test used?*
4 How often was it feasible to blind these individuals?	*Was it feasible to blind clinicians/participants/outcome assessors to the test used?*
5. Was the primary analyses conducted appropriately, to minimise the effects of selection bias and attrition bias?	*Could the outcome be measured in all randomised participants?* *Was the outcome measured consistently in all arms?* *Was the primary analysis complete?* *Were outcome responses missing?* *How many were missing in each arm?* *Were reasons for missing data reported?* *What methods were used to deal with missing data?* *Did investigators exclude participants from the analysis?* *How many were excluded in each arm?* *Were the reasons for exclusion reported?* *Were patients lost to follow-up?* *How many were lost in each arm?* *Were the reasons for loss to follow-up reported?* *Did trialists report an 'intention-to-treat' analysis?* *Were patients analysed according to their randomised allocation, regardless of the intervention actually received?*
6. How did studies determine sample size?	*Was a power calculation described?* *What was the estimated target sample size?* *Was the outcome on which the power calculation was based reported?* *Was this the same as the primary outcome?* *If estimated sample size was not achieved, what were the reported reasons?*

Table 6.1: Items extracted to evaluate the methodological quality of test-treatment RCTs.

assessment was only performed for standardised criteria that are commonly used to evaluate methodological quality, namely the quality of sequence generation methods, allocation concealment methods, presence of blinding, and documentation of numbers analysed for the primary outcome.

Neither reviewer was blinded to publication details (authors, institutions, journal) or to trial results, since the benefits of doing so have not been consistently proven[100,241–242]. Cohen's kappa statistic was calculated to convey chance-corrected agreement. The results presented below reflect the consensus reached between the author and independent reviewer, where applicable. Disagreements in quality assessment were identified by the author, and each instance discussed with the independent reviewer to reach consensus.

6.1.4 Appraisal of Methodological Quality

Sequence generation and allocation concealment

Randomisation and allocation procedures were extracted for each trial. Methods were appraised for their adequacy in preventing selection bias using the rigorous criteria recommended for the evaluation of treatment RCTs by the Cochrane Collaboration[123] (Table 6.2). Quality of methods were categorised as 'unclear' if the information provided was insufficient to judge the presence of a random element in the generation of the allocation schedule, or the overall predictability of the allocation sequence.

Blinding conduct

Studies were examined for clear reports of whether participants, care–providers (defined as those responsible for patient management) and outcome assessors had been masked to the identity of tests used for decision-making during the trial. All such attempts were extracted to characterise how blinding had been achieved in these studies. Since it is theorised that blinding may not always be possible, the reasoning trialists provided to explain the absence of blinding was examined.

	Adequate	Inadequate	Unclear
Sequence Generation	Clear description of a method to randomly generating numbers, e.g. • computer random number generator • random number table • tossed coin • shuffled cards/envelopes • throwing dice • drawing lots • minimisation	Clear description of a number generation method that is partly or fully systematic, e.g. • alternate assignment • birth date • consultation date • hospital number • judgement of clinician • patient preference • test results • availability of intervention	Missing data or obscure description with unclear indication of random component to sequence generation, e.g. • generic reference to 'randomisation' • use of a randomised list • randomisation performed by computer • random assignment • randomisation schedule
Allocation Concealment	Clear description of an attempt to conceal the order of allocation from study recruiters and patients, e.g. • central remote-site randomisation procedure (e.g. telephone, independent trial office etc) • concealment of allocation instructions (sequentially numbered, sealed and opaque envelopes)	Clear description of a predictable order of allocation, e.g. • open random allocation schedule (e.g. selection of next random number by treating staff) • unconcealed assignment envelopes (e.g. unsealed, transparent or not sequentially numbered) • non-random sequence generation which is easily predictable (e.g. alternate/rotational assignment, hospital number, birth date etc)	Missing data or obscure description that does not allow the predictability of schedule to be judged, e.g. • generic reference to a masked or concealed allocation process • unclear safeguards for assignment envelopes • centralised procedure with no reference to remote-site, or of unclear location

Table 6.2: Criteria for appraising methods of sequence generation and allocation concealment.

Blinding feasibility

An attempt was also made to judge the feasibility of blinding in all included studies. This subjective assessment was made on a case–by–case basis, since it was recognised that the ability to mask individuals is highly dependent on the clinical and comparative context. For participants and care–providers the ability to blind was determined by reference to the individual clinical setting, including the similarity in characteristics of tests administered

Outcome Type	Examples
Patient reported	Pain, quality of life
Patient–outcome assessor contact required	Walking speed, function
Patient–outcome assessor contact not required	Appearance of joint structure (X-ray)
Clinical events and therapeutic outcomes determined by interaction between patient and clinician (physician-driven data)	Length of hospital stay, treatment failure
Clinical events and therapeutic outcomes assessed from data on medical forms	Death, treatment prescription

Table 6.3: Criteria for categorising outcome assessors[243].

and the nature of their comparison (replacement, add-on or triage). Since the ability to blind outcome assessors also depends on the endpoint being measured, feasibility was judged separately for each outcome taking into account how the outcome was measured, the identity of the outcome assessor and whether the measurement was determined objectively or subjectively. The identity of the outcome assessor was extracted directly, or surmised from description of measurement methods where possible. Each outcome was categorised according to the type of contact between participant and outcome assessor required for measurement to take place, using predetermined criteria developed by Boutron and colleagues for assessing the quality of non-pharmacological RCTs (Table 6.3)[243]. This approach was conducted in order to facilitate subsequent assessments of the feasibility of blinding outcome assessors. The degree of interaction between participants, care–providers and outcome assessors is particularly convoluted in trials evaluating complex interventions, where multiple phases of decision-making may be the subject of evaluation. Boutron's method was selected since it was designed to categorise outcomes and investigate the feasibility of blinding in trials of non-pharmacologic interventions which, though not test-treatment RCTs, are nonetheless complex intervention trials and thus considered suitable for the present analysis.

Judgements were grouped into three categories, following published methods[243]: blinding was judged as 'feasible' if the means required to blind were considered common or could be applied simply; blinding was judged 'difficult' if blinding could have been conducted, but would have required the implementation of solutions that differ considerably from normal clinical practice; and lastly blinding was considered 'impossible' if the reviewer thought it not physically or ethically practicable to mask the identity of the test, even using creative solutions. These classifications are illustrated with examples in Box 6.1.

Appropriateness of the main analysis

Appropriateness was judged according to five criteria: 1. whether outcomes were measured in the whole study population (and not a subgroup), 2. whether outcome measurement was consistent across trial arms, 3. whether patients were analysed in the groups to which they were randomised, 4. whether analyses were complete, and if not why 5. whether trialists conducted an 'intention–to–treat' analysis.

Subgroup analyses

Published analyses were examined to determine whether the denominators used for calculation constituted the whole randomised population. Primary outcomes measured in subgroups of the study population were considered inappropriate due to the ensuing risk of selection bias when comparing subgroups that may reflect non–random differences in composition[135].

Consistency of outcome measurement

Details of outcome measurement were examined to establish whether the same method of ascertainment was used for all arms in each trial. If methods differed, the type of endpoint (see chapter 5 p.130–134) was used to judge the comparability of resulting findings. For example test performance outcomes (e.g. diagnostic yield or therapeutic yield) by definition must be measured using the tests under comparison, and so were judged to

Box 6.1: Definition of the blinding feasibility judgement categories used, with examples.

Feasible **Clinical setting and nature of comparison easily accommodate blinding, e.g.**

PARTICIPANTS: *Comparison of tests that are very similar or do not carry any risk of procedural morbidity, for example d–Dimer blood test vs. ultrasound for diagnosis of deep vein thrombosis(T3).*

CARE–PROVIDERS: *Interpretation of test results are not made by the treating physician such that generic reports can be produced, for example in the comparison of two laboratory biochemical tests (CK/CK-MB vs. Troponin I) to identify chest pain patients at high risk of adverse events(T26).*

OUTCOME ASSESSORS: *Outcome measurements do not involve contact with patients, for example ascertaining the rate of colonic rebleeding during follow–up from medical notes(T30).*

Difficult **Blinding possible, but requires substantial modification of normal clinical practice involving 'creative' solutions, e.g.**

PARTICIPANTS: *Using simulated test procedures, for example comparing the replacement of clinical examination with computed tomography (experimental test) for the management of possible intracerebral injury, which would involve giving comparator arm patients sham CT scans(T68).*

CARE–PROVIDERS: *Providing treating clinicians who are traditionally involved in test interpretation with sham test results by a third party, for example requiring non–treating physicians to take samples, interpret results and produce generic reports when comparing Bronchoalveolar lavage and quantitative culture with Endotracheal aspiration and nonquantitative culture to identify the organism causing pneumonia(T62).*

OUTCOME ASSESSORS: *Outcome measurements generally conducted by treating physicians that need to be rearranged so they are conducted by independent physicians, for example measuring the incidence of retained products of conception during follow–up using gynaecological examination and transvaginal ultrasound(T76).*

Impossible **Blinding is not physically possible or ethically acceptable, e.g.**

PARTICIPANTS: *Comparison of invasive tests that render sham procedures unethical, for example endoscopy vs. 13-Carbon urea breath test for detecting Helicobacter pylori(T44).*

CARE–PROVIDERS: Test *and treat stages are performed during the same procedure, for example comparing white–light cystoscopy with fluorescence cystoscopy for the detection and removal of bladder cancer(T1).*

OUTCOME ASSESSORS: *Patient–reported outcomes where blinding patients is impossible, for example dyspepsia symptom relief measured on patient questionnaires to evaluate two complex, partly invasive testing strategies where it is not practical or ethical to mask patients(T40).*

have been measured appropriately. Conversely, the measurement of disease rates should be made using the same test, and so were judged to have been ascertained inappropriately if different tests were used in different arms.

Incomplete analysis: patient exclusions, missing data and differential attrition

Participant flow data were examined to reveal whether the main analysis was complete, to quantify the amount of data missing from study groups, and to determine the likelihood of bias arising from differential attrition between comparative arms.

For each trial the number of patients randomised was compared with the number analysed to determine the magnitude of attrition for all study groups. Analyses were considered incomplete if the number analysed was less than the number randomised. Attrition was defined as the absence of data (for any reason) for the primary outcome measured at the primary time-point, if specified. Otherwise the time-point used in sample size calculations was used, and if this also was not available then the most complete analysis was selected. When more than one primary outcome was reported, patient outcomes were preferentially extracted to reflect their primacy over process outcomes as measures of clinical effectiveness.

Differential attrition was arbitrarily considered at two levels, as ≥5% and as ≥20% difference between arms, following the approach advocated by the Centre for Evidence Based Medicine when judging the quality of comparative evidence of effectiveness[244].

Reasons for missing data were extracted, and classified as either investigator-determined exclusion post-randomisation or loss to follow-up. As recommended by CONSORT[121], reporting of exclusions was examined to determine whether the apparent absence of attrition corresponded with an explicit statement of no exclusions.

Analysis according to randomisation

Participant flow data for each trial were examined to establish whether deviations from protocol had occurred, and if so whether these patients had been analysed according to their allocated groups regardless of the test actually received.

Use of intention–to–treat principle

Reports clearly stating that primary analysis was by the 'intention–to–treat' (ITT) principle were extracted. These were compared against both components needed in an ideal ITT analysis[123], namely whether all study patients were analysed as randomised, and whether analyses were complete.

Determination of sample size

The assessment of sample size quality was limited to whether sample size calculations were reported, extraction of target numbers, appraising whether these calculations used primary trial endpoints, and establishing whether target sample sizes were reached. Reasons for any deficits were also extracted. Preliminary studies suggest that power calculations for test-treatment RCTs are likely to require estimates of test sensitivity and prevalence of the target condition[176–177]. The author initially attempted to trace these data for one included trial[(T62)] in order to replicate the reported power calculation and so determine its adequacy, however published information could not be found. In view of the wide variety of diagnostic settings encountered in the included studies, and extensive bibliographic searching needed to identify the required diagnostic performance parameters for each trial, it therefore became necessary to limit this analysis to the presence of power calculations and the distribution of attained sample size across included trials.

6.1.5 Analysis

The objectives of this review are to describe the frequency with which test-treatment trials were found to have used adequate methods to limit bias and enhance the validity of results. Consequently, the analysis presents a descriptive synthesis of these findings using percentages that reflect the categorical nature of the data. Comparisons between these frequencies were used to enhance the description of findings, and aimed to highlight potential associations between particular methodological items and aspects of the study settings, as characterised in chapter 4. The author did not intend to evaluate specific hypotheses regarding these associations, but rather to generate hypotheses for how easily the methods of trial design could be employed to conduct reliable and informative test-treatment RCTs. As a consequence testing for the statistical significance of these comparisons was not appropriate.

Data were exported from the extraction database to Microsoft Excel 2007 for sorting and analysis.

6.2 Included studies

As for the review of reporting quality presented in chapter 5 (p.122–123), methodological quality was assessed for the same 103 trials with full results publications[*]. These evaluated 119 experimental and 105 comparator interventions.

6.2.1 Inter−observer reliability

For the sample of 66 independently assessed trials, agreement between reviewers was substantial for assessing the adequacy of both sequence generation and allocation concealment methods (Table 6.4). During the subsequent consensus meeting, it became apparent that most disagreements (12/13 for sequence generation and 9/11 for allocation concealment methods) concerned differences in opinion as to whether more meagre descriptions were sufficient to allow a quality judgement to be made, while the remainder were due to errors in data extraction.

[*] Citations for included test-treatment RCTs are prefixed with a ‘T’

2nd Reviewer		Author				*A_{Obs}	†A_{Exp}	κ	95% CI
		Adequate	Inadequate	Unclear	Total				
Sequence Generation	Adequate	33	0	11	**44**	0.809	0.483	**0.630**	**0.5–0.8**
	Inadequate	0	1	0	**1**				
	Unclear	1	1	21	**23**				
	Total	**34**	**2**	**32**	**68**				
Allocation Concealment	Adequate	24	0	10	**34**	0.838	0.459	**0.701**	**0.5–0.8**
	Inadequate	0	2	1	**3**				
	Unclear	0	0	31	**31**				
	Total	**24**	**2**	**42**	**68**				

Table 6.4: Inter-reviewer agreement in assessing quality of randomisation and allocation concealment methods.

*Observed agreement

†Expected agreement

2nd Reviewer		Author			*A_{Obs}	†A_{Exp}	κ	95%CI
		Blind	Not blind	Total				
Patients	Blind	3	0	**3**	1.000	0.913	**1.000**	**x**
	Not blind	0	63	**63**				
	Total	**3**	**63**	**66**				
Care-providers	Blind	3	3	**6**	0.955	0.872	**0.645**	**0.3–1.0**
	Not blind	0	60	**60**				
	Total	**3**	**63**	**66**				
Outcome assessors	Blind	16	0	**16**	0.970	0.617	**0.921**	**0.8–1.0**
	Not blind	2	48	**50**				
	Total	**18**	**48**	**66**				

Table 6.5: Inter-reviewer agreement in assessing the conduct of blinding.

x – not calculable

*Observed agreement

†Expected agreement

Agreement was perfect when judging the presence of patient blinding, near–perfect for outcome assessor blinding and substantial for blinding care–providers (Table 6.5). All discrepancies were due to inaccuracies in data extraction; the three disagreements regarding whether care–providers had been masked owed to the misidentification of whether personnel described as blind were treating physicians performing the experimental or comparator test.

6.3 Results: Do trials adequately control for selection bias?

Evaluation of the methodological quality of included trials began with appraising the adequacy of their methods to limit the bias in the creation of study groups (Appendix E.1).

6.3.1 Adequacy of sequence generation

Approximately half the trials (59/103) reported robust methods of sequence generation, judged as likely to have permitted a truly random order of patient allocation. Two studies used non–random and hence inappropriate methods, either allocating systematically by the sum of the day and month of birth (even number assigned control)(T23), or employing 'random sampling' to achieve cohorts with similar sizes(T28). The remainder (41%, 42/103) failed to report their methods of sequence generation in enough detail to enable an independent judgement of methodological quality (Figure 6.1). Of these, 23% (10/43) made no reference to sequence generation, 30% (13/43) reported the term 'randomisation' without describing the method used, while 44% referred to either 'randomisation by computer' (7/43), 'block randomisation' (8/43) or a 'centralised system' of randomisation (4/43) without reference to whether the schedule was generated using a random component.

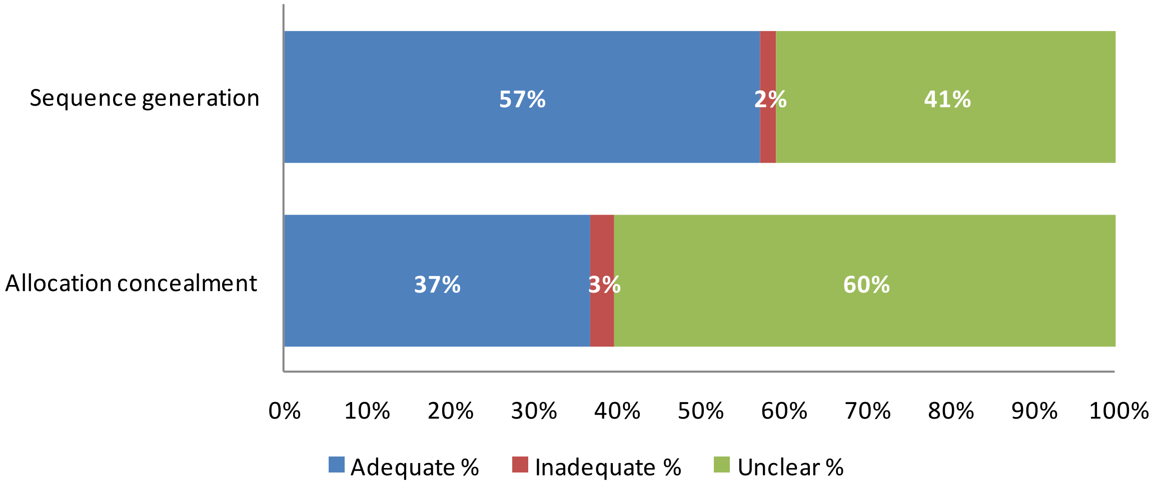


Figure 6.1: Adequacy of quality for sequence generation and allocation concealment methods.

6.3.2 Adequacy of allocation concealment

Approximately one in three trials (38/103) were judged to have adequately concealed the randomisation schedule from study recruiters (generally physicians) and patients. Three studies (3%) overtly employed inadequate methods; the inadequately randomised trial that allocated patients by date of birth(T23) used a clearly predictable schedule, while two cluster randomised trials did not conceal established cluster allocations from participant recruiters until all patients had been recruited(T48,T97). However the great majority of studies, 60% (62/103), did not provide sufficient detail to make an independent judgement of methodological quality. Over half of these (34/62) did not refer to allocation concealment. The balance provided incomplete descriptions, including unclear safeguards for assignment envelopes (19/62), opaque reference to a centralised procedure with no reference to a remote-site location (6/62), or statements that concealment was carried out without reporting the methods used (3/62).

Group identified as blind	Number of trials reported as:					
	Blind	(%)	Not blind	(%)	Unclear	(%)
Patients	5	(5)	30	(29)	68	(66)
Care-providers	4	(4)	33	(32)	66	(64)
Primary outcome assessors*	22	(21)	14	(14)	67	(65)

Table 6.6: Frequency of blinding in test-treatment RCTs (N=103)
* judged as blind if measurement of ≥ 1 primary outcome conducted by a blinded assessor.

6.4 Results: Do trials control for performance & ascertainment bias?

6.4.1 Conduct of blinding

Reports of blinding were few in number and of poor quality. Overall authors failed to indicate whether blinding had been conducted in approximately two-thirds of studies (Table 6.6). Participants, care–providers and primary outcome assessors were masked to the identity of the diagnostic strategy used for treatment decisions in 5%, 4% and 21% of trials respectively. A summary of judgements is provided in Appendix E.1 (p.410).

Participants

Patients were blinded to the test used in four add-on comparisons(T69,T98,T100,T103) and one triage comparison(T74) (Table 6.7).

All four add-on trials evaluated the addition of single, straightforward non-invasive strategies. Patient blinding was achieved either by administering experimental and control tests to all participants and masking them from test results(T69,T100,T103), or by conducting both tests in the same sample of tissue (T98). In the latter case, the tests being compared were laboratory-based examinations of preimplantation embryos, hence the test technologies are ordinarily performed and interpreted in the absence of patients. To preserve blinding for the remainder of the study however, trialists prevented patients from entering laboratories and did not reveal information on the number or quality of embryos transferred during IVF.

The triage trial blinded patients to a more complex protocol, introducing a non-invasive test to select which patients would proceed to the invasive control test. Patients with suspected pulmonary embolism (PE), requiring a confirmatory diagnosis to direct further treatment, were randomised to management by the standard ventilation-perfusion (V/Q) scan or initial triage for ruling out PE(T74). This combined 'BIOPED' test, comprised two laboratory assays (D-dimer and alveolar dead space measurement) and a clinical prediction rule (Well's seven-variable clinical model), sought to eliminate PE as a possible cause thus sparing these test-negative patients the more invasive V/Q scan. While test-negative patients could be safely redirected to other treatments, those with a positive 'BIOPED' score would proceed to further V/Q testing for a definitive diagnosis. In order to blind the triage stage, the BIOPED was given to all patients. Since receipt of a V/Q scan risked revealing the allocation, all patients also underwent a V/Q scan. In order to ensure that rule-out occurred only as a result of the triage test results, BIOPED negative patients received a sham V/Q procedure, while test positives received a true V/Q scan.

Care–providers

Attempts to blind treating physicians to the identity of the allocated test strategy were reported in four trials (4%)(Table 6.6), three of which also blinded patients(T74,T98,T100).

Blinding was easily achieved during the evaluation of preimplantation embryo analysis, as treating physicians (gynaecologists) were not involved in the conduct or interpretation of either test(T98) and, as with patients, they were prevented from entering laboratories and or receiving information on the number or quality of embryos transferred during IVF.

Comparison (Control vs. Experimental)	Purpose of the experimental test	How patients were masked	How care-providers were masked	Trial Ref.
V/Q scan vs. BIOPED* ± V/Q scan	To rule out suspected PE and identify the subgroup of patients with PE that will benefit from further investigation	All patients given real BIOPED, results kept from patients and care-providers. EXP arm: BIOPED negative patients receive sham V/Q, while test positives receive true V/Q scan. Sham results disseminated to patients and care-providers to maintain masking.	For BIOPED negative patients in EXP arm receiving a sham V/Q, clinicians were sent a sham V/Q report indicating no PE. BIOPED positive patients received a real V/Q scan, as did all patients in the comparator arm.	T74
Routine clinical work-up† vs. Echocardiogram + Routine clinical work-up	To diagnose target organ damage in patients with essential hypertension and thus identify those who will benefit from treatment (angiotensin II receptor antagonists).	All patients given real Echocardiogram, performed by non-treating cardiologists who were masked to group allocation and instructed not to inform patients about echo findings.	Care-providers not masked.	T69
Embryo morphologic score vs. genetic screen + Morphological score	To rule out chromosomal abnormalities in embryos of women at risk, thus identifying suitability for implantation.	Testing took place in laboratories, at a distance from patients and treating physicians. "..To maintain blinding, they [patients and physicians] were not allowed to enter the laboratories and were given no information about the number and quality of the embryos to be transferred." (10:2)		T98
Standard tests‡ vs. NT-proBNP + standard tests	To detect heart failure in patients with dyspnoea and direct towards appropriate treatment	All patients receive NT-proBNP prior to randomisation.	Care-providers reported as blind, but referring to adjudicators of clinical outcomes. Treating clinicians not blind.	T100
Clinical consultation vs. MRI + clinical consultation	To rule out organic disease in patients with knee disorders in order to identify those in whom arthroscopy could be avoided	All patients were given real MRI scans. No further details provided on the likely success of blinding.	Treating physicians could not be masked (since the comparison was whether MRI changed decisions based only on clinical consultation).	T103
V/Q scan vs. CTPA	To confirm suspected PE and identify those in need of further investigation (leg vein US)	Patients not masked	Treating physicians issued generic pulmonary imaging reports that indicated positive for PE, nondiagnostic, or no evidence for PE.	T91

Table 6.7: Summary of trials that attempted to blind participants and/or care–providers. EXP: Experimental arm; BIOPED: Bedside Investigation of Pulmonary Embolism Diagnosis; CTPA: computed tomographic pulmonary angiography; MRI: magnetic resonance angiography; NT–proBNP: N-terminal pro-B-type natriuretic peptide; PE: Pulmonary embolism; US: Ultrasound; V/Q: ventilation–perfusion scanning;

* d-dimer, wells clinical prediction rule, alveolar dead space fraction

† History, physical examination, routine blood tests, electrocardiogram, urinalysis and 24hr ambulatory blood pressure monitoring

‡ reporting incomplete: "standard diagnostic tests such as electrocardiogram, chest x-ray and standard blood tests"

In a simple replacement comparison for managing patients with suspected PE, physicians were provided with generic diagnostic reports in order to conceal test identity(T91). Again, this particular clinical setting is conducive to masking an element of the diagnostic process since both competing tests, CT pulmonary angiography and V/Q scanning, must be performed by nuclear radiologists; hence treating physicians could simply be provided with the interpretation ("positive for pulmonary embolism", "non-diagnostic study" or "no evidence of pulmonary embolism") whilst safely remaining ignorant of the test's identity.

Moving one step further away from diagnostic decision-making, treating physicians in the aforementioned BIOPED trial were also masked to the triage stage throughout the duration of the study(T74). Interpretation of BIOPED results was passed to third party investigators, who used them to decide which patients should go on to receive a V/Q scan. In order to maintain blinding, a fake negative nuclear medicine report was subsequently sent to the physicians of patients who had received a sham V/Q scan.

Although reported as 'double-blind', a trial evaluating whether NT-proBNP gives incremental value to managing patients with suspected acute heart failure(T100) cannot have blinded its treating physicians, since NT-proBNP results were randomly disclosed in order to assess differences in clinical decision-making. The authors were instead referring to non-treating cardiologists who produced independent final diagnoses without knowledge of the add-on NT-proBNP test results, but did not contribute to management decisions.

Outcome assessors

Primary outcome assessors were blinded in one-fifth of trials (22/103, 21%), considerably more often than blinding of patients or care–providers (Table 6.6).

Measurements were blinded by using independent expert panels(T6,T14–15,T38,T63,T91,T96,T99,T101), clinicians not involved in care provision(T32,T55,T60,T69,T74,T92), or research assistants(T18,T33,T42,T57,T66–T67,T72).

Outcome category	**Number of trials (N=103)**	**(%)**
1. Patient–reported	20	(19)
2. Patient – Outcome Assessor contact	14	(14)
3. Outcome Assessor, no contact	2	(2)
4. Patient – Care-provider contact	25	(24)
5. Medical form data	39	(38)
Unclear	5	(5)
No outcome defined	6	(6)
Total	103	

Table 6.8: Frequency of the five types of outcome measurement in included trials.

Ascertaining the identity of the outcome assessor, in order to subsequently determine whether or not they had been blinded to diagnostic interventions, was somewhat difficult. This was due to the inadequate reporting of primary outcomes, which generally provided insufficient details to establish how, and importantly by whom, they had been measured. This information was frequently absent from reports as authors focused on listing the measurements taken and legitimising their selection of endpoints. Establishing precisely who measured the relevant outcomes was therefore often deduced implicitly from the outcome type and descriptions of measurement methods. Ultimately, the outcome assessor's identity could not be discerned in 11 studies (11%), including six that did not define a primary outcome and a further five that did not provide enough information on how the outcome was measured(T9–10,T56,T94,T108).

Most trials measured outcomes requiring contact with the patient (57% (59/103), of which a third were reported by the patients themselves (20/59, Boutron category #1), almost half were driven by patient contact with treating physicians (25/59, 42%; Boutron category #4), and relatively few requiring contact with non-treating clinicians (14, 14%; category #2) (Table 6.8). The most common method of measurement was the collection of data from medical records (43, 29%; category #5). Very few examples of complementary

Boutron outcome criteria	Example	Source
Patient-Reported (1)	Dyspeptic symptoms recorded daily in a personal calendar by the patient	T46
	Coping responses to cancer recorded using the Mental Adjustment to Cancer Scale	T24
Patient – Outcome Assessor contact (2)	Maximal exercise endurance on the treadmill	T41
	Pregnancy rate, defined as positive urine or serum test in association with presence of an intrauterine gestation sac on ultrasound	T73
Outcome Assessor, no contact (3)	X-ray appearance of the hip	T15
	Diagnostic yield of PCR, defined as the % of embryos with a diagnosis	T93
Patient – Care-provider contact (4)	VTE rate, patients with suspicious signs/symptoms on follow-up given venography or compression ultrasonography of the proximal deep veins for confirmation of diagnosis	T63
	Time-to-first-recurrence, bladder cancer detected during follow-up with cystoscopy and cytology (confirmed histologically)	T90
Medical form data (5)	Embryo implantation rate (number of foetal sacs per embryo transferred)	T87
	Futile thoracotomy rate, defines as futile if an intended curative thoracotomy ended as explorative surgery without tumour resection, or a resected patient died from lung cancer or had recurrent disease during follow up	T17

Table 6.9: Examples of primary outcomes extracted for each outcome type
PCR – polymerase chain reaction
VTE – venous thromboembolism

investigations (3, 2%; category #3) were encountered, namely tests conducted by non-treating care–providers involving no patient contact. Table 6.9 presents examples for each category.

Outcomes requiring contact with patients were less frequently blinded than other categories (Table 6.10). Patient–reported outcomes were never blinded. Almost half the outcomes (13/31, 42%) measured by independent assessors but involving contact with patients were reportedly blind, however upon closer examination most (12/13) of these assessments could not have been blind and so constituted misreporting of trial methods. These outcomes were measured in four trials(T18,T66,T69,T92), of which three(T18,T66,T92) had blinded the outcome assessor but not the patients, leading one to question how successful such attempts at removing ascertainment bias may have been. For example, one trial(T18)

Assessor category	Blind	(%)	Not Blind	(%)	Unclear	(%)	Total	(%)
1.Patient-reported	0	(0)	13	(31)	29	(69)	42	**(28)**
2.Patient-3rd party assessor contact	13	(42)	0	(0)	18	(58)	31	**(21)**
3.Third party assessor, no contact	1	(33)	0	(0)	2	(67)	3	**(2)**
4.Patient-physician contact	9	(35)	2	(8)	15	(58)	26	**(17)**
5.Medical form data	11	(26)	7	(16)	25	(58)	43	**(29)**
Unclear	0	(0)	1	(20)	4	(80)	5	**(3)**
Total	**34**	**(23)**	**23**	**(15)**	**93**	**(62)**	**150**	**(100)**

Table 6.10: Frequency of attempts to blind primary outcome assessments, as reported by trialists. Denominator is the number of primary outcomes (n=150).

measured eight aspects of health–related quality of life, reflecting each of the eight domains from the Short–Form 36. These outcome data were obtained through telephone interview with unmasked patients, by research assistants who were blind to intervention assignments. A single trial(T69) conducted a fully blinded patient–assessor outcome measurement, conducting a follow–up ultrasound in masked patients using masked, independent assessors.

A third (9/26, 35%) of outcomes measured by treating clinicians were reported as blind, although again one of these trials identified the wrong outcome assessor as being masked. Examining the utility of adding a point–of–care ultrasound for investigating trauma patients in the ED, time–to–operative–care was recorded by unmasked treating physicians, and abstracted from medical notes by blinded researchers(T72). Since the outcome value could have also been influenced by the physician, it cannot be claimed to be free of bias.

The remainder (8/9) claimed to have achieved blinding by using masked independent experts to adjudicate outcomes initially measured by non–blinded clinicians. For example,

Number of trials blinding objective vs. subjective outcomes	**Blind**	**(%)**	**Not Blind/ Unclear**	**(%)**	Total
Objective	16	(27)	44	(73)	60
Subjective	7	(18)	32	(82)	39
Unclear	0	(0)	7	(100)	7
Total	*22	(21)	†81	(79)	103

Table 6.11: Frequency of blinding in trials assessing objective versus subjective primary outcomes.
* 1 trial measured a mixture of objective and subjective outcomes(T18)
† 2 trials measured a mixture of objective and subjective outcomes(T24,T53)

in three trials the frequency of venous thromboembolism in patients managed for suspected pulmonary embolism was initially established by treating physicians according to the results of various follow–up investigations (compression ultrasound or venography or pulmonary angiography or spiral CT or V/Q scan)(T63,T74,T91). These findings were subsequently evaluated by an independent committee, blind to initial assignment, in order to confirm the occurrence of a true outcome event. The nuclear medics and treating clinicians who performed these follow–up tests were not blind, however, and since interpretation of the resulting images involve an element of subjectivity it is theoretically possible that ascertainment bias may not have been adequately eliminated in these studies.

The majority of trials (60, 58%) used objective outcomes as their primary measure (Table 6.11), including 'hard outcomes' like all–cause mortality and healthcare cost but also measures of health response assessed using standardised methods of observation, for example assessing patients' maximal endurance to exercise on the treadmill by calculating their mean number of metabolic equivalents using a predefined protocol(T20,T41,T92). Although only 27% (16/60) of these trials performed blinded evaluations, the risk of introducing ascertainment bias should have been relatively low since these measurements are less prone to the influence of opinion. Subjective outcomes are prone to these influences, yet were less frequently blinded (7/39, 18%). Moreover, as discussed above, blinding is unlikely to have been successful in at least 2 of these trials due to the failure to mask both patients and interviewers(T18,T66). These findings raise the distinct possibility that at least a third of the entire cohort of trials (32/103) are at risk of having produced biased primary results.

6.4.2 Rationale for not blinding

The reasons trialists provided to explain why blinding was not carried out are now reported.

Patients

Of the studies that did not blind, only 23% (7/30) provided specific reasons stating that it would not be ethical to do so(T5), that it would not be 'pragmatic' (T12,T19,T107) or that the characteristics of the comparative tests prevented either sham diagnostic procedures or the administration of both tests to all patients(T6,T48,T70). Nine trials(T11,T18,T24,T47,T60–61,T66,T83,T92) stated simply that it was 'not feasible' to blind patients, while the remaining 15(T10,T21,T27–28,T49,T51,T53,T55–56,T68,T71–72,T91,T97,T106) provided no discussion and described the study as an open-label or unblinded RCT.

Care–providers

Similarly, only 27% (9/33) of trials that explicitly did not blind care–providers provided specific reasons for doing so, asserting that blinding would prevent an assessment of the test's impact on clinical decision-making(T92,T107), that is would not be 'pragmatic' (T12,T19), that the nature of the comparative tests prevented clinicians being blinded to their results(T6,T24,T54,T70) or that it would have been unethical to blind those planning treatment (T58). Six studies alluded to the impossibility of blinding(T11,T18,T47,T57,T60–61), while 18 provided no discussion beyond a simple statement of absence(T5,T10,T21,T23,T27–28,T30,T50,T53,T55,T62–63,T66,T71–T72,T83,T97,T106).

Outcome assessors

Comparatively fewer trials stated explicitly that they did not blind outcome assessors (14/103, 14%) than either patients (30/103, 29%) or care–providers (33/103, 32%). Two trials appealed to the impossibility of masking, as the measures required to capture health impact were of necessity subjective, patient-reported appraisals conducted in a setting where it was either 'not feasible' (T11) or 'not pragmatic' (T107) to mask patients. The remaining 12 trials (86%) failed to provide any reasoning for not blinding assessors, although four sought to legitimise their methods by emphasising that open assessment would not have influenced outcome results(T10,T21,T26,T53). Yet two of these trials(T10,T53) measured subjective endpoints which may have been influenced by the assessor's knowledge of group allocation. For example, seeking to evaluate whether initial investigation in a specialist syncope unit improved the diagnostic yield of patients with syncope of undetermined cause, Shen et al compared the number of patients receiving a diagnosis in each arm(T53). Outcome assessors in this case are likely to have been treating physicians who are inextricably implicated in the success of their respective interventions. Consequently, knowledge of the patient's allocation (in this case unavoidable due to the tests being compared) may have influenced the decision to establish a diagnosis. Measurement details were not reported, however a more objective assessment could have been made by specifying the full diagnostic criteria required in the trial protocol.

6.5 Results: Can trials control for performance & ascertainment bias?

Since it was theorised that blinding may be impossible to achieve, a subjective assessment of whether blinding could have been performed was carried out for all trials, regardless of whether they reportedly blinded or not (summarised in Appendix E.1, p.410).

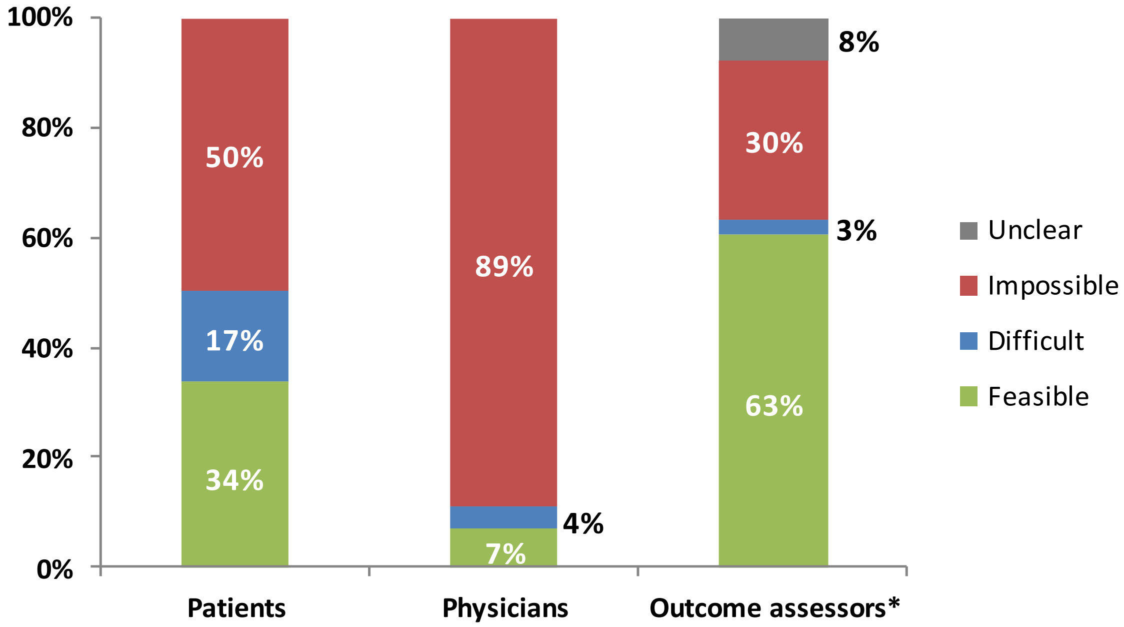


Figure 6.2: The feasibility of blinding patients, care-providers and outcome assessors in test-treatment RCTs

* Trials with multiple primary outcomes were entered once if all outcomes fell into the same categories. For four trials outcomes fell into 2 feasibility categories; see table 6.14

6.5.1 Feasibility of blinding

The subjective feasibility assessment suggests it was not always possible to perform fully blinded studies. Taking into account all elements of the clinical setting for each study, the author judged that it would rarely have been possible to blind care–providers (11/103, 11%), while 50% (51/103) of trials could accommodate patient blinding. Masking outcome assessment was most often deemed feasible, though even then only 66% (68/103) of trials could have achieved this (Figure 6.2). The methods required to blind successfully were considered to be difficult to implement when considerable additional resources would be required, such as employing additional teams of clinical staff to perform tests for the purposes of blinding clinicians(T3,T14,T62), attempting sham procedures to blind patients to invasive tests(T70,T99,T105) or to multiple phases of testing(T65), or when measuring patient outcomes in trials that were considered difficult to also blind patients(T24,T65,T68). Methods were judged to be difficult to blind clinicians in a third (4/11, 36%) of these studies and to blind patients in 33% (17/52) of trials.

Masking outcome assessors was judged to be more easily attainable in most comparisons (65/68, 96%) however this varied according to the types of outcome measured. Objective assessments were easier to blind (90%, 70/78) than subjective ones (25%, 18/72) (Table 6.12), while patient-reported responses were far more difficult to blind (category #1: 10%, 4/42) than routine data collections (category #5: 86%, 37/43) (Table 6.13). Though few in number, outcomes assessed by independent clinicians or researchers (category #3) were more often amenable to blinding, for example assessing the recovery of shoulder mobility after invasive surgery for the diagnosis of breast cancer spread(T24) or exercise endurance after treatment for coronary stenoses(T20). Assessments made during treatment could be masked if independent adjudicators were used, for example to measure the rate of venous thrombosis in symptomatic patients previously managed for suspected pulmonary embolism presenting during follow-up(T3), which could have been achieved for 73% (19/26) of these outcomes. Contrary to expectations, outcomes assessed from data on medical forms could not always be made in blinded fashion since the tests received by patients could be revealed, for example when collating resource-use for the calculation of diagnostic and/or treatment costs(T5,T12,T19,T55,T70,T83). This pitfall was avoided in trials using a randomised disclosure design since differences in cost could be calculated for the period after randomisation, and hence after testing had taken place(T100).

Overall, trialists would have found it impossible to blind patients as well as care–providers and outcome assessors in at least a fifth of all trials (23, 22%)(T4,T11,T12,T13,T18–19,T24,T27,T30,T34–37,T39–40,T44,T46,T51,T56,T64,T77,T88,T107). This rate excludes five studies in which blinding appeared impossible for patients and care–providers, though the feasibility of blinding outcome

Feasibility of blinding outcome assessor	**Objective outcome**		**Subjective outcome**		**Total**	
	N	**%**	**N**	**%**	**N**	**%**
Feasible	61	(78)	14	(19)	75	(50)
Difficult	9	(12)	4	(6)	13	(9)
Impossible	8	(10)	52	(72)	60	(40)
Unclear	0	(0)	2	(3)	2	(1)
Total	**78**	**(100)**	**72**	**(100)**	**150**	**(100)**

Table 6.12: The feasibility of blinding according to subjectivity of outcome measurement. Individual outcomes used as denominator.

Blinding Feasibility	**Outcome category**													
	1	**(%)**	**2**	**(%)**	**3**	**(%)**	**4**	**(%)**	**5**	**(%)**	**Unclear**	**(%)**	**Total**	**(%)**
Feasible	2	(5)	12	(39)	3	(100)	19	(73)	37	(86)	2	(40)	75	(50)
Difficult	2	(5)	10	(32)	0	(0)	1	(4)	0	(0)	0	(0)	13	(9)
Impossible	38	(90)	9	(29)	0	(0)	6	(23)	6	(14)	1	(20)	60	(40)
Unclear	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	2	(40)	2	(1)
Total	**42**		**22**		**3**		**26**		**43**		**5**		**150**	

Table 6.13: The feasibility of blinding individual outcomes in test-treatment RCTs. Note: since feasibility is contingent on the type of measurement taken, the total number of outcomes assessed is used as a denominator.

assessors could not be determined due to poor reporting of the methods of outcome measurement(T108) or primary outcome definition(T2,T29,T79,T82).

Frequency of blinding when judged feasible

When these frequencies are examined according to trialists' reported attempts to blind (Table 6.14), it is apparent that test-treatment trials did not always blind when they could have done so. Almost half (47/103, 46%) could have improved their designs by blinding patients, including 29% (30/103) that could have done so easily since the nature of the test–treat comparison would have enabled all tests to be given ethically to all patients, as real or sham procedures. For example, trialists assessing the value of an additional MRI in patients with suspected fracture of the scaphoid could have given control patients a sham MRI(T10). Another 17 trials could have blinded patients, though with some degree of

(A) Patients	**Blind**	**(%)**	**Not Blind**	**(%)**	**Total**	**(%)**
Feasible	5	(100)	30	(31)	35	(34)
Difficult	0	(0)	17	(17)	17	(17)
Impossible	0	(0)	51	(52)	51	(50)
Total	*5*	*(100)*	*98*	*(100)*	*103*	*(100)*
(B) Care-providers						
Feasible	2	(50)	5	(5)	7	(7)
Difficult	1	(25)	3	(3)	4	(4)
Impossible	1	(25)	91	(92)	92	(89)
Total	*4*	*(100)*	*99*	*(100)*	*103*	*(100)*
(C) Outcome assessors						
Feasible	22	(100)	43	(53)	65	(63)
Difficult	0	(0)	3	(4)	3	(3)
Impossible	2	(9)	29	(36)	31	(30)
Unclear	0	(0)	8	(10)	8	(8)
Total	**22*	*(100)*	*†81*	*(100)*	*103*	*(100)*

Table 6.14: Feasibility of blinding patients (A), care–providers (B), and outcome assessors (C) tabulated against attempts at blinding.

Trials with multiple primary outcomes were entered once if all outcomes fell into the same categories. For four trials outcomes fell into 2 different categories:

* Outcomes in two categories for each of two trials (T18,T55)

† Outcomes in two categories for each of two trials (T24, T30)

difficulty, for example by giving sham CT or sham plain X–ray in addition to the allocated test to patients under investigation of acute abdominal pain within one–hour of presentation(T85).

Similarly, around half the trials examined (46, 45%) could have blinded outcome assessors but failed to do so, almost all of which (43/46, 93%) could have achieved this by simple methods, such as blinding patients(T41,T68,T71) and/or employing an independent assessor to take follow–up measurements(T58,T75,T90).

Though rarely possible, approximately one in twelve trials (8/103, 8%) could have blinded care–providers but failed to do so. This includes five trials that could have done so with ease, for example by providing clinicians with standardised diagnostic reports when comparing exercise ECG with stress ECG for the identification of patients with coronary artery disease requiring immediate treatment(T89). Another three trials (3%) may have achieved blinding through more convoluted means, for example by employing an independent team of physicians to take aspirate samples from patients when comparing two techniques of aspirate sampling and culture processing for identifying the causative organism in patients with ventilator–associated pneumonia(T62). In order to complete the masking procedure, care–providers could subsequently be given test results in standardised format.

6.5.2 Mistaken reports of blinding

Three discrepancies were noted between the subjective assessment of feasibility and trialists' reports of blinding. In two trials blind outcome assessments could not have been successful due to the impossibility of masking the patient in addition to the interviewer during assessment of health-related quality of life(T18), or the impossibility of masking assessors to the identity of tests used when collating resource use and cost data from the medical record(T55). The third trial compared two methods for selecting healthy embryos for in–vitro fertility treatment. While the gynaecological care–provider was blind, in this case the individual making the treatment decisions (i.e. responsible for selecting the embryos) was the laboratory clinician carrying out the tests, who therefore could not have been masked(T98).

6.6 Results: Do trials adequately control for attrition bias?

The validity of trial findings is threatened if study groups are no longer similar at the point of analysis, and this was posited to be a particular threat in test-treatment RCTs if the number of interventions given result in increased drop–out rates. The fourth aim of this review therefore assessed whether main analyses adequately limited the influences of attrition bias (summary data provided in Appendix E.2, p.418).

6.6.1 Appropriateness of the main analysis

Subgroup comparisons

The first criterion for an unbiased primary analysis is that the outcome must be measured in all randomised participants. Thirteen studies (13%) used a primary outcome that could only be measured in a subgroup of participants. Nine trials (9% of all trials) used the subgroup as the denominator for calculating effect differences (Table 6.15), and so are at risk of having produced distorted measures of effect by comparing two groups that are not analogous in patient characteristics. In order to evaluate the impact of the whole test–treatment strategy, the proper approach requires observed events to be calculated using all randomised participants as the denominator, as performed by the remaining 4 trials(T17,T61,T78,T102).

When event rates were recalculated using the full denominator, the results of one trial changed enough to require a different interpretation of impact. The trial evaluated whether subfertile women attending for intrauterine insemination should all be given diagnostic laparoscopy as standard prior to treatment, or instead only be investigated for biological causes of infertility (e.g. tubal pathology or endometriosis) after failure of initial intrauterine insemination(T27). Hypothesising that laparoscopic abnormalities would be more frequent amongst women failing to get pregnant, the authors measured the proportion of participants with abnormal laparoscopic findings requiring treatment and/or leading to a change in fertility treatment. Using the proportion of participants undergoing investigative laparoscopy as the denominator, a non–significant increase was reported (experimental 13/23, control 31/64; OR=1.4 [95%CI: 0.5–3.6]); however when the full study population is used a significant *decrease* in the proportion of women receiving a change in treatment as

Trial	Primary outcome	Subgroup	Proportion of patients excluded from primary outcome*		
			Total	Expl	Comp
T32†	Appropriateness of initial ED triage decision (Inappropriate discharge home of pts with ACI)	Patients with a final diagnosis of ACI ***(false-negative rate)***	87%	86%	87%
T26	Rate of in–hospital cardiac catheterisation	Patients who were admitted ***(treatment subgroup)***	86%	82%	90%
T10	Days unnecessarily immobilised	Patients with a final diagnosis of no fracture ***(false-positive rate)***	72%	64%	76%
T72	Mean time from arrival in ED to direct transfer to operative care	Patients receiving operative treatment ***(treatment subgroup)***	71%	74%	68%
T27	Rate of abnormal laparoscopies leading to a change in Rx	Patients receiving laparoscopy ***(treatment subgroup)***	44%	70%	17%
T60	Negative appendectomy rate	Patients who had their appendix removed ***(false-positive rate)***	46%	46%	46%
T7	Proportion prescribed antibiotics for non-pneumonic acute U/LRTI	Patients with a diagnosis of non-pneumonic U/LRTI ***(treatment subgroup)***	20%	28%	13%
T91	Frequency of VTE in patients in whom PE was excluded	Patients with a negative test result for PE ***(false-negative rate)***	17%	19%	14%
T74	Recurrence rate of VTE (in pts not taking anticoagulants)	Patients with negative test result for VTE ***(false negative rate)***	16%	17%	15%
T32†	Appropriateness of initial ED triage decision (inappropriate hospitalisation of pts without ACI)	Patients with a final diagnosis of no ACI ***(false-positive rate)***	13%	14%	13%

Table 6.15: Test-treatment RCTs using subgroup measurements for primary analysis.

* calculated as a proportion of the numbers analysed, if different from the numbers randomised. This is done in order to differentiate from other reasons for exclusion from analysis.

† This trial had two subgroup primary outcomes, which if measured together would have constituted an appropriate full analysis.

a result of undergoing testing after failure of treatment is observed (experimental 13/77, control 31/77; OR=0.3 [95%CI: 0.14–0.64]).

A third trial used two subgroup outcomes to capture differences in the primary study aim: the appropriateness of decision–making(T32). Evaluating the addition of single-photon emission CT to standard care for the management of suspected acute cardiac ischaemia in an emergency setting, patients with a study diagnosis of ischaemia were admitted, whilst those in whom ischaemia was ruled out could be discharged. The appropriateness of these decisions were measured by independent reassessment of all participants; discharged patients were considered incorrectly managed if independently diagnosed as having ischaemia, while hospitalised patients were inappropriately managed if found to be free of the target condition on independent examination testing. Although these two outcomes are complementary and so could have been combined to be analysed as a full–group outcome, they were kept separate and so open the results to selection bias. Whether the results really are biased could be examined by checking for baseline imbalances within each subgroup, though unfortunately the authors did not do so.

Consistency of outcome measurement

The large majority of trials used appropriate methods (87/103, 84%), either clearly using the same method across all study arms (67, 65%) or assessing test performance outcomes for which use of different tests was appropriate (20, 19%)(Table 6.16). Nine additional trials provided unclear descriptions, however were considered at low risk of bias since the nature of the outcome being measured suggests that ascertainment methods are likely to be the same(T4,T9,T33,T50,T56,T62,T98,T100,T104,T108). Two, for example, assessed the ‘clinical pregnancy rate’ in women managed for infertility(T98,T104); though methods of assessment were not reported, women were followed–up at the same time point within each trial and the tests under comparison (biochemical analysis of embryos(T98) and hysteroscopy(T104)) could not have been used.

Consistency	No. of trials	(%)
Consistent	67	(65)
Inconsistent	2	(2)
Unclear – low risk of difference	9	(9)
Unclear – high risk of difference	2	(2)
Not relevant – test performance measure	20	(19)
Primary outcome not defined	6	(6)
Total	***103**	

Table 6.16: Consistency of outcome measurement across trial arms.
* 4 trials with multiple outcomes fell into more than one category: Unclear–low risk (Health perception) and Yes (Function, Psychological morbidity, Quality of life, Satisfaction, Symptom rate)(T4); Yes (Recurrence rate) and No (Recurrence rate)(T30); Yes (Clinical status) and Test performance (Diagnostic yield)(T93); Unclear–low risk (length of stay) and Yes (Cost)(T100).

Two trials used inappropriate methods likely to have lead to bias(T10,T32), and two others provided unclear descriptions judged to be at high risk of bias if measurements were not identical(T16,T91). One trial compared the number of days patients with suspected scaphoid fracture were unnecessarily immobilised (also a subgroup comparison, see above) following investigation by ‘standard care’ or with an additional wrist–MRI(T10). All patients were initially immobilised, and those found to have a fracture on the basis of these tests subsequently had the plaster removed. Attempting to compare the inappropriate treatment rate, defined as fracture–free patients who were initially immobilised, measurement of this outcome requires knowledge of the true disease rate (fracture). However the disease rate was determined by two different testing strategies, and so the resulting rates are not comparable. In order to achieve true comparability, the same test would have to be used to determine the true disease rate, as was accomplished in the four other trials that also assessed the appropriateness of treatment decisions(T17,T32,T60,T102). A cardiovascular trial comparing rates of venous thromboembolism following management for suspected pulmonary embolism provides an example of unclear reporting at high risk of bias(T91). Patients were randomised to computed tomography pulmonary angiography or V/Q scan

Primary analysis	Reported	(%)	Not reported	(%)	Total	(%)
Complete	27	(35)	14	(56)	41	(40)
Incomplete	51	(65)	8	(32)	59	(57)
Unclear	0	(0)	3	(12)	3	(3)
Total	78	(100)	25	(100)	103	(100)

Table 6.17: Frequency of test-treatment RCTs with complete primary analyses. Incomplete analyses defined as those analysing fewer patients than were randomised

for confirmation of the suspected diagnosis, but at post–treatment follow–up the authors report that venous thromboembolism was ascertained "using either CTPA [computed tomography pulmonary angiography] or V/Q scanning". This statement does not make clear whether use of the two tests differed systematically according to study group; if it did then the rates of disease can be expected to differ as a consequence of different accuracy, which is inappropriate for a downstream outcome.

6.6.2 Completeness of outcome data

Two thirds of trials (59/103) analysed fewer patients than were randomised due to the exclusion of participants after randomisation and/or missing outcome data (Table 6.17). Seventy–six percent of trials (78/103) explicitly reported whether exclusions and missing data had occurred, with 35% (27/78) of these claiming their analyses to be complete. Trials that did not explicitly report these details were more likely to appear complete (14/25, 56%), raising the possibility that real losses may have occurred but were not reported. In three trials poor reporting precluded an assessment of whether the analysis was complete(T49–50,T86).

Exclusion of participants after randomisation

In total, 30 trials (29%) excluded participants after randomisation due to: protocol deviation (11)(T33,T37,T39,T46,T51,T58,T70,T85,T88,T104,T108), withdrawal of patient consent (10)(T6,T11,T27,T30,T68,T70,T74,T77,T88,T92,T95), subsequent re–evaluation of eligibility (7)(T1,T5,T49–

Deviations from randomised allocation	No. of trials	(%)
Protocol deviations excluded	11	(11)
Protocol deviations included	42	(41)
Stated no protocol deviations	18	(17)
Implied no protocol deviations	16	(16)
Unclear	16	(16)
Total	103	(100)

Table 6.18: Procedures for handling participants who did not received the allocated intervention.

50,T52,T72,T86,T88), contraindications to receiving the allocated test (3)(T4,T39,T95), patients randomised twice (1)(T68), treatment refusal (1)(T90), negative or severe test results (1)(T90), the identification of an outlier at analysis (1)(T5) or for no specified reason (2)(T44,T51).

Overall half the trials (53/103, 51%) documented a deviation from protocol that meant some participants did not receive their allocated test and/or treatment intervention, although only a minority (11/103) failed to analyse these participants according to their randomised groups (Table 6.18). A third of trials (32, 31%) did not report whether protocol deviations had occurred, half of which appeared to have no ensuing patient exclusions and the other half for whom it was impossible to establish whether either had occurred due to very poor reporting.

Missing outcome data

Data were missing from primary analyses due to losses during follow–up and/or missing responses in 52 studies (51%), 15 (29%) of which also excluded some participants after randomisation. Reporting was so poor for two studies that it was impossible to determine – even implicitly – whether data were missing or not(T50,T86). Overall one third of trials with missing data (16/52) provided specific reasons for all missing responses in each comparative arm (Table 6.19), and 21% (11/52) provided no descriptions whatsoever. The majority (48%, 25/52) provided partial accounts that were insufficient to determine whether

Reasons for missing outcome data	No. of trials	(%)
Full reasons reported	16	(31)
Partial reasons reported	25	(48)
No reasons provided	11	(21)
Total	52	(100)

Table 6.19: Quality and frequency of reporting the reasons for missing outcome data

reasons for missing data might be associated with the test–treat strategy, studies for example stated simply that patients were 'lost to follow–up' (T3,T39,T49,T56,T89,T91,T101) or that they had suffered 'missing data' (T11,T32,T96).

Procedures for handling missing responses were very poorly reported. Most trials failed to report any method and excluded all missing responses to perform a complete case analysis (69%, 36/52)(Table 6.20). Poor reporting of participant flow by four trials with missing data meant that it was unclear what their approach had been(T7,T41,T49,T82). Only nine trials (17%) imputed all missing values, while three others imputed partial responses but excluded wholly missing records(T4,T57,T99). Imputation methods were reported by 33% of these studies (4/12) as the last observation carried forward(T69), use of an earlier or later outcome response(T4), allocation of a poor outcome(T57) and censoring(T99). Six trials implicitly allocated a poor outcome by including participants with a missing response in the denominator but not the numerator. The method could not be discerned in the remaining two studies(T10,T76). No trials reported using multiple imputation methods.

When the quality of reporting was examined against approaches used to deal with missing data, over half the trials (29/52, 56%) presented inadequate accounts of methods that could potentially create bias if mishandled; namely those that excluded participants, or that were unclear regarding whether exclusions had occurred, and failed to fully report the causes of data loss for each comparative strategy (Table 6.21).

Missing outcome data	No. of trials	(%)
Stated none missing	33	(32)
Appeared to have none missing	15	(15)
Unclear whether data missing	3	(3)
Excluded all missing	36	(35)
Imputed some missing	3	(3)
Imputed all missing	9	(9)
Unclear whether missing data imputed	4	(4)
Total	103	(100)

Table 6.20: Procedures for handling missing outcome data.

Reasons for missing outcome data:	**Missing data:**						**Total no. of trials**	**(%)**
	Included	(%)	*Excluded	(%)	†Unclear	(%)		
Fully reported	2	(22)	14	(36)	0	(0)	16	(31)
Part reported	6	(67)	19	(49)	0	(0)	25	(48)
Not reported	1	(11)	6	(15)	4	(100)	11	(21)
Total	**9**	**(100)**	**39**	**(100)**	**4**	**(100)**	**52**	**(100)**

Table 6.21: Reporting quality cross–tabulated with method used to deal with missing data.

* studies presenting a mixed approach (including some missing data, but excluding other missing data) are considered to have excluded overall.

† denotes that the trial's method to manage missing data (i.e. to include or exclude) was unclear.

6.6.3 Quantification of incomplete analyses & differential loss

In addition to the three poorly–reported trials where the reviewer was unable to ascertain whether analyses were complete, the missing proportion could not be quantified in four trials that report exclusions but failed to report the numbers randomised(T7,T82) and/or analysed for each arm(T41,T64). In total, therefore, 55 trials (53%) excluded between 0.1%(T62,T32) and 46%(T60) of randomised participants from primary analyses (median: 7.0%, IQR: 1.4%-17.6%), while 24% of all included studies (25/103) excluded more than 10% of the original study population (Table 6.22).

Total missing from randomised study population	No. of trials	%
<1%	52	(50)
1–10%	21	(20)
11–20%	12	(12)
21–30%	5	(5)
31–40%	3	(3)
41–50%	3	(3)
Unclear	7	(7)
Total	**103**	**(100)**

Table 6.22: Quantification of the total number of participants excluded from primary analyses for each trial.

Comparative between-arm attrition	Attrition greatest in: Experimental arm	Control arm	Equal	No. of comparisons	(%)
within 5%	24	19	2	45	(68)
5-9.9%	9	2	0	11	(17)
10-19.9%	8	1	0	9	(14)
>20%	1	0	0	1	(2)
Total	**42**	**22**	**2**	**66**	**(100)**

Table 6.23: Degree of differential attrition in trials with some attrition. Note: the denominator used is the number of comparisons.

Participant analysis	ITT	(%)	No ITT	(%)	ITT not reported	(%)	Total	(%)
Analysed as randomised	41	(67)	2	(67)	16	(41)	59	(57)
Not analysed as randomised	5	(8)	1	(33)	5	(13)	11	(11)
Unclear	15	(25)	0	(0)	18	(46)	33	(32)
Total	61	(100)	3	(100)	39	(100)	103	(100)

Table 6.24: Trials analysing all participants as randomised compared to their reports of an intention–to–treat analysis.

Attrition differed by more than 5% between arms in 21 comparisons made by 16 trials (16%)(T4,T20,T27,T34,T37,T39,T40,T46,T51,T65,T66,T88,T90,T92,T92,T94), and in almost all cases experimental interventions lost the most participants (Table 6.23). Attrition differed by more than 20% in only one trial(T27). This latter study excluded over half the randomised population, including 70% of experimental group participants, and 17% of comparator arm participants. The trial sought to evaluate the clinical utility of performing diagnostic laparoscopy routinely in infertile women before intrauterine insemination, compared to standard management that performs laparoscopy to detect tubal abnormalities only after a first phase of insemination has failed. Acknowledging that in excess of 1000 women would be required to demonstrate an impact to the clinically important outcome (pregnancy rate), a process measure of diagnostic yield was selected in which the rate of abnormal laparoscopies leading to a therapeutic intervention was used as the primary measure of effectiveness. However, in the primary analysis trialists excluded all women who failed to receive a laparoscopy, a perplexing approach which is at odds with the nature of the comparator intervention that did not require testing unless treatment had previously failed. Thus all participants who became pregnant during intrauterine insemination treatment were excluded from the effectiveness measurement. It is clear that data were not missing at random from this study, but being 'missing' was in fact directly linked to a prognostic variable: since excluded individuals might also be expected to have fewer tubal abnormalities, the trial's results are very likely biased.

6.6.4 Conduct of an intention–to–treat analysis

An intention-to-treat analysis was reportedly conducted by 61 trials (59%), 3% (n=3) reported that they had not performed one, while the remaining 38% (n=39) did not report the type of analysis conducted. Two–thirds of studies (41/61) describing an intention–to–treat analysis complied with the principle of analysing randomised participants in their assigned groups (Table 6.24). Half of these (21/41) also had no missing data or imputed missing data, and so comply with the most rigorous definition of intention–to–treat analysis[123].

6.7 Results: Do trials adequately control for type II error?

Test-treatment trials are likely to require substantially larger study populations, hence they risk being underpowered to detect patient health effects, while it may not be feasible to recruit the necessary number of patients in adequately powered trials. The final aim therefore assessed the size of study populations, whether power calculations were used to estimate numbers needed, and how often target sample sizes were reached (summary data provided in Appendix E.2, p.418).

6.7.1 Use of power calculations

Eighty-one trials (79%) reported an *a priori* justification of sample size, including all 6 cluster RCTS. One study offered a *post hoc* computation of power(T4), while the remaining 21 trials provided no justification(T1–2,T7,T13,T20,T28–29,T31,T35,T37,T41,T54,T59,T76,T79,T82–83,T87,T90,T104–105). These trials were also less likely to define primary endpoints (16/22, 27%) than trials with adequate reporting of sample size (81/81, 100%), severely compromising the utility of their results.

The outcome parameter used in power calculations was reported by all but two of the 81 trials (79/81, 98%)(T18,T89). Twelve trials (15%) did not report any primary study endpoints other than those used in the power calculation, precluding further appraisal. Fifty-four studies (67%) correctly used the primary study outcome as a basis for sample size estimates, while six (7%) powered on a single variable when the study evaluated multiple primary outcomes(T26,T39,T55,T93,T100,T107). The power variable did not match the primary outcome in one trial(T70), although the trialists had consciously done so in order to increase

Achieved sample size	2-arm trials	(IQR)	>2-arm trials	(IQR)	Cluster	(IQR)	All trials	(IQR)
Median per trial	301 (n=87)	148-615	383 (n=10)	219-851	577 (n=6)	364-760	309 (n=103)	153-731
Median per study arm per trial	152 (n=83)	72-314	124 (n=8)	65-220	227 (n=5)	148-250	166 (n=96)	72-297

Table 6.25: median achieved sample sizes of test-treatment RCTs. 25th and 75th quartiles used.

the study's power to detect differences in secondary patient outcomes whilst keeping resource use (cost) as the main study aim.

6.7.2 Attainment of target sample size

Overall, achieved sample sizes ranged from 20(T87) to 5341(T67) participants in individually–randomised trials (median: 305, IQR: 152–740), and 145(T66) to 972(T48) participants in cluster RCTs (median: 577, IQR: 364–760). On average (median) 166 participants were recruited to each intervention arm across all trials (n=96, IQR: 72–297)(Table 6.25). Trials reporting power calculations had considerably larger study samples (median: 408, IQR:157–782) than those omitting such description (median: 212, IQR: 108–304), suggesting that reporting may be a good surrogate for rigorous *a priori* methodological planning. Trials using patient primary outcomes had slightly larger median study samples (median: 348, IQR: 163–772) compared to those using process outcomes (median 247, IQR: 138–500).

Of the 79 trials in which a comparison between target and achieved sample sizes could be made (two did not provide power calculation results(T26,T32)), 41 (52%) achieved 95–105% of the estimated target, 25 (32%) achieved >105% and 13 (16%) failed to reach within 5% of their target (Figure 6.3). Seven studies exceeded their targets by more than 20%(T5,T10,T12,T46,T75,T77,T94), and three of these by more than 100%(T5,T10,T77), although in no report were the reasons for such over–recruitment documented or rationalised.

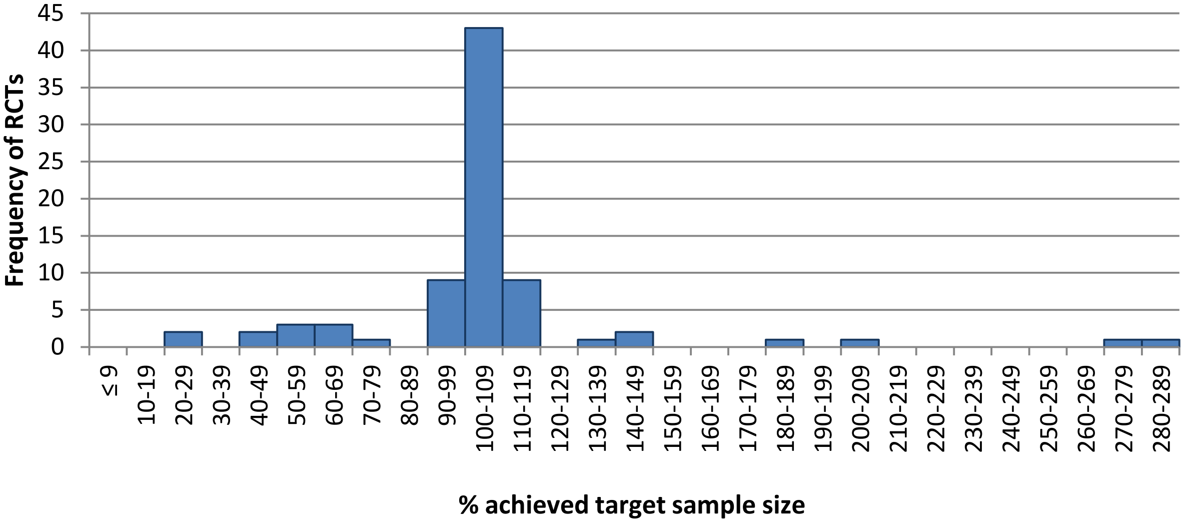


Figure 6.3: Distribution of achieved sample sizes as a proportion of the target sample size in RCTs providing power calculations and sample size estimates (n=79)

Conversely, 11 trials failed to reach 75% of their estimated targets(T17,T38,T40,T42,T53,T63,T67,T73,T74,T103,T108), including four that recruited less than 50% of the required participants(T38,T42,T63,T73). Of the trials with a ≥5% deficit, four (31%) were stopped early as routine practice had evolved during the study period to incorporate the experimental strategy(T40,T42), or because interim examination exposed a very low outcome rate(T63) or a very high outcome rate(T67). None of these provided details of any stopping rules. Another five (38%) reported difficulties in recruiting the desired numbers(T30,T38,T53,T74), while three provided no explanation for their failure to recruit the intended numbers(T15,T103,T108).

In total five studies reported modifying their initial power calculations, four revising down due to a higher than hypothesised disease prevalence in the study population(T91), a lower than expected primary outcome rate(T103), updated observational data(T15), or due to difficulties in recruiting(T36), while the fifth produced a higher estimate on the basis that disease prevalence was lower than expected(T3).

6.8 Discussion

This chapter has assessed the adequacy of methods and reporting quality of a cohort of 103 test-treatment RCTs. Its overall aim was to establish the extent to which these complex intervention trials are susceptible to the particular methodological challenges that have been hypothesised to affect them. To answer this charge, the review's key findings are now discussed by contrasting and comparing them to those observed for the quality of standard treatment RCTs and complex interventions trials.

Trial conduct was poorly reported

As found in chapters 4 and 5 test-treatment RCTs were generally poorly reported, this time with regard to the methods used to ensure internal validity. At times these accounts inaccurately represented actual conduct, reflecting the experiences of previous researchers reviewing the quality of standard treatment RCTs[118,152]. Fewer than half of all trials provided the methods used to conceal allocation or whether blinding had been conducted. A quarter failed to include a power calculation, or to provide the information necessary to establish the completeness of the primary analysis, while slightly fewer reported methods used to randomise participants or whether an intention–to–treat analysis had been carried out.

It is clear that this level of reporting can neither sustain a proper appraisal of methodological quality, nor can it support an independent verification of the effectiveness of test–treatment strategies through the interpretation of a trial's results. Such sub–optimal reporting raises the concern of overestimated treatment effects as well as potentially spurious findings, which have been associated with poor reporting[97,106–108].

6.8.1 Trials often fail to protect against selection bias

Sequence generation and allocation concealment are not influenced by the clinical setting, and so should have been performed adequately in all included RCTs. This was not the case; approximately 40% provided inadequate or unclear methods of randomisation, while 2 in 3 trials presented inadequate or unclear concealment of allocation. This finding, particularly the low rate of adequate concealment, raises the concern that many trials have failed to protect adequately against selection bias. Since an open allocation process allows clinicians the opportunity to select which diagnostic strategy to use in each participant, this approach risks falsely underestimating test–treat effects if the experimental technology is preferred for more challenging cases who are destined for poorer health outcomes. Of course, the direction of bias will vary according to the nature of clinicians' preconceptions regarding the efficacy of competing tests.

These rates are generally better than those found in cohorts of general RCTs. In the largest study, a meta–analysis of seven meta–epidemiological reviews found 23% of trials to have adequate allocation concealment[108], while in two more recent populations rates of 56%[117] and 28%[104] were reported. Adequate methods of randomising allocation also appear less frequently in other populations, ranging between 24%[104] and 43%[100] in individual meta–analyses but reported as only 25% in the large meta–epidemiological study[108]. Test-treatment RCTs also performed better than trials of surgical interventions, which randomised adequately in 41% and concealed adequately in 25%[239], and non–pharmaceutical RCTs, performing adequately in 54% and 18% respectively[240]. Seen within the context of common suboptimal performance across clinical disciplines and intervention types, these inadequacies in evaluation are not related to test-treatment interventions.

6.8.2 Trials very rarely control for performance bias and ascertainment bias

Blinding is rare in test-treatment RCTs

Using reporting as a proxy for methods, blinding was conducted by very few test-treatment trials. Patients were masked in 5% and care–providers in 4% of included trials. Assuming that the lack of blinding exerts the same influence on test–treatment effects as general cohorts of trials, these findings indicate that approximately 95% of test-treatment trials risk producing results that are distorted by performance bias. Blinded outcome assessments were considerably more frequent, though still only implemented by one in five test-treatment trials. What is more, subjective outcomes were less frequently masked than objective outcomes; since existing meta–epidemiological evidence strongly suggests that subjective treatment effects are the most strongly distorted in open trials, one can infer that at least 30% of the test-treatment trials examined in this review (32/103) are at high risk of ascertainment bias.

This picture is very similar to that produced for other complex intervention RCTs. A review of 158 surgical intervention RCTs found masking of patients in 8%, care–providers in 0, and outcome assessors in 17% of trials[239], highly comparable to the rates observed in the current review. Boutron and colleagues directly compared frequencies of blinding in RCTs evaluating pharmaceutical interventions and non–pharmaceutical interventions for the treatment of hip or knee osteoarthritis[243]. They also found that blinding was significantly less frequent in non–pharmaceutical trials with patients, care–providers and outcome assessors blinded in 24%, 6% and 36% of these studies compared to 97%, 82% and 98% of drug trials[243].

Comparison to reviews of general cohorts of trials is more difficult. Most reviews appraise the frequency of ‘double–blinding’, and so rarely provide insights into the frequencies with which each of the three separate constituent partakers of trials are masked. Moreover, definitions of ‘double–blinding’ vary widely both with regard to the meaning ascribed by reviewers[108], and to that originally intended by trialists. Chan and colleagues[152] for example encountered nine variations in the recipients of blinding by trials describing themselves as ‘double–blind’. A recent study that combined seven meta–epidemiological reviews of methodological quality reported double–blinding to have been conducted by 56% (590/1057) of their cohort of RCTs, which consisted mostly of pharmacological trials[108]. Taking the most common definition of double–blinding used, that ≥2 of either patients, care–providers or outcome assessors should be blind, this frequency is far higher than that observed in the current review, where only 5% of test-treatment RCTs (5/103) blinded at least two categories of individual. In fact, only 24% (25/103) could be described as ‘single–blind’, namely blinding either patients or care–providers or outcome assessors. Within the context of existing research, this review confirms that blinding is far rarer in test-treatment RCTs than single intervention treatment trials.

6.8.3 Blinding is not always feasible in test-treatment trials

Scarcity of attempts to blind are almost certainly a reflection of the practical and ethical difficulties involved in performing sham diagnostic procedures, or indeed in masking real test results from patients and their clinicians. This is supported by the results of the feasibility analysis, which confirms existing suspicions that blinding is frequently impossible in test-treatment RCTs. Only half the trial settings were amenable to blinding participants, around one in ten could accommodate care–provider blinding, and two–thirds could have masked the primary outcome assessment. These proportions broadly equate with the findings of a similar review, which concluded that blinding in RCTs of complex therapies for osteoarthritis was possible for patients in 52% of trials, care–providers in 14% and outcome assessors in 50%[243]. This similarity stems from the nature of interventions examined by the two reviews: complex therapies for osteoarthritis, as well as test–treat strategies, are commonly performed by care–providers, can be invasive, and are typified by active patient participation. As found by Boutron[243], these features create difficulties in ensuring that the interventions being compared are indistinguishable to both patients and their physicians.

Barriers to blinding patients and clinicians

Feasibility was found to hinge on an interplay between four features of the study setting, characterised in chapter 4 (p. 92–106): the types of tests being compared, circumstances surrounding administration of tests, the type of test comparison and the study question. Though it is tempting to categorise feasibility according to these features, the complexity of these trials was such that it is not possible to derive simple ‘rules’ of feasibility; rather the author found that it was more commonly the balance of differences and similarities between study settings that determined when blinding might be possible. Table 6.26 summarises study characteristics that appeared more or less conducive to blinding.

In order for patients to be masked successfully, they must remain unaware of which test is being used to direct their treatment decisions until follow–up is complete. To achieve this participants must either receive all testing strategies, or receive a convincing ‘sham’ for the test that they were not allocated to. As illustrated in table 6.26, similarities in the physical characteristics of comparative tests were key to blinding patients, in particular the risks of procedural morbidity and mode of administration.

Patients could most easily be masked when comparative tests were not invasive, thus allowing them to be administered safely as real or sham procedures to all patients, as occurred in all four trials that did blind. On the other hand, invasive tests could accommodate patient blinding in certain settings, for example if the competing techniques involved the same method of administration, were conducted in the absence of the patient (such as biochemical testing of samples) or under general anaesthetic. The comparison of white-light with blue-light cystoscopy for the visualisation of bladder tumours serves as a

	Patient	Clinician
Feasible	**Comparative tests have no/minor associated direct harms and the same physical characteristics, or route of administration that could easily accommodate shamming or can ethically be given to all study participants e.g.** *white light cystoscopy vs. white light + fluorescence cystoscopy conducted during same procedure for detection of bladder cancer (T90)*	**Simple comparisons where all tests are interpreted by non-treating physicians in a manner conducive to production of standardised reports, e.g.** *stress test vs. cardiac troponin-I to rule out myocardial infarction (T106)*
Difficult	**Comparative tests differ in physical characteristics but are not associated with significant harms, e.g.** *addition of partogram to clinical notes, where both are left at the bedside (T81)* *Non-invasive ultrasound for investigation of peripheral artery disease vs. contrast-enhanced MR angiography that requires contrast injection (T70)* **Complex comparisons, such as >2 study arms and/or triage trials comparing tests otherwise amenable to blinding, e.g.** *Prompt endoscopy vs. medical treatment ± endoscopy vs. carbon urea test ± endoscopy for management of dyspepsia, where endoscopy is the accepted gold standard (T65)* **Tests very different but at least one is conducted under anaesthesia, e.g.** *Empirical therapy vs. gastroendoscopy for detection and treatment of H.pylori (T105)*	**Tests that produce similar results amenable to standardisation, usually conducted by treating clinicians but ethical for non-treating physician to conduct, e.g.** *Endotracheal aspiration with non—quantitative culture vs. Broncho—alveolar lavage with quantitative culture to identify organisms causing infection (T62)*

Table 6.26: Characterising the feasibility of blinding patients and clinicians to test use (continued across page).

	Patient	Clinician
Impossible	**At least one diagnostic strategy carries risks of significant harm that are too different from comparative strategies to accommodate ethical blinding, e.g.** *comparison of non-invasive 13-Carbon urea breath test and invasive gastroendoscopy for detecting Helicobacter pylori infection (T44)* *conservative pharmacological treatment compared to invasive percutaneous coronary intervention or open heart surgery for the management of coronary artery disease (T6)*	**Comparative tests produce different types of results which cannot be standardised, e.g.** *stress test vs. SPECT to detect cardiac ischaemia (T32)* *continuous fetal pulse oximetry vs. cardiotography to identify fetal distress (61)*
	Timing of diagnostic decision-making differs between interventions, e.g. *Early vs. delayed, selective hysteroscopy in infertile women (T73)*	**Timing of diagnostic decision-making differs between interventions, e.g.** *Early vs. delayed, selective hysteroscopy in infertile women (T73)*
	Patient response is an integral part of at least one comparative test, e.g. *Clinical examination to detect postoperative morbidity (T37)*	**Treating clinician must administer the test, e.g.** *Clinical examination to detect postoperative morbidity (T37)*
	The comparison takes place across different care settings, e.g. *hospital inpatient investigation vs. emergency specialist chest pain unit investigation (T16)*	**The comparison takes place across different care settings, e.g.** *hospital inpatient investigation vs. emergency specialist chest pain unit investigation (T16)*
	Knowledge of test results is central to the proposed benefit of at least one test, e.g. *addition of genetic test for familial hypercholesterolaemia to standard lipid profiles to change patients' perceived control over disease (T51)*	**Test and treatment are conducted by the clinician in the same procedure, e.g.** *two modes of endoscopy for the detection and treatment of Cholelithiasis (T9)*
		At least one test produces visual results used in subsequent treatment procedure, e.g. *duplex ultrasound used to locate varicose vein for subsequent removal (T58)*
		Nature of comparison is too complex to produce standard reports, e.g. *4-arm triage comparison of testing to select treatment in known coronary artery disease patients (T92)*

Table 6.26 cont.: Characterising the feasibility of blinding patients and clinicians to test use

good illustration, both tests being performed under general anaesthetic, in a virtually identical manner using similar equipment(T90) with the only difference in method being the instillation of a photodynamic dye, which could have been simulated with a placebo in white-light arm patients. Though more difficult to achieve, certain minimally-invasive imaging modalities could be given as sham procedures to conscious patients; for example when comparing duplex ultrasonography with contrast-enhanced MR angiography to determine the extent of stenosis in peripheral arterial disease, patients allocated to the latter could be administered placebo contrast injections as part of the sham angiography(T70).

Instead, blinding was likely to be impossible due to ethical concerns when the harms associated with undergoing diagnostic procedures differed between arms, such as when comparing conventional imaging to the implantation of a loop recorder to establish the cause of recurrent syncope(T56).

Certain types of tests were impossible to blind regardless of other aspects of the comparison, such as when patient response formed part of the test. This was often found to be the case with clinical examinations, for example the comparison of different formats for clinical consultation to assess the need for urgent referral in ED patients with neurological symptoms. In this trial the experimental consultation entailed video-conferencing with real-time visualisation of the patient, while the two comparator approaches were based on case-conference only, and so neither sham tests nor giving all consultations to all patients is possible(T75). Similarly, comparing dissimilar test procedures – whether invasive or not – that involved patient participation would not accommodate blinding, as was observed in an evaluation of three modes of tilt-table testing for diagnosing the cause of syncope(T2). In this example each test aimed to induce syncope by administering assorted pharmaceutical stimulants and tilting patients for varying lengths of time.

The range of comparisons in which care–providers can be blinded appears to be more constrained as successful methods rely on the ability to disguise the identity of test *results*, and in order to do so it is necessary to provide them with generic results or sham results. Based on the examination of the 103 included trials, this solution does not appear to be possible in the greater majority of test-treatment comparisons since all tests across all study groups would need to be administered and interpreted by a third party, as summarised in table 6.26. Accordingly, care–providers cannot be masked in any comparison where at least one test must be administered by them, including clinical examinations(T11,T37,T57,T75) or settings where testing and treatment must be conducted in the same procedure, for example potentially curative endoscopy(T1,T9,T24,T30,T35,T73,T88,T90,T95,T104,T108). This is also true for subjective test results that must be interpreted directly by the treating clinician, particularly when they were used to guide subsequent treatment, for example using preoperative duplex US to locate and map pathological veins for subsequent surgical removal(T58).

The type of diagnostic information produced by tests was equally important, and it would not be possible to create generic reports if the comparative tests described differing aspects of the target condition or provided different indicators for subsequent treatment. When comparing CT angiography to conventional stress testing for ruling out coronary artery disease as the cause of chest pain(T82) for example, stress testing allows the risk of stenosis to be determined while angiography would also quantify the *degree* of stenosis present, and so potentially inform the treatment approach in a different way. For similar reasons, comparisons of test versus no–test (T18,T34,T45,T49,T63,T71,T76,T102,T105), or comparisons evaluating the consequences of expedited decision-making, such as the provision of more rapid microbacterial identification in the management of bacterial infection(T23), both also preclude blinding.

The complexity of diagnostic strategies was a third barrier to blinding care–providers, due either to the sheer convolution of doing so for extensive batteries of tests(T34,T40,T42,T46,T92), or the impossibility of disowning diagnosticians from the decision-making needed to select patients for further testing or treatment in evaluations of multiple phases of decision-making, such as the introduction of triage tests (T16,T24,T33,T36,T44,T46–47,T60,T65,T73,T82,T92,T108). The role of the new test in the existing pathway also appeared to impact on the viability of masking patients, though clearly for different reasons. Here, blinding was more often possible in add–on (21/28, 75%) than either replacement (27/60, 45%) or triage (5/17, 29%) comparisons, probably due to the differences in the organisation of care that are at a minimum in add-on comparisons since the control intervention must be replicated in both arms. In fact, similarity in the organisation of care between interventions was crucial to the prospect of blinding both patients and care–providers. Of the four trials assessing the impact of new diagnostic units for the provision of specialised diagnosis and management(T16,T29,T48,T53), none could have masked patients or care–providers to the differing care settings. The use of different treating staff was often a component of these new programmes of care. For example, a new chest pain observation unit, situated adjacent to routine care in the ED, was staffed by specialist nurses who had received additional training in the management of acute undifferentiated chest pain and were not available during routine care(T48), thus necessitating a fully open trial.

Barriers to blinding outcome assessors

The finding that it was impossible to blind outcome assessments in a third of trials is at odds with existing hypotheses that such blinding should always be possible[22,174]. The biggest barrier was the impossibility of masking care–providers and patients, who were frequently directly involved in the assessment of primary outcomes (table 6.27).

Though common, patient–reported outcomes were particularly difficult to blind, yet these tend to be the most subjective and hence open to the influence of patient (or clinician)

	Outcome Type	Examples	Feasibility of blinding the outcome assessor
1	Patient reported	Pain, quality of life	Same as blinding the patient
2	Patient –outcome assessor contact required	Walking speed, function	Feasible if patient is blinded Impossible if patient is not blinded
3	Patient–outcome assessor contact not required	Appearance of joint structure (X-ray)	Always feasible
4	Clinical events and therapeutic outcomes determined by interaction between patient and clinician	Length of hospital stay, treatment failure	If measured by clinician: same as blinding the clinician If measured independently: always feasible
5	Clinical events and therapeutic outcomes assessed from data on medical forms	Death, treatment prescription	Feasible if relevant records do not contain information on diagnostic tests administered Difficult if records include tests administered

Table 6.27: Characterising the feasibility of blinding outcome assessors to test use

expectations. When possible, the use of independent assessors or adjudication panels could overcome the inability to mask clinicians, yet whether this approach succeeds in removing ascertainment bias is tricky to evaluate. Outcome data given to blind panels may have been recorded by unmasked staff; when this information is based on subjective interpretation (say an MRI report for example), the possibility of distortion due to ascertainment bias theoretically remains. This raises an interesting issue: who should trialists be blinding to eliminate ascertainment bias in test-treatment trials? In some settings, such as the venous thromboembolism trials discussed earlier (p. 170), blinding the individual who performs the follow–up test may be key to ensuring that ascertainment bias has been avoided.

Implications for not blinding in test-treatment RCTs

According to the tenets of standard trial design, these findings present worrying implications for the validity of some test-treatment RCTs. If trials cannot blind participants, clinicians or outcome adjudicators, then little can be done in these cases to protect from the threats of performance and ascertainment bias. On the other hand, this review suggests there is considerable room for improvement. Approximately half of all trials could have introduced measures to blind patients and outcome assessors, and a more modest 8% could have attempted to blind physicians. Therefore, though the methodological quality of existing test-treatment trials is sub–optimal, many more comparisons could control for performance and ascertainment bias than is currently attempted.

This conclusion assumes that the consequences of blinding will always serve to *minimise* performance and ascertainment bias. Yet there are indications that this may not always be the case in trials of test-treatment interventions.

The cohort of trials included some for which blinding patients or care–providers could have altered or even eliminated the desired treatment effect. This hypothesis is directly related to the proposed benefits of the new test-treatment strategy. For example when the value of the new test lays in its ability to alter how patients respond to their management, trials would not be able to measure this effect if participants were no longer aware of which tests were used to inform their future treatment. This was observed in one of the back pain trials where in order to measure whether the addition of MRI to standard orthopaedic consultations would reassure patients suffering with acute lower back pain as to the absence of serious disease, the investigators gave MRI to all randomised participants [(T4)]. Similarly, the methodological benefits of blinding care–providers can also be uncertain, particularly when there is no rigid link between test results and management decisions. In order to mask treating clinicians we must remove their ability to interpret test results, and depending on the types of tests involved also to make the diagnosis, therefore blinding could prevent investigators from fully evaluating the test's impact on clinical decision–making.

This suggests that blinding may play a very different role in test-treatment trial methodology, compared to treatment RCTs. Test-treatment interventions share fundamental similarities with therapeutic complex interventions, in that behavioural change can be integral to how test-treat strategies are expected to benefit downstream patient health interventions, and are not a peripheral, incidental effect[245]. Since test-treatment RCTs commonly evaluate the consequences of patient or clinician behaviour, such as the impact of decision-making in the face of new or earlier information, balancing the need to evaluate particular effects with the need to minimise performance and ascertainment bias is likely to pose considerable difficulties to the design of test-treatment RCTs.

6.8.4 Trials often exclude patients & fail to protect from attrition bias

The choice of outcomes used raises some concern as to the usefulness of test-treatment trials. Although the majority of trials analysed patient outcomes as a primary question, 56% of these (30/53) were surrogate measures of health, namely clinical endpoints and disease rates. It would seem, therefore, that when test-treatment interventions are evaluated using randomised trial designs they do so in the main to answer intermediate questions of process and short–term health impact, whose relevance to long–term health may be questionable[246]. This issue is examined further in chapters 7 and 8.

Inappropriate comparisons

A small proportion of studies (9%) conducted inappropriate analyses by using subgroup denominators to analyse their primary measure of effect. Such comparisons are at risk of producing misleading results that are not full and true reflections of an intervention's effectiveness. Firstly, selecting outcomes that are targeted at a particular subgroup ignores the consequences of testing and treatment in patients excluded from the subgroup. Although one can appreciate the temptation, particularly in test-treatment evaluations where the clinical need to examine the effects of the care process on individuals who have been misdiagnosed (false–positive or false–negative patients) is justified, these otherwise legitimate comparisons should remain as secondary explanatory additions to a primary aim that evaluates a strategy's impact to all randomised participants[135]. By looking primarily at the false–negative rate for example, extreme differences in negativity rates between the testing strategies would not capture the consequences of overtreatment. As a consequence, the 13 studies using such subgroups, albeit a minority of the cohort, must be considered to have produced partial and therefore unreliable accounts of downstream treatment effects.

Secondly, subgroup analyses strongly risk providing distorted effect sizes by comparing two groups that are not random samples of the original study group, and so may well vary in important prognostic factors. Moreover, unlike treatment trials where subgroup analyses are generally identified using baseline characteristics[247], in test-treatment trials the subgroups are created by different tests and so the composition of subgroups will always represent unequally selected populations. The impact of this selection bias was illustrated in at least two included trials where the author's recalculation of endpoints based on all randomised participants served to change the trial's primary conclusion.

In a related issue, a small number of trials used differing methods to measure the primary outcome across study groups. This approach is virtually guaranteed to cause bias in test-treatment trials; since different tests will identify different patients in each arm the outcome is based on a systematically different ascertainment method which could distort results. This issue was only encountered in between 2–4 trials, suggesting that it does not pose a frequent threat to the validity of published test-treatment RCTs.

Intention–to–treat, missing data and attrition bias

The majority of trials reported an intention–to–treat analysis (ITT), similar to the performance of other cohorts of RCTs, which have found reported rates of 48%[248], and more recently 54%[249] to 62%[250]. Test-treatment trials appear to perform considerably better than surgical trials where only a third (36%) reported an ITT analysis[239]. Although most trials appeared to adhere to the principle of analysing patients in the groups to which they had been assigned, most test-treatment analyses proved to be inconsistent with the most rigorous definition of ITT[123], with only 21% also including all randomised participants. Previous studies have discovered similar rates of true ITT analyses in reviews of trials (24%[248]; 28%[104]) though a more recent analysis shows a higher rate of 39% suggesting practice is improving in general[250]. Boutron and colleagues[240] found that pharmaceutical RCTs performed better than non–pharmaceutical RCTs in this respect (35% vs. 25%), which along with the present findings may indicate that full ITT analyses are more difficult to carry out in complex intervention trials. This may be a consequence of the practical difficulties involved in maintaining a proper ITT analysis, whereupon the risk of losing participants after randomisation is increased due to undergoing numerous interventions. This is somewhat substantiated by the high rates of incomplete analyses due to missing outcome data. The common inability to blind in these trials is also likely to have driven drop–out rates and missing responses, since patients dissatisfied or disillusioned with their diagnostic allocations may have been less motivated to comply with the trial follow–up protocol. This is demonstrated in at least one trial, where participants with suspected fractures randomised to receive an additional MRI scan to the usual X–ray were more likely to return their questionnaires than those who knew they had not received the new technology[(T5)].

Over half the included trials presented incomplete analyses due to exclusions and/or missing responses. This is lower than the 75% of trials published in four high impact factor journals during 1997 that reported an ITT analysis[248], though akin to more current reviews finding 58% in a group of randomly selected RCTs published in 1999[251] and 60% in general RCTs published in 2002[250]. A quarter of test-treatment trials excluded at least 10% of participants, corroborating early previous research[248] though considerably higher than the more recent reviews that both report rates of around 10%[250–251]. This comparably high proportion is of concern, particularly since so few studies stated their strategy for handling missing data in their analyses.

The most common approach, deduced implicitly from the numbers reported, was a complete case analysis in which protocol deviations or missing responses are excluded. This method is inadequate since it damages the distribution of prognostic factors ensured through a proper randomisation process, creating a subgroup comparison which will produce biased results unless participant data are missing at random[252]. Trials performing such exclusions are shown to produce larger treatment effects with smaller p–values, though the direction of bias is variable[105]. Imputation methods require assumptions of missing data that may be difficult to justify, introducing uncertainty around treatment effects and potentially themselves introducing bias to results[121]. Nevertheless some approach to include missing data is required since these are likely to be related to diagnosis, treatment response and other prognostic characteristics[253]. Sensitivity analysis using multiple imputation is the preferred approach as, though difficult, it allows the impact of several imputation strategies on the treatment effect to be quantified[254]. No included study reported that they had done this, although it is worth highlighting that these trials largely predate the widespread dissemination of multiple imputation methods.

Disappointingly, most trials provided insufficient reporting of the reasons for missing data for the reader to determine whether the resulting subgroups were likely to differ systematically. A single trial openly discussed end composition, stating: “Patients who were lost to follow-up were older and more severely depressed than patients with complete follow-up"(T66), with clear implications for the validity of its results.

The threat to internal validity was particularly high in the 16% of trials that displayed attrition differing between arms by more than 5%. Experimental interventions were more likely to lose higher proportions of participants and, given that these patients were probably the least likely to respond positively to treatment, this finding would indicate that attrition bias may have served to favour the new strategies in these studies. Although serious threats to internal validity were only found in one trial, with a differential attrition of over 20%, the commonplace nature of nonetheless significant attrition in test-treatment RCTs attests to a poor general level of methodological quality and hence questionable validity of trial results. Unfortunately the author could find no published data on differential attrition rates in general cohorts of RCTs with which to compare this interesting finding.

6.8.5 Trials may be chronically underpowered

Performing *a priori* power calculations is a necessary aspect of methodological quality, since it demonstrates that the required balance between risk of type II error and sample size has been deliberated in advance of a trial[121], and ensures that recruitment ceases in line with this predetermined target rather than in response to favourable interim analyses[128]. Present in almost 80% of trials, such reporting compared favourably to other cohorts of trials in which generally half or fewer provided power calculations[118,152,240,249].

As discussed in chapter 1 (p. 38), attaining a sufficient sample size with enough power to detect differences is an important consideration in test-treatment RCTs. Although it was outside the scope of this review to assess the adequacy of published sample size calculations, the median achieved sample sizes are somewhat smaller than would be expected for target populations expected to be so much bigger than standard RCTs. In a review of parallel–group RCTs (any intervention type) published during the same timeframe as the present cohort (2005–6), median sample size was considerably larger at 425 per arm (IQR: 158–1041)[130] than found in the present test-treatment trials (median 166, IQR: 72–297). Moreover, trials primarily aiming to evaluate the impact of interventions on patient health did not recruit significantly greater numbers than those primarily interested in process outcomes. Although rather a crude interpretation, these preliminary findings may suggest that test-treatment RCTs are chronically underpowered to detect clinically important differences in downstream health effects.

This study also reveals that test-treatment RCTs were found to suffer from practical difficulties in achieving target sample sizes, and several trials reported considerable problems in recruiting sufficient numbers, or completing recruitment before the experimental practice became routine. That over 10% of included RCTs failed to recruit at least 95% of the target number of participants provides further indication that these studies are likely to suffer from a heightened risk of type II errors.

6.8.6 Study limitations

Due to time constraints, several interesting and important aspects of test-treatment trial validity unfortunately remain unexplored. For example, the review did not trace protocols in order to appraise selective outcome reporting, which is empirically demonstrated to be associated with overestimations of treatment effect since unpublished outcomes are less likely to be statistically significant findings[152].

Perhaps the most important omission of this study was its failure to directly address hypotheses that test-treatment trials are at increased risk of type II errors. Though the seemingly small sample sizes suggest included trials were underpowered to detect differences in downstream health outcomes, this cannot be confirmed without re–estimating power calculations using the suggested inflation factor[176] to adjust for the unreclassified fraction. The difficulties encountered by the author upon initial attempts to find parameter estimates for one study prohibited this more comprehensive approach, and so it must be left to future studies to verify this hypothesis rigorously. However, this obstacle may also suggest that performing power calculations may be particularly challenging for trials of test-treatment interventions, where empirical evidence of diagnostic efficacy is less likely to exist than the evidence required to estimate power for trials of pharmaceutical treatments.

As discussed in the limitations to the review of reporting quality (p. 142), although the project cohort does not contain all relevant test-treatment RCTs published from 2004–2007, it is arguably unlikely that their examination would drastically change this review’s conclusions. Even if missed trials were of better quality than encountered in this cohort, this would not invalidate the considerable difficulties in achieving the rigorous methodological standards that were observed in many included trials.

The quality of reporting trial methods, and by association methodological quality, have improved over the last 15 years[118,153] so it is possible that trials published since 2007 are of better quality. However if this is the case it is only likely to apply to items found to be equivalently achievable in test-treatment and general RCTs, since methods to address those that may require different or more exerted efforts for test-treatment evaluations have not yet been disseminated.

Most of the existing studies used for comparative discussion, other than those specifically selected to observe the quality of complex interventions, did not select trials for review on the basis of the types of intervention evaluated. Consequently complex intervention trials did form part of these cohorts, though always constituted a small minority. The comparison between such ‘general’ cohorts and purely complex intervention cohorts is thus likely to remain accurate and valid.

Variations in the criteria used by previous reviews to appraise the adequacy of reporting and methods mean that not all reviews may be entirely comparable to the current dataset. Somewhat paradoxically, these reviews commonly fail to define their appraisal criteria explicitly, thus limiting a proper judgement of whether their results are consistent with those of others[255]. In order to limit an unfair comparison, the methods used to appraise reporting and quality in the current review were chosen to reflect the highest standard of methodological appraisal as recommended by leading groups of methodologists[121,123]. Efforts were also made to compare findings against reviews using these methods, however this was not always guaranteed for reviews of complex interventions. Consequently, the rigour with which current methods were applied may have reduced the apparent performance of test-treatment RCTs if compared to more leniently appraised cohorts of trials. Albeit the case, the precise methods used have been detailed to allow future researchers the independent verification of this review's results.

Ultimately, the review has relied on the common approach of using trial reports as a proxy for actual trial design and conduct, and this has been empirically shown to conceal true methodological quality[134]. Since poorly–designed trials that are reported well will reveal their true methodological quality, this is likely to have led to an underestimation of the quality of test-treatment RCTs. Nonetheless, this confounding factor affects all such reviews of trials, so as a comparative measure of test-treatment RCT performance this review's findings stand. Moreover, adequate reporting remains key to the 'usefulness' of trials: study findings, and the methods used to produce them, are predominantly accessed through journal articles and so the ability to understand trial processes, interpret results and consider reproducing them in practice relies on the content transmitted by them.

Finally, the subjective nature of quality appraisal would caution that the results presented here are proposed as tentative results in need of validation through future analyses of test-treatment trial cohorts. The high levels of interobserver agreement for adequacy of randomisation methods and blinding status do suggest the results are reliable. However, not all methods were subjected to this analysis. Assessing the feasibility of blinding was a particularly subjective exercise, at times impeded by poor reporting of trial conduct, and so could have been influenced by the author's preliminary hypothesis derived from the literature that blinding is difficult to achieve in test-treatment RCTs. It would be interesting to explore the degree of agreement in these judgements by comparison to a second independent assessor. Moreover, the present author's clinical inexperience cannot preclude the possibility that certain assessments may have been misjudged. While efforts were made to research the clinical practices encountered, the relevance of findings and conclusions drawn must be evaluated by the physicians for whom test-treatment RCTs have been designed.

6.9 Conclusions

To conclude, this review finds a clear need for improvement in the methods used to conduct test-treatment RCTs. Several weaknesses in design were observed that, by reference to the findings of methodological quality meta–reviews, suggest this cohort of trials are at risk of several types of bias. The widespread exclusion of noncompliant cases and inappropriate use of subgroups may have produced misleading trial results since there is no guarantee that a like–for–like comparison is being made. In particular, the rarity of blinding – in particular outcome assessors – suggests these trials are more likely to be measuring the expectations of trialists and preconceived notions of participants than the true effects of test–treatment strategies. Meanwhile other limitations, including inadequate randomisation and concealment procedures, may have caused trials to create study groups that were systematically different at the outset and thus also produce biased results.

These failings are partly explained by the suboptimal standards of methodological quality generally found in all RCTs[97,107]. However, the differential performance of particular methodological items when compared to standard, non–complex intervention trials does appear to confirm that certain methodological problems are more prominent in test-treatment trials. Specifically:

1. While the number of RCTs presenting incomplete analyses was equivalent to frequencies reported by reviews of treatment RCTs, test-treatment trials tended to have higher proportions of missing data.
2. A seemingly high proportion of trials (16%) presented with attrition that differed between arms by more than 5%.
3. Despite theoretically needing considerably higher sample sizes, test-treatment trials had far smaller study populations than trials of treatments, suggesting they may be chronically underpowered.
4. Blinding patients and care–providers was far more rarely conducted, and generally impossible to perform. Contrary to expectations, blinding outcome assessors was also not always feasible.

The following chapter completes the analysis of these trials by examining the extent to which they fully evaluate test-treatment interventions.

This chapter has been published in part:

Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests – A framework for designing and evaluating trials. BMJ 2012:344:e686.

7

How tests change patient health:

Development of a conceptual framework

Understanding how diagnostic tests influence patient health is a matter of significant relevance to trial design and interpretation, yet the relationship between the two is fundamentally indirect and complex. This paper develops a theoretical framework that sets out the common mechanisms by which tests can change patient outcomes. It builds on existing thought by adding new mechanisms identified through an analysis of how test-treatment interventions have been observed to impact on patients. The framework is tested, refined and explained using the project cohort of published test–treatment RCTs, identified in Chapter 2.

7.1 Introduction

The reviews of reporting quality (chapter 5) and methodological quality (Chapter 6) concluded that insufficient documentation of test-treatment interventions, including decision–making, is a key impediment to the utility of trial reports, since it hinders the interpretation of results by concealing what is being evaluated. Another vital requirement for useful trials is the selection of endpoint measurements that will capture all the intended effects of an intervention; this is hypothesised to be particularly challenging for test-treatment comparisons where tests are indirectly related to patient health.

In order to evaluate this contention, the ideal analysis should appraise the *appropriateness* of outcome selection in included trials. Performing a reliable judgement would require extensive clinical expertise to be able to identify the most important outcomes for the full range of diagnostic settings included. Critically, the appraisal should also be made by reference to a solid theoretical model of how test-treatment interventions cause treatment effects. Yet examination of existing diagnostic research frameworks suggested such a standardised theory was lacking and in urgent need of development.

Several authors had attempted to delineate the specific contribution of testing to patient health, commonly expecting that changes would be driven by improvements to test accuracy[14,19,21,49–51,77,176,193,256] or reductions in procedural harms[19,45–46,48,77,193], however no existing study provided a complete overview of the ways in which changes to patient outcomes could occur as a consequence of introducing a new test.

Upon further investigation the absence of such a conceptual model, or rather the under–development of the existing theoretical perspective, was found to have been voiced as an important barrier to reaching a consensus on which aspects of tests should be measured, how they should be measured, and why[13,77,187,257–259]. Given that the relationship between testing and health outcomes is fundamentally indirect, others had described the ensuing difficulty in disentangling a test's contribution to observed changes in patient outcome from intermediate elements of the test-treat process[6,59,175,260]. This principle is recognised in complex intervention guidance[83], where the importance of developing a good theoretical grasp of how the use of a complex intervention may impact on patient health is considered essential to constructing a sound scientific rationale for the intended behaviour of the intervention. Without a similarly comprehensive understanding for the use of tests, it is impossible to evaluate whether outcomes measured in a trial capture the true effects of test-treatment interventions, and therefore to determine how thorough the interpretation of results has been.

In order to address this important question, the author sought to develop a framework of thought that sets out the relationship between diagnostic tests and patient outcomes. By building on existing theoretical notions, its aim is to identify and explain the mechanisms by which a diagnostic test can contribute to changes in patient health. The new, comprehensive framework is presented as a graphical structure that models the test–treatment care pathway and illustrates how its components interrelate to change different measures of patient health.

7.2 Methods

The main aim of this study was to formalise and expand existing knowledge of how diagnostic tests can affect downstream patient health. This required an explorative and theory–generating approach, in which existing theoretical assumptions could be summarised and broadened into a comprehensive explanatory structure which specifies all the potential consequences of testing to patient health. In light of the limited current theoretical understanding, the cohort of published test-treatment RCTs (identified in chapter 2) were considered key to developing the explanatory structure since they might shed light on the diagnostic determinants of health in a variety of clinical settings. In this way, the theory could be grounded in empirical comparisons of diagnostic strategies in order to gain as accurate a representation of reality as possible.

This approach followed the principles of analytic induction[261] whereby theories are generated from the observation and analysis of empirical data, in contrast to the deductive reasoning more commonly employed in evidence–based health care where data are selected and interpreted using existing theory. Key to analytic induction is the role of ‘disconfirmation’ whereby after tentative explanatory hypotheses are constructed, they are developed and refined by finding cases that do not entirely adhere to the new theory[262]. This is an iterative process in which theory is under constant revision whilst all available evidence is examined, and where the development of concepts and the relationships between them are primarily driven by new insights provided by the analysis of ‘deviant cases’[263].

The method was selected for two reasons. First, though existing research frameworks address the issue of how tests are linked to patient health obliquely, the notions contained therein are nonetheless highly relevant to such a discourse, and so it was considered important to use these existing principles as a foundation for developing a generalised, explanatory theory. Second, the availability of 103 published test-treatment RCTs provided the opportunity to gain empirically–based insights into how tests have been shown to influence health outcomes, and so these primary studies should be used to test the strength of existing explanations and elaborate current thinking.

Following these preliminary deliberations, the explanatory framework is developed in three steps:

7.2.1 Theoretical premise

The first step was to define a basic theoretical premise of how tests are connected to patients. With reference to discussions presented in the existing theoretical literature, the key clinical processes that a patient proceeds through from the point of being referred for testing to their response to treatment were identified.

7.2.2 Preliminary explanatory features

Secondly, existing research frameworks were reviewed for implicit assumptions or explicit mention of how these key clinical processes may be of value to downstream patient health. These existing notions were added to the initial premise to create a preliminary generic model of the diagnostic healthcare process, containing a set of factors which might explain differences in how patients respond to treatment.

7.2.3 Explanatory model revision using 'deviant cases'

The validity of this representation was examined by using it as a structure to interpret how differences in observed health outcomes had been created in test-treatment RCTs. For each trial competing test–treatment processes were reconstructed and used to populate the generic model. Focussing on patient outcomes measured by the trial, the model was used as a tool to consider where and how observed differences in patient health might have originated, in a systematic manner. Any trial that appeared to be conceptually different, a ‘deviant case’, was analysed in depth. Where existing explanatory concepts failed to account for observed findings, the model was supplemented by new hypotheses. In this way, the framework was applied to each trial in order to be sure that it could explain all available test-treatment comparisons. This allowed the theoretical framework to be developed and modified through an iterative process of testing and retesting specific hypotheses using the cohort of published test-treatment trials, in order to produce a revised and comprehensive framework of thought.

7.3 Framework development

The current theoretical understanding of how tests create treatment effects is derived from three main strands of conceptual research. Foremost in both number and the attention they have received are the frameworks outlining a ‘phased evaluation of tests’, in which an idealised set of evaluative stages are proposed as necessary to introduce a test into practice[15,44–54,57]. As outlined in Chapter 1 (p. 7–8) these organisational structures contain between four and six stages of evaluation, which tend to be arranged hierarchically on the premise that tests which are *not* efficacious at a given evaluative stage will not be capable of efficacy at higher and more complex phases of evaluation.

Closely related to these is a group of more theoretical studies that primarily discuss the methodological basis of test evaluation, often within the context of an analytic infrastructure of test evaluation[6–7,14,19,51,55,59,77,84,175–176,193,257–260,264–265]. The now seminal work by Fryback and Thornbury[14] is one such conceptual framework, which though traditionally considered as one of the founding ‘phased evaluation’ frameworks was in fact intended as a more general conceptual discourse seeking to draw together the study designs available to evaluate aspects of test performance, and the methodological issues presented therein[57,265]. A second important contribution to this group are the proceedings of an international workshop[77,175,193,257,259–260,265] convening diagnostic practitioners and methodologists in test evaluation to explore the methodological challenges involved in assessing the outcomes of diagnostic testing. The resulting set of papers provide a particular insight into the conceptual difficulties regarding how to determine which endpoints are necessary to demonstrate that a test is efficacious[259].

The remaining theoretical frameworks address the extent to which accuracy studies and simplified designs can replace direct evidence gained through RCT designs, and so are largely concerned with developing methods to synthesise and appraising existing evidence[19,21,91,93,256].

While none of the existing frameworks is specifically directed at explaining how tests influence patient health per se, each reveals valuable insights into the conceptual basis for this theory, often through assumptions authors make regarding the link between testing and health outcomes when discussing appropriate methods to evaluate tests.

7.3.1 Basic premise: defining how tests are linked to patient outcomes

A central premise of the current theoretical understanding is that the relationship between tests and measures of patient health is fundamentally indirect. Acknowledging that diagnostic tests are not administered in isolation, but form part of a broader clinical process in which a period of testing is followed by decision-making, management planning and treatment implementation, existing research frameworks are all inherently founded on the principle that tests are linked to patient outcomes by the intervening clinical processes.

The basic relationship between tests and patients may therefore be conceived as a series of intermediate steps occurring between the two. Figure 7.1 depicts this relationship as an adaptation of a diagram presented at the 1999 international workshop by Bossuyt and Lijmer[193], who of all reviewed frameworks most clearly delineated the process. It portrays a

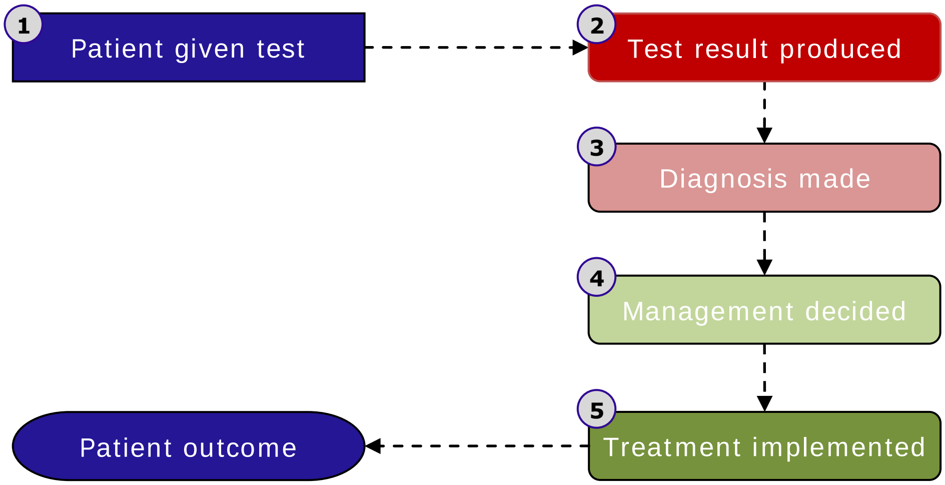


Figure 7.1: Simplified test-treat pathway showing each step in the pathway as a component of the patient's management which can influence patient health: (1) Patient given test, (2) Test result reported, (3) Diagnosis made, (4) Management decided, (5) Treatment implemented. (*Adapted from Bossuyt and Lijmer 1999*[193])

simplified patient care pathway, in which a diagnostic test is applied to a patient in order to identify a condition [1], the result of the test is considered [2] along with other evidence to decide a diagnosis [3], from which a course of treatment is identified [4] and implemented [5]. Considered in this way, each step along the test–treat process becomes a component that could contribute to improvements in a patient's health.

7.3.2 Preliminary explanatory features

The conceptual papers revealed three theories regarding how aspects of the patient care pathway might impact on patient outcomes. There was a general consensus that the key prerequisite for a new test to affect downstream patient health is its capacity to change diagnostic decision–making, a task achieved by providing superior accuracy[6–7,14–15,19,21,44–51,53–55,77,84,91,93,176,187,193,256–257,260,265]. Furthermore, these changes should alter subsequent treatment decisions for a difference in health outcomes to be observed. This causal link is reflected in the hierarchical structure of the 'phased test evaluation' frameworks. Each of the five care pathway components in figure 7.1 may be likened to a corresponding phase of test evaluation, whereby each phase produces outcomes measuring the utility of a test that can be likened to diagnostic attributes. As these attributes share a hierarchical relationship, we can posit that one will have an effect on another further up the hierarchical chain.

So, at the bottom of the hierarchy lay evaluations of technical efficacy and diagnostic accuracy, attributes of a test that reflect its ability to produce classificatory information reliably (precision) and accurately (a test's ability to identify or exclude disease compared to a reference standard)[14,44–46]. Further up the hierarchy are evaluations of 'diagnostic thinking efficacy' producing measures of diagnostic yield, defined by Fryback and Thornbury[14] as the appraisal of how far test results influence the diagnoses patients receive. Patient management decisions are conceived as separate decision–making events since tests that succeed in changing diagnostic thinking may not always affect therapeutic management. Thus a test's 'therapeutic efficacy' must be assessed separately by measuring therapeutic yield[14,44–46]. At the summit of most of these frameworks lays the evaluation of patient outcomes, variously referred to as 'clinical outcome efficacy' or 'patient outcome efficacy', the patient's health. Here the impact on patient outcome is described as the composite effect of all above elements[45–46,49–50,176,193,256].

A second common assertion was the direct impact that tests can have on patient health[6,19,21,45–46,48,77,84,93,176,187,193]. Conceived in the main as procedural harms, tests were noted to influence immediate health outcomes regardless of whether subsequent test results correctly identify patients with or without disease.

Thirdly, several authors drew attention to the potential for test results to reassure or cause anxiety to patients, or otherwise affect their perceptions of health and disease[19,48,175,193,265]. These values of testing are seen to lay outside the tiered efficacy model and are defined

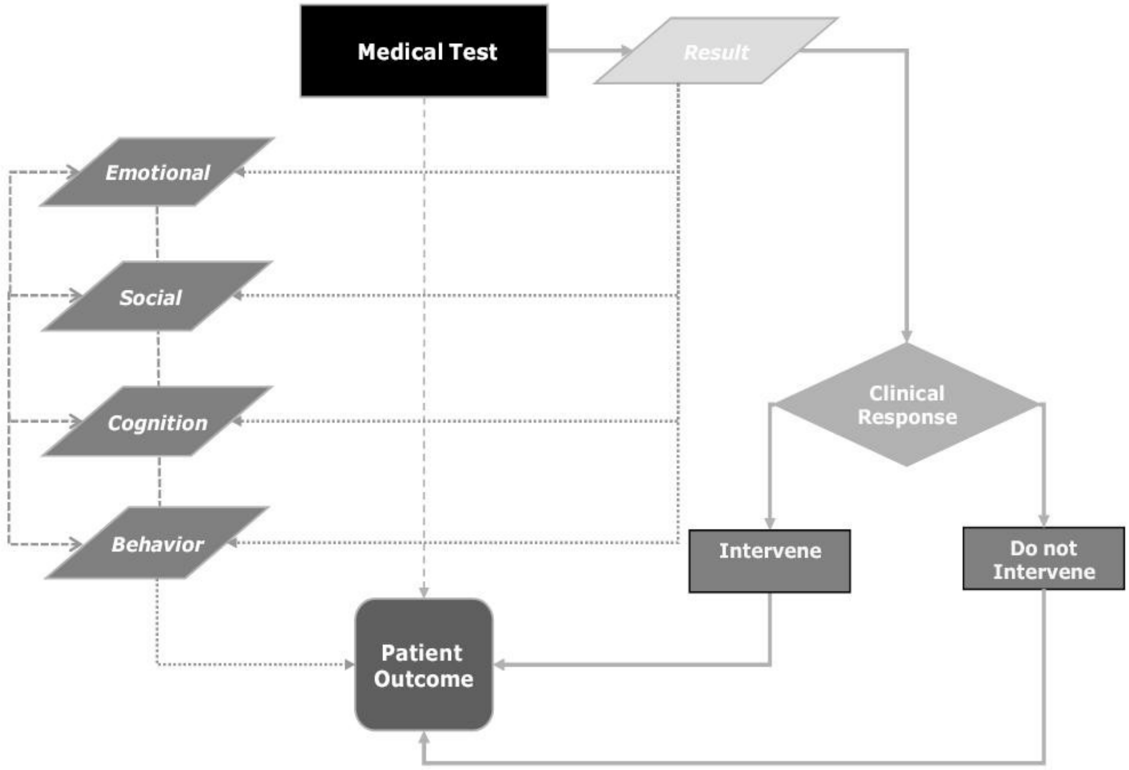


Figure 7.2: Framework for conceptualising the psychological impacts of testing, designed by Bossuyt and McCaffery[264]. Figure reproduced with permission.

as ‘non–decisional’ attributes, set apart from the effects of guiding more appropriate therapy through better diagnostic accuracy[14,175]. These concepts are explored comprehensively by Bossuyt and McCaffery[264] who delineate the cognitive, emotional and behavioural changes that can occur both directly, and through a patient’s knowledge of test results (Figure 7.2). Their model incorporates the understanding that both the test procedure and the information it produces can affect patient outcomes in a way that is mediated through a patient’s cognitive facilities, influencing emotional responses, treatment compliance and perceptions of self and health state. Guyatt and colleagues noted that these effects were not limited to patients, but emphasise that the degree to which physicians are reassured by a given test, perhaps even falsely so, could influence the adequacy of treatment and consequently patient outcome[44].

Pulling these concepts together, the current theoretical position conceives tests to affect patient health in three ways: 1) through the provision of improved decision–making, 2) by reducing the harmful effects or increasing the beneficial consequences of undergoing testing, and 3) through providing different diagnostic information that modifies a patient's perception of their health state.

7.3.3 Expanding the existing model

In order to test the preliminary model, the first task for each trial in the cohort was to represent its test-treatment processes of a trial diagrammatically. Using the patient care pathway, the comparative tests and diagnostic and treatment decisions were mapped onto the diagram representing a simplified trial algorithm. The properties of each test and its output were positioned at the beginning of the pathway, and patient outcome added at the terminus. This representation then allowed the author to conceptualise which elements may contribute to an observed change in patient outcome effect, as described above. An example of a 'deviant' trial is worked through in Box 7.1. During this process certain difficulties arose as the number of contributing factors and interrelationships became more complex. It became necessary to think very carefully about each specific clinical and study setting in order to track the care pathway correctly, paying particular attention to the type of diagnostic comparison, the diagnostic and treatment purposes of the test(s) under evaluation, the target condition, patient group and care setting. In rare cases where the rationale for the intended impact of tests was provided, this tended to be couched in more ephemeral descriptions (generally found in introductory statements) regarding the proposed benefits of using one diagnostic strategy as opposed to another.

No cases were found to disconfirm the relevance of any previously identified attributes. However this process proved very fruitful for the expansion of the framework, in that it identified additional diagnostic attributes that might drive changes or otherwise facilitate

Box 7.1: **Worked example of a deviant trial:** Antibiotic selection patterns in acutely febrile new outpatients with or without immediate testing for C–reactive protein and leucocyte count [(T7)]

Patient Group: Outpatients with acute febrile condition (clinically relevant fever of >37.5˚C) suspected of having an infection

Test Comparison: Clinical consultation with immediate C–Reactive Protein (CRP) response and white blood cell count (WBC) results versus clinical consultation only

Comparison Type: Additional

Diagnostic question: Does the patient have an infection, and if so is it bacterial or non–bacterial?

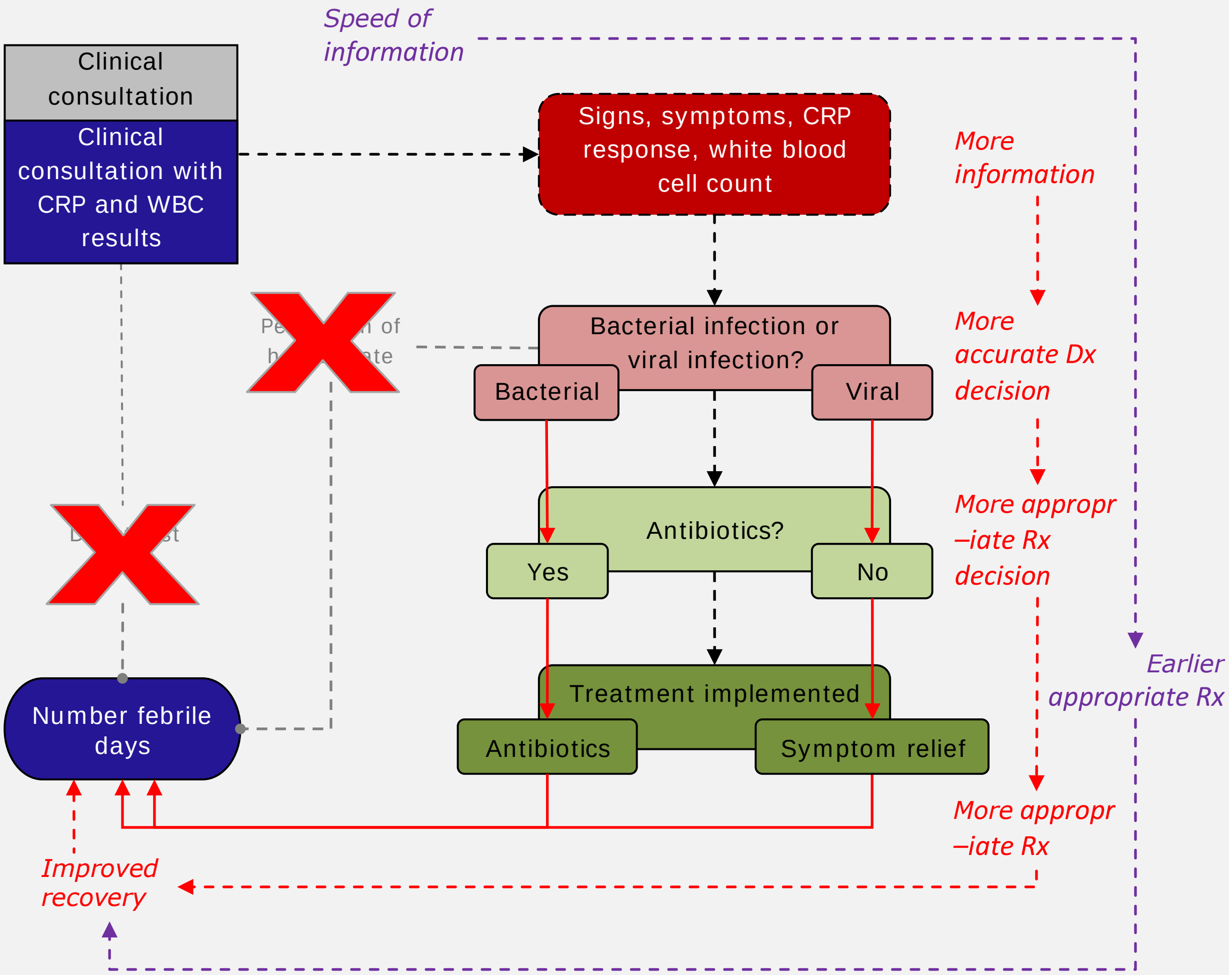
This trial was conducted to evaluate whether the immediate availability of two biochemical tests (CRP and WBC) at the time of initial clinical consultation would improve the management of patients presenting with a possible infection. Patients randomised to the control arm received a consultation in which clinicians did not have recourse to biochemical test results, however clinicians were free to order these tests whereupon results would be available at a later consultation.

The primary outcome was not defined clearly, but was extracted as the proportion receiving antibiotic treatment which corresponded with the study aim reported in the article's introduction. The aim of this exercise was to determine how tests impact on *health*, hence the patient outcome measured by trialists is used here: the number of febrile days. The figure below illustrates the reconstructed patient care–pathway for the trial. The process of elucidating how the addition of immediate biochemical testing could influence the number of febrile days began by considering whether existing factors could be responsible, after which the author conceptualised whether previously unidentified factors could also cause any differences. This rationale is now explained.

How could differences in the duration of fever be created?

Direct test effects are not relevant in this comparison, since neither of the testing strategies is invasive. The availability of additional information could possibly alter the dynamic of the initial consultation and thus influence the patients' perception of their health state, although in this setting it is unlikely that this factor could change the duration of fever. This outcome is a measure of patient recovery, and so is likely to be influenced predominantly by the administration of appropriate treatment. Since the new strategy provides the clinician with additional information, which according to the authors has been shown to differentiate accurately between bacterial and non–bacterial origins of infection, one could expect it to improve patient recovery by allowing more patients to be diagnosed accurately, more of whom would then receive appropriate treatment (red arrows). Diagnostic uncertainty is also important here, and as the biochemical tests are familiar the new strategy is expected to maximise appropriate decision–making by improving

Box 7.1 continued



***Comparative patient care pathways mapped onto the preliminary model**. CRP – C–reactive protein response; WBC – White blood cell count; Dx – diagnosis; Rx– treatment.*

confidence that the true aetiology of infection has been found.

Looking at the comparison more closely, it became apparent that improvements to patient health may have a second important origin, which could not be explained using existing identified factors. Clinicians using the control strategy were free to order additional tests to aid their decision–making, although clearly these results would be delayed with respect to the experimental strategy. Thus an important comparison hidden within the study design is that of the *speed* of appropriate decision–making (purple arrows): the new strategy allows important information to be available more quickly, which in the context of acute infection could certainly influence a patient's recovery by enabling the right treatment to be given at an earlier stage. This factor was tested, developed and refined as a causal concept through examination of subsequent trials.

improvements to downstream patient health. Importantly the process also allowed a fuller conceptualisation of how diagnostic attributes interrelate to create these changes.

As qualities of a process that interacts with patients, these diagnostic attributes were perceived to act as *mechanisms* through which a test might improve an aspect of patient health. Moreover, treatment effects appeared to be triggered by changes to the workings of these mechanisms as a result of introducing a new or different test. Indeed, the fundamental methodological notion that evidence-based medicine relies on to provide a meaningful measure of clinical effectiveness is that of comparison to a second experimental intervention. So these mechanisms become *comparative changes* rather than fixed properties, whereby introducing a *change* in testing or a *change* in management can produce an improvement or decline in the performance of certain mechanisms, and as a result alter the effectiveness of a diagnostic strategy.

Mechanisms were observed to share common interdependencies, forming causal networks which would then determine how and to what extent to the experimental strategy could lead to differences in patient outcomes. This process was observed to occur along four sequences of interactions between mechanisms, which are illustrated below using examples from the cohort of test-treatment trials:

1. Direct route (direct test effects),
2. Decisional route (altering decisions and actions),
3. Temporality (changing timeframes)
4. Perceptions (influencing patient and clinician perceptions).

Figure 7.3 depicts the final framework schema, illustrating all 14 mechanisms and how they relate to their parent components in the care pathway; Table 7.1 defines all confirmed and newly–identified diagnostic attributes.

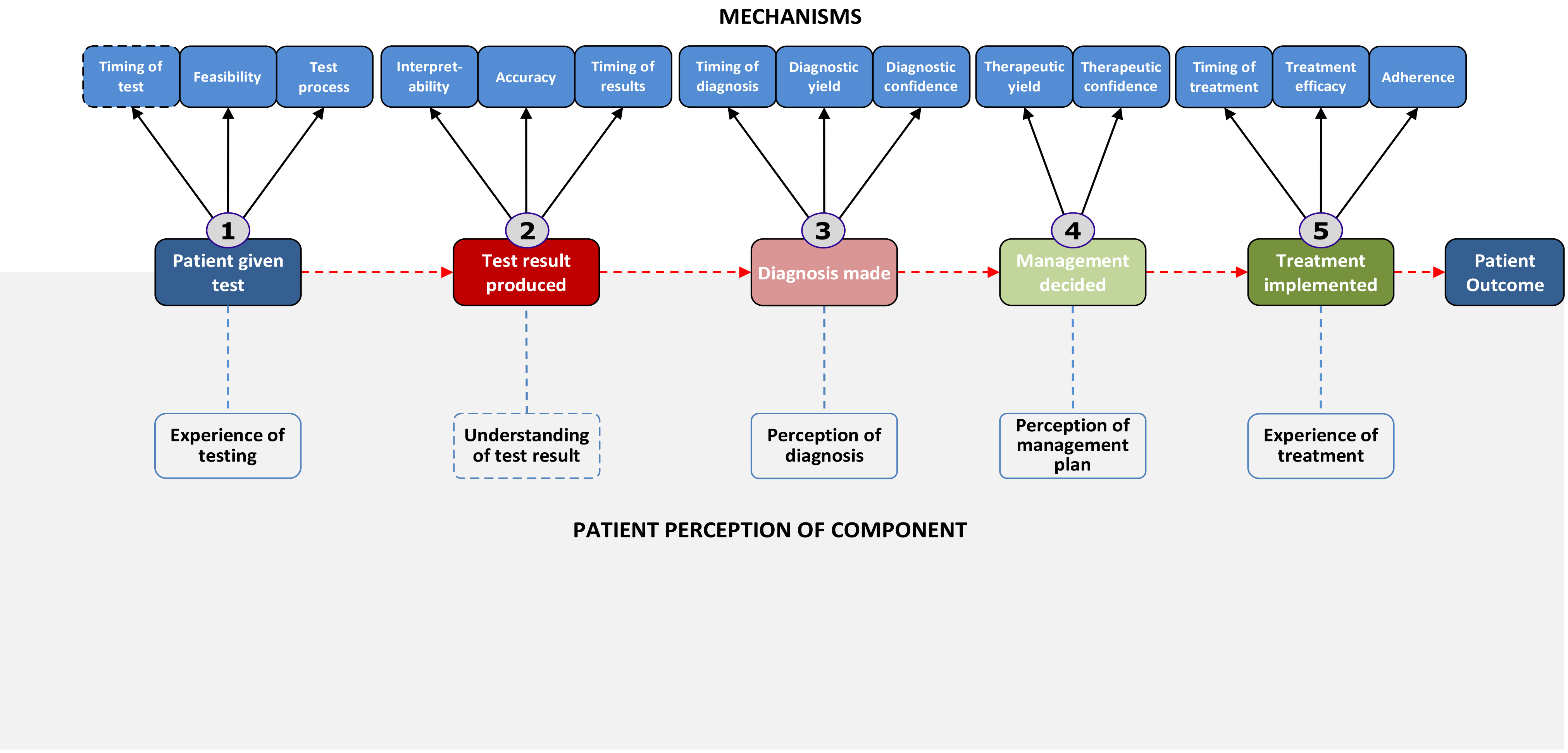


Figure 7.3: Schema illustrating the 14 causal mechanisms as attributes of the test-treat pathway.

Care Pathway Component	Mechanism	Definition
1. Test Delivery	Timing Test	The rapidity of performance of a test within the management strategy.
	Feasibility	Completion of the test process, where reasons for non-completion are: **a)** Counter-indication (Clinician refusal to administer test) **b)** Patient acceptability (Patient refusal to have test) **c)** Technical failure (Ability of diagnostic equipment to produce data)
	Test Process	Patients' interaction with the test procedure, potentially causing physical or psychological harms or benefits.
2. Test Result	Interpretability	After successful completion of the test process, the degree to which test data can be used to inform a diagnostic classification.
	Accuracy	The ability of a test to distinguish between diseased and non-diseased patients.
	Timing Result	The speed with which test results are available.
3. Diagnostic Decision	Timing Diagnosis	The speed with which a diagnostic decision is made.
	Diagnostic Yield	The degree to which the test contributes to a patient diagnosis in any form, including: • Provision of definitive diagnosis • Confirmation of suspected diagnosis • Ruling out a working diagnosis • Distinguishing between alternative diagnoses with different treatment implications Differentiated from 'Accuracy' in that it also incorporates any other information used by a clinician to formulate a diagnostic decision (such as prior test results).
	Diagnostic Confidence	The degree of confidence clinicians have in the validity or applicability of a test result.
4. Treatment Decision	Therapeutic Yield	The degree to which diagnostic decisions influence therapeutic plans.
	Therapeutic Confidence	The certainty with which a clinician pursues a course of treatment.
5. Treatment Implemen–tation	Timing Treatment	The speed with which patients receive treatment.
	Treatment efficacy	The ability of the treatment intervention to improve patient outcomes
	Adherence	The extent to which patients participate in the management plan, as advised by their physician, in order to attain the therapeutic goal.

Table 7.1: Definitions of attributes of each components that may influence the effectiveness of a test-treat strategy

7.4 Elaboration and Illustration

Below, each sequence is illustrated by taking each mechanism defined in the schema and considering it to produce a *difference* between how one test operates compared to another.

7.4.1 Direct Route: impact of the test process

Test Process

As documented in previous research, some test procedures can directly impact on a patient's health independently of subsequent diagnostic or treatment decisions, hence those that offer a reduced procedural-related morbidity will be of immediate benefit to patients. In the new schema, this effect is defined by the 'test process' mechanism and was a common observation. For instance, amongst early breast cancer patients being investigated for metastatic spread, significantly fewer were demonstrated to suffer from postoperative arm swelling, seroma formation, numbness and paresthesia if initially triaged with sentinel lymph node biopsy (SLNB) instead of immediate full axillary lymph node dissection (ALND). This direct physical benefit was due to the considerable proportion of node-negative SLNB patients who were subsequently able to avoid the more invasive ALND(T24) (Figure 7.4).

Tests may also have a direct therapeutic value when the act of experiencing a test can confer immediate health benefits. These psychological effects can be understood as patient perspectives, and are discussed separately below.

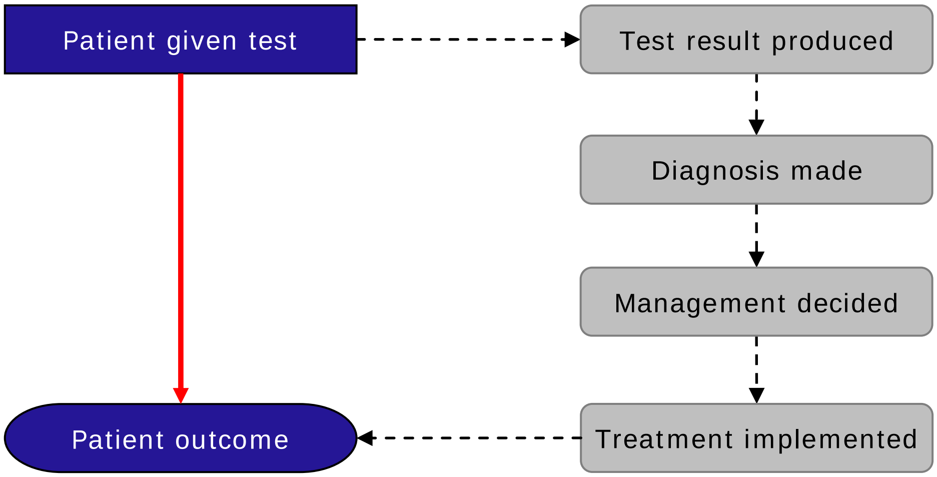


Figure 7.4: Direct impact of diagnostic tests. Attributes of the test process can influence patient outcomes independently of differences in accuracy or diagnostic decision–making.

7.4.2 Decisional Routes: impacts of diagnostic information and decision–making

Feasibility and Interpretability

The contribution that a test's results can make to decision-making is first mediated by its ability to produce a diagnostic output (feasibility) that can be interpreted clearly (interpretability). These issues were rarely addressed by included trials, yet differences in either property could prompt a succession of unfavourable changes to subsequent mechanisms. Feasibility in particular was generally not captured by these trials since the failure of a test to produce results tended to be listed as a reason for patient exclusion(T67), either before(T89) or after randomisation. Nonetheless failed procedures, whether as a result of counter-indication to testing (for example claustrophobia in individuals randomised to receive an MRI(T4) or peripheral arterial disease preventing catheterisation in patients awaiting angiography(T6)) or technical malfunction (for example contaminated blood culture samples(T23)), as well as indeterminate results could invite additional investigations, increase the total diagnostic time or decrease diagnostic and therapeutic yields through incorrect decision-making and poor diagnostic confidence. For example, in a trial evaluating the diagnosis of coronary artery disease (CAD), patients with acute chest pain who were allocated to exercise ECG were significantly more likely to be referred for further investigation (coronary angiography) than patients receiving a stress echocardiogram. They were also significantly more likely to be diagnosed as having intermediate post-test probability of CAD. These findings were caused by the higher frequency of failed procedures and inconclusive results produced by the exercise ECG, whereby patients with uncertain diagnoses were arbitrarily classified as at intermediate risk of CAD(T89).

Test Accuracy, Diagnostic Yield, Therapeutic Yield and Treatment Efficacy

The most widely recognised impact on patient health is a composite of the interaction between the mechanisms of accuracy, diagnostic yield therapeutic yield and treatment efficacy. As posited in the existing literature, diagnostic reclassification afforded by a test with greater accuracy could lead to a change in treatment through better-informed decision-making. A clear example is provided by a trial evaluating whether photodynamic diagnostic cystoscopy (PDD) in addition to standard white light cystoscopy could reduce the risk of bladder cancer recurrence. Due to its enhanced accuracy in the detection of smaller carcinomas and ability to more clearly define tumour borders PDD identified and treated substantially more lesions at initial diagnosis in the experimental group, leading to a significant reduction in the frequency of recurrence(T1) (Figure 7.5). Of course, improvements to accuracy, diagnostic yield and therapeutic yield will not materialise unless the available treatments are efficacious. This may be one reason why a trial evaluating three modes of tilt testing to better diagnose the cause of previously

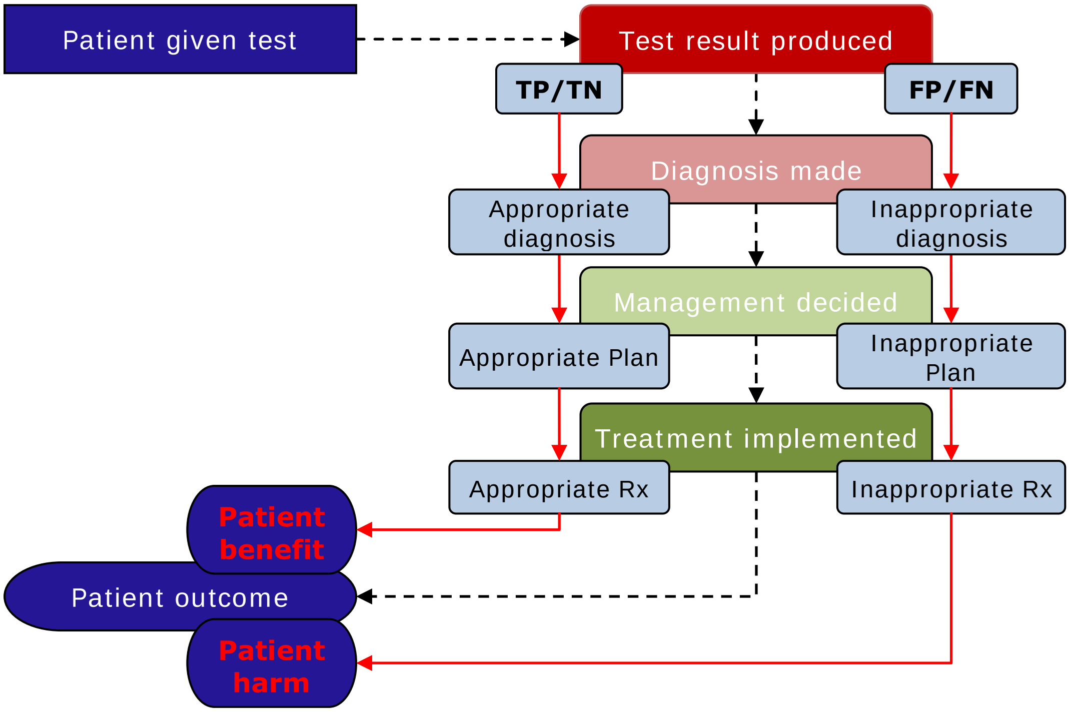


Figure 7.5: Decisional impact of tests. The schematic illustrates how information produced by a test may influence patient outcomes through how it changes diagnostic decisions, treatment decisions and treatment implementation.

unexplained syncope failed to find any difference in the time–to–syncope recurrence(T2). As acknowledged by the authors, there are no effective treatments to eliminate syncope and so, despite wide differences in the accuracy of the three tests, observing an improvement in the primary outcome would have been very unlikely in these patients.

Diagnostic Confidence

While diagnostic yield generally increases with accuracy, it is also independently influenced by a physician's diagnostic confidence. Clinician confidence in test results were observed to contribute to the overall effectiveness of test–treatment strategies in more ways than anticipated by the preliminary model. Tests that induced greater confidence in their results could benefit patients by reducing further investigations (T12), 266 (decreasing

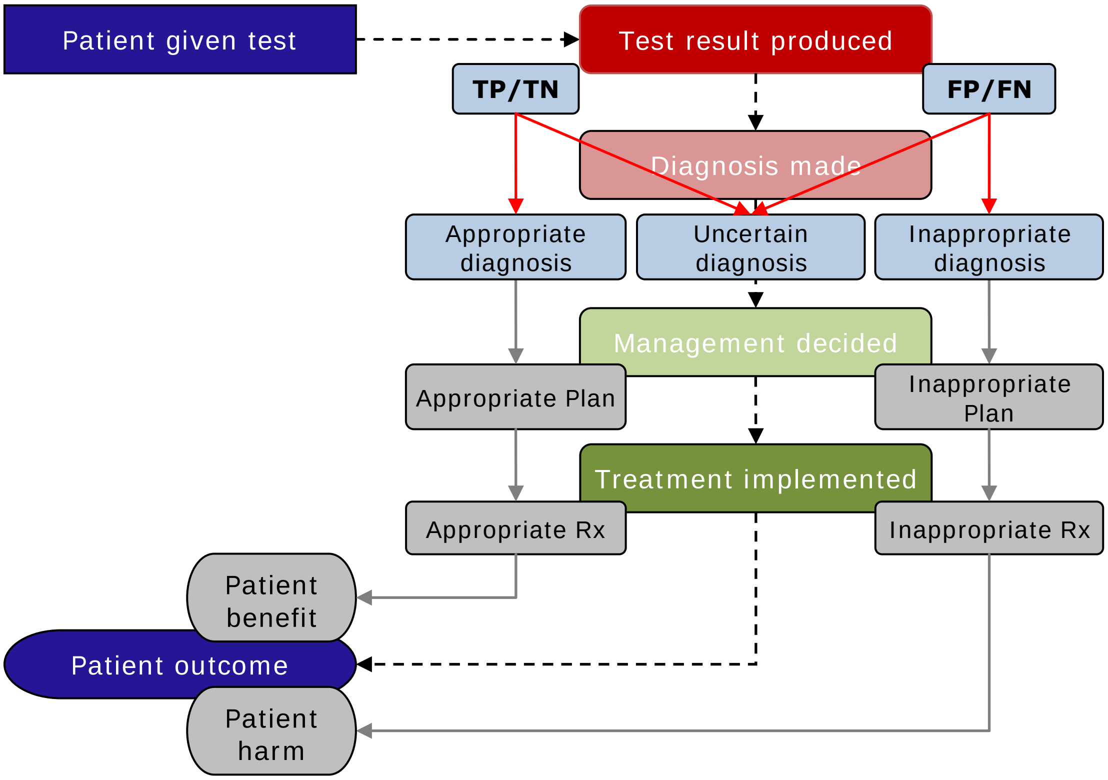


Figure 7.6: Decisional mechanisms may fail to cause a change to patient health. Here a more accurate test is provided (red arrows), however poor confidence in the discriminatory ability of the test may mean the new test is ignored, and consequently the potential gains to health from better decision–making are not realised (grey arrows).

any associated procedural harms) and expediting the time to treatment, though such changes would only benefit patients if the new test is also at least as accurate as the existing test.

A trial evaluating the benefit of adding positron emission topography (PET) to the pre-operative staging of patients with an established diagnosis of non-small cell lung cancer demonstrates how diminished physician confidence in the ability of a test can over-ride the benefits of improved accuracy if test results are ignored[(T52)](Figure 7.6). As remarked by the authors, PET results could have changed downstream management in a quarter of participants through the enhanced detection of mediastinal disease, thereby avoiding the

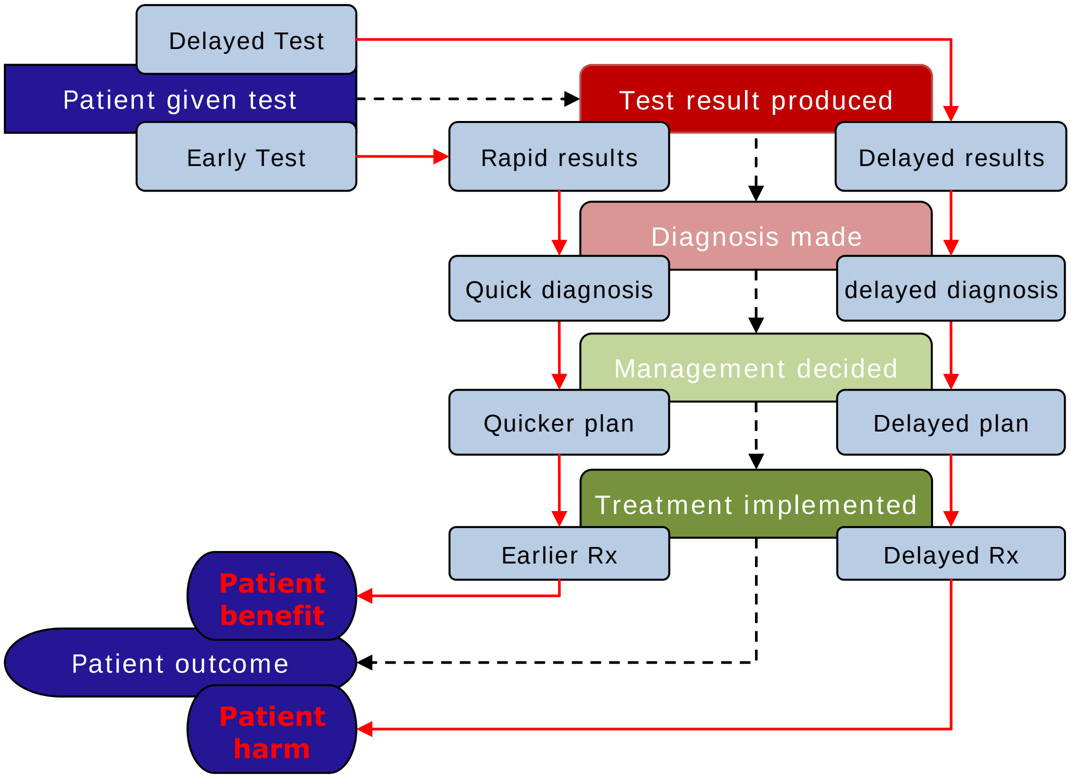


Figure 7.7: Changes in the temporality of the test–treat process impact on health outcomes. Hastening the point of testing, the production of results, or the time of diagnoses can all indirectly improve patient health through producing earlier treatment (red arrows).

need to proceed to thoracotomy in incurable patients. Ultimately no difference was observed in the proportion of patients avoiding a thoracotomy (the primary outcome) as surgeons preferred to confirm PET findings with operative staging, the standard test they were accustomed to using. The observed treatment effect does not likely reflect the real accuracy of PET, but the physician response to the introduction of a 'new' diagnostic technology. Clinicians may not have been confident in its accuracy or, more importantly, in how its results should be interpreted leading to so-called errors of implementation(T52). Had the surgeons trusted the PET results, downstream management could have changed to avoid thoracotomy in the quarter of participants where PET had detected incurable mediastinal disease.

That increased diagnostic confidence does not automatically confer an increase in diagnostic yield is also suggested by the results of some before–after studies, where observed increases in diagnostic confidence do not always translate to changes in diagnostic decision–making[267].

Therapeutic Confidence

Physician confidence in the ensuing success of a treatment plan could affect the success of treatment by influencing the approach to treatment. Therapeutic confidence is most clearly exemplified by tests that directly inform surgery. Digital subtraction angiography (DSA) and multi-detector row computed tomographic angiography (MDR–CTA) are used to determine the location and degree of vascular narrowing in patients with symptomatic hardening of peripheral arteries (atherosclerotic Peripheral Arterial Disease) who have been referred for revascularisation. In this setting, a key determinant of improved patient health may be the confidence with which a surgeon approaches the revascularisation. An RCT evaluating this comparison discovered that physicians using DSA were significantly more confident of plans for surgery as a direct consequence of the test's clearer vascular images, while CTA images were found to obscure interpretation and decrease confidence in the presence of vessel wall calcifications(T12).

7.4.3 Temporality: impacts of timing

Temporality is conceived as a property of each of the five components in the patient care pathway. Differences in the speed with which diagnoses are produced and treatment administered to patients can provoke changes to patient health regardless of changes in decision-making. Strategies that hasten the administration of a test can be of benefit to patients, particularly if complemented by earlier treatment (Figure 7.7). The provision of coronary angiography (CA) on average 57 hours earlier in a patient's management was, for example, found to decrease the combined risk of death, non-fatal cardiac events and rehospitalisation (the primary outcome) when compared to a delayed strategy* in patients with unstable angina and non-ST segment elevated myocardial infarction(T14). As both strategies employ the same diagnostic test there can be little impact from more accurate decision-making, and the key mechanism improving patient outcome is likely to be the more rapid provision of treatment as a consequence of the earlier testing. Reducing the turnaround time in the production of test results can also shorten the time-to-treatment. For example, patients with ventilator-associated pneumonia who received a rapid antimicrobial susceptibility test received definitive reports a mean of 2.8 days earlier than those receiving the standard susceptibility test, and suffered significantly fewer days of fever, bouts of diarrhoea and days on mechanical ventilation(T59).

Either of these two mechanisms mainly influences patient outcomes by triggering an earlier diagnosis and earlier treatment, although of course this relies on test results and ensuing decision-making being deployed equally promptly. Failure to do so can nullify any impact of timing, as was demonstrated in an RCT evaluating the addition of Polymerase Chain Reaction (PCR) to conventional testing for discerning between viral and bacterial aetiologies in lower respiratory tract infection. While PCR results (for the detection of virus) were produced earlier than bacterial culture results, the strategy failed to decrease time-to-treatment (in this case time to the discontinuation or modification of antibiotics) as physicians were unwilling to base treatment decisions solely on PCR, preferring to wait for slower bacterial results as well(T21).

* The control strategy entailed initial medical treatment followed by referral for CA if clinically indicated. Consequently 97% of immediate testing patients received CA in a median of 22 hours, while 51% of delayed strategy patients received CA in a median of 79 hours.

7.4.4 Perceptive dimensions: impacts of the patient and clinician experience

Each mechanism has been described from an objective clinical perspective, however this understanding may differ from the subjective perceptions of individuals involved. Both the patient's perspective, and more recondite aspects of the clinician's perspective, are far less predictable and hence could mediate the effects of the diagnostic strategy in unexpected ways. These unpredictable responses could eliminate potential improvements gained from other mechanisms.

Patients

As a key negotiator of treatment decisions, the patient and their perspective of the healthcare process represents an all-pervasive dimension in the explanatory framework, capable of mitigating the final impact that mechanisms may have on health outcomes. Patients' perceptions of testing, their experience of the testing process and their understanding of the test result may all influence patient outcomes. A large number of studies show social, emotional, cognitive and behavioural effects of testing across a wide range of clinical conditions[264].

Seen from the patient's perspective the degree to which a test succeeds in producing results (its feasibility) relies on their willingness to undergo a procedure. Poor acceptability of the procedure is most likely to affect patient outcomes in multiple-testing situations, where an unpleasant first test could negatively influence patients' willingness to attend follow-up testing or treatment.

The experience of undergoing diagnostic procedures can also influence patients' illness cognitions. For example patients with non–Q–wave myocardial infarction who received immediate angiography demonstrated significant improvements in angina stability, treatment satisfaction and disease perception than those managed with a non–invasive stress test(T20).

Patient perceptions of testing can also impact on downstream measures of health by means of a diagnostic placebo effect, where the impression of a thorough investigation encourages improvements in perceptions of health status. This may account for significant improvements in health utility (EQ-5D) reported by patients with acute undifferentiated chest pain diagnosed in a specialist Chest Pain Unit compared to those diagnosed in emergency departments, despite equivalent treatment and rates of adverse cardiac events(T48). How patients process test results and react to diagnoses can be unexpected and difficult to predict however, and responses are likely to be specific to the tests being used as well as the presenting symptomatology and severity of conditions to be detected or ruled out. Other studies demonstrate the risk of somatic fixation as a result of undergoing testing and receiving a 'diagnosis'. For example acute lower back pain patients receiving an X–ray were found to display worse overall health status 3 weeks later than those receiving only standard consultation(T28), suggesting that being labelled with a clinical diagnosis, albeit one of minor consequence (e.g. age–related degeneration of the spine), may serve to legitimise illness beliefs, rather than reassure as to the absence of severe disease as intended.

Earlier knowledge of one's diagnosis could also have behavioural and health consequences in certain situations. Psychological benefits may occur if serious disease is ruled out more promptly, by dispelling anxiety or providing earlier reassurance. These effects could also be detrimental; confirming the presence of disease may increase anxiety if further investigations are warranted (disease staging), or serve to propel negative behaviours. Although not measured in any of the 103 included test–treatment trial, examples of the psychosocial impacts of earlier testing are seen in screening comparisons[268].

Adherence

Experiences and perceptions of the care pathway can also impact on the patient's willingness or motivation to adhere to medical advice[269–271]. Non-adherence with the agreed treatment can mitigate prospective improvements to health gained from advances in yield, confidence or other clinical mechanisms. Negative perceptions or experiences of testing and diagnosis are responses that could cause patients to lose confidence in the diagnosis or management plan, thereby instigating a reluctance to undergo subsequent testing or planned treatment.

Due to its distance from the diagnostic intervention, adherence was seldom measured or considered in the interpretation of test-treat trials however it is likely that increasing adherence could maximise the potential benefit gained from treatment and preceding improvements to the diagnostic strategy. While enhanced adherence has been positively correlated with improved clinical outcomes[272], there is also an increasing recognition that there are multiple influences on adherence and its relationships with health response are complex. A recent systematic review found, for example, that interventions succeeding in raising adherence did not always improve clinical outcomes[269].

Clinicians

It is of course important to consider the impact of the clinician's perspective, emotional, cognitive, social or behavioural perspectives which are external to objective medical concerns but can nonetheless impact on decision-making. Referring physicians have been shown to modify their prescriptions of downstream management according to the nature of their relationship with their patient, for example to satisfy their patient's expectations of investigation and treatment[273–274], as a response to unstable relationships with their patients[275], or to prevent a perceived threat of malpractice[276–277]. The most common response to such situations is a request for additional diagnostic information, so-called 'defensive medicine' [277], which can serve to raise the diagnostic threshold needed to trigger a change in management. Limited evidence has begun to indicate that certain personality traits may predispose to an overzealous approach to the prescription of antibiotics, such as ‘zeal’ and ‘a readiness to serve’ [275], with indications that this may also serve to increase investigative referrals[278]. If these additional tests are less accurate, then diagnostic and treatment yield will be adversely affected, potentially cancelling out any gains to patient health incurred from preceding mechanisms. If additional tests are more harmful to the patient or lead to considerable delays in treatment, then patient health could be directly harmed. The effect of extra-clinical concerns on clinical behaviour requires more investigation, however at least one study has demonstrated that fear of malpractice is an important predictor for the number of tests ordered by clinicians with those so affected ordering on average 25% more tests than clinicians not so concerned[277].

Complex public health interventions are known to be sensitive to organisational structures[83], and the same is likely to apply to test-treat settings with local differences in the characteristics and prioritisation of health services as well as variations in resource availability and clinical protocols contributing to variations in decision-making and hence patient outcome effectiveness. While a discussion of these influences remains beyond this discussion, their potential impact on patient health through the channelling of resources should be kept in mind.

7.5 Discussion

This study has proposed a framework of thought to explain how diagnostic tests affect patients. Building on existing concepts this analysis has defined the key clinical processes that link tests to patient health, and revealed it to be composed of a complex network of mechanisms through which changing a test might create differences in downstream patient outcomes.

The framework uses the concept of the 'patient care pathway' to make explicit all the ways in which health outcome differences might be generated by the introduction of a new testing strategy. As the patient moves through the test-treatment process, each component can influence the patient's health trajectory through several mechanisms, as well as through the patient's perception and experience of that component. Each mechanism may independently trigger improvements to patient health either directly, or indirectly by interacting with other mechanisms as shown in the four sequences above. These sequences may be considered as causal pathways of effect, where the potential advantages offered by improvement to one mechanism demonstrably fail to improve patient health if an interrelating mechanism produces a contrary effect to that desired.

Founded on a wide body of existing conceptual research, it is no surprise that many of the ideas contained in this framework are not new. The founding principle that tests and outcomes are indirectly related, and likely to be mediated through intervening treatment and other clinical processes, is commonly–held and indeed the 'patient care pathway' structure was directly informed by the previous work of Patrick Bossuyt and Jeroen Lijmer[193]. Driven by accuracy, the role of diagnostic decision–making as a key factor mediating the clinical utility of tests is central to current thought, as is the notion that differential adherence to management will mitigate the end success of a testing pathway[19,77]. That tests will have direct effects was similarly widely acknowledged, while more recently the psycosocial context of testing and its consequences has been characterised[264].

Nonetheless, this schema does differ considerably from previous frameworks in three key respects: the number of ways in which tests influence downstream patient health, the complexity of how this is achieved, and finally the condition that superior accuracy is a necessary precondition for clinical utility.

7.5.1 More numerous mechanisms

The examination of test-treatment RCTs highlighted that diagnostic tests are capable of influencing patient health in ways more numerous and complex than currently conceived. With a total of 14 individual mechanisms, the framework identifies as least seven additional attributes of the test-treat process that are rarely considered to influence clinical effectiveness. Not only may health be influenced indirectly by accuracy-driven changes to patient management or directly by the test apparatus, but elements of decision-making behaviour and other technological properties of tests can modify the extent to which this occurs. Tests that are more often feasible, produce results that are more easily interpretable or that engender higher therapeutic confidence could each result in more effective healthcare. The most enlightening contributions are the four timing mechanisms, which although common considerations in the screening literature[23] were not considered by any of the reviewed frameworks.

7.5.2 More complex interactions

Perhaps the most important contribution to existing thought is the elaboration of how these mechanisms interact to create (or prevent) differences in treatment effect. Previous frameworks conceived the diagnostic determinants of health outcomes to be hierarchically ordered, such that improvements in mechanisms earlier in the test–treat process are necessary but not sufficient to deliver a favourable downstream outcome. This linear approach does not allow for any downstream benefits to occur if earlier diagnostic attributes fail to be improved. Conversely, the new framework highlights the importance of considering *all potential differences* in competing test-treatment processes. Within the schema each mechanism can trigger improvements to patient health either directly or indirectly by acting on other mechanisms with which it is commonly interdependent. In this way the new framework is more holistic and less linear than preceding structures. Not only does it encompass a greater number of processes and mechanisms, but it emphasises that these processes do not occur in series as is often conceived, but rather as a set of interrelated mechanisms governed by the specific clinical setting. Moreover, rather than focussing on the singularity of changing decision–making behaviour as a vehicle for improving health outcomes effects are created along four causal pathways which can act in parallel in any single comparison to influence patient health.

Of equal importance to explaining how tests might influence health outcomes is the exposition of why intended effects might not materialise. Existing theory acknowledges that this can occur. Some authors noted that more accurate tests would not always produce benefits in decision-making or patient outcome[14,44–46], however none extrapolated *why* this might occur, beyond the difficulties of conducting and interpreting test-treatment RCTs. By reconceptualising test-treatment cause and effect as causal pathways of mechanisms, the new structure has been able to demonstrate how downstream health benefits are only realised if the potential advantages of a mechanism are not nullified by inadequacies in ensuing mechanisms. Namely, these causal pathways can break down and fail to realise intended improvements if at least one mechanism in the causal chain produces a contrary effect. That PCR failed to reduce time to treatment[(T21)] is illustrative of the role that successive mechanisms may play in mediating a test's impact, since potential improvements in patient health derived from earlier targeted treatment were cancelled out by a decreased confidence in the diagnostic capacity of the new test.

7.5.3 From accuracy−driven change to complex system interactions

The review revealed that existing notions regarding the clinical utility of tests have traditionally placed accuracy as the central linchpin of clinical effectiveness[19,21,44–46,49–50,77,91,193,256]. This stems from the premise that if patients are treated appropriately they will have better outcomes[59], and so the main value of tests is seen as the mediating role they have in guiding treatment decisions[193]. A caveat of existing frameworks is that whilst superior accuracy is not sufficient for clinical effectiveness, it is nonetheless necessary.

There can be no doubt that diagnostic accuracy is a key driver of clinical effectiveness, however this analysis of test-treatment trials highlights that it can no longer be considered the *sine qua non* of diagnostic performance. Since differences in patient health are not solely driven by accuracy but also created through improving the delivery of non-decisional aspects of care, superior diagnostic accuracy is not necessary for a test to demonstrate superior clinical utility. A diagnostic strategy no different in its ability to influence decision-making, yet which enables patients to be diagnosed and treated more promptly would be unfairly represented in its potential to improve patient health by looking only at its accuracy. Nor do improvements to patient health necessarily follow improved accuracy, given that several other decisional mechanisms are also accountable for any downstream effects. It is only by looking at the whole picture that we can begin to decide which outcomes, surrogate or otherwise, will best capture the true impacts on patient health.

7.5.4 Study limitations

This study has applied the principle of analytic induction to identify the key mechanisms by which test-treatment strategies change patient health, using a cohort of test-treatment RCTs to elaborate the model. It is however possible that future studies or missed examples may reveal additional mechanisms that are highly specific to diagnostic settings not represented in this data set. Therein lies the main drawback of the current method; since theory was developed through the use of disconfirmation, technically no theoretical construct can ever be considered ‘final’ since the existence of other deviant cases not captured by the study cannot be ruled out.

Equally, the focus on RCTs may have excluded empirical evidence of other ways in which tests can impact on patients. As an experimental study design it is less suited to evaluating more subjective aspects of clinical management for example. Nonetheless, RCTs encompass the whole test-treatment process and so have presented a useful starting point for developing this framework.

There are also certain diagnostic intricacies that have necessarily been simplified in order to produce a generalisable model of the test-treat process. Most relevant is the nature of the diagnostic decision, which in practice is generally not the simple dichotomy alluded to here but is multiplicitous; a diagnostic test is not only used to rule a condition in or out, but to discern between multiple differential diagnoses. By focusing on the target condition the framework may fail to identify the impact a new test has on the downstream management of test-negative patients, such as receiving an immediate alternative diagnosis or being channelled more effectively towards other diagnostic tools to evaluate their condition further. Similarly, the framework does not specifically address how the test-treat process can react to incidental diagnostic findings, although by their nature these likely affect a small proportion of patient management. In an attempt to highlight the importance of the patient and clinician perspective, it is also recognised that the current approach is limited to that expounded in existing research as well as the author’s individual academic and clinical perspective. Consequently there may well remain aspects of the patient perspective, and diverse clinical viewpoints, that the author has failed to distinguish as important modifiers of downstream health.

7.6 Conclusions

This analysis highlights the multitude and complexity of interactions occurring within test-treatment interventions. It is presented as a preliminary framework, open to deliberation, modification and further development as methodological research into patient outcome efficacy progresses.

Although theory is crucial to an understanding of how tests change patient health, any theory is only truly useful if it can be used to underpin and progress empirical evaluation. In this regard an important advantage posed by the new framework is that it addresses a fundamental problem of test-treatment trials: that of our ability to disaggregate effects of the test from effects of the treatment. The following chapter considers the specifics of how the conceptual structure may underpin and progress the design and interpretation of test-treatment RCTs.

This chapter has been published in part:

Ferrante di Ruffano L, Hyde CJ, McCaffery KJ, Bossuyt PMM, Deeks JJ. Assessing the value of diagnostic tests – A framework for designing and evaluating trials. BMJ 2012:344:e686.

8

Towards a full evaluation of test-treatment interventions:

Developing a method to select outcomes

In order to fully evaluate test-treatment interventions it is necessary to identify all the ways in which a new strategy might impact on patient health. The theory to underpin this rationale was developed in Chapter 7, yet it suggests that elucidating all likely processes of change is likely to be a highly complex endeavour for which there is currently no guidance. Chapter 8 was designed to address this deficit by developing the framework into a practical tool. The tool is presented as a checklist with an accompanying graphic, and its value to the design, interpretation and appraisal of test-treatment RCTs is illustrated by worked examples, derived from the project cohort.

The preceding chapter developed a theoretical framework that sets out the common mechanisms by which diagnostic tests, and their subsequent actions, can change patient health. During the detailed examination of all included trials undertaken to build this generalised framework, the author's experience was that very few reports provided useful accounts of how the experimental strategy was expected to change patient health. This severely curtails the ability to appraise whether trialists have selected all important outcomes. Critically, if this lack of reporting reflects an incomplete deliberation of how test-treatment interventions create change, then these trials strongly risk having conducted incomplete evaluations of effect.

The importance of developing a clear scientific rationale is a well established tenet of study design[121]. The principle is given particular emphasis in evaluations of complex interventions, where it is recognised that their multiple and interacting components engender a wider range of effects than expected from 'simple' interventions[83]. Developing a 'coherent theoretical basis' is therefore considered essential to capture and interpret these effects, however there is currently little guidance on how to formulate this rationale, particularly for test-treatment interventions. Chapter 7 suggests that this is likely to be a difficult task to perform for test-treatment interventions; the framework highlights the need to consider multiple causal pathways made up of complex interactions in order to ensure that all possible effects have been identified. However, since we now have a generalised concept that models the relationship between tests and outcomes, the framework could be harnessed to generate a theory of expected change for any given test–treatment strategy under evaluation.

Accordingly, this chapter was developed to address the shortfall in existing guidance by designing a standardised tool that researchers can use to discern the scientific rationale of how test-treatment strategies cause change. This chapter describes how the framework was developed into a tool, presents it as a preliminary version, and discusses its value to underpinning reliable evaluations of tests by reference to examples. The chapter concludes that its use could benefit four aspects of test-treatment evaluation: designing trials, establishing the need for a trial, interpreting trial results and appraising trial quality.

8.1 Method: developing the tool

The author took the central premise that in order to be useful the tool should enable users to conceptualise their test-treatment comparison clearly, to think through each framework mechanism, and to iterate how all relevant mechanisms might interact to cause change to patient health. Furthermore, to be comprehensive and reliable the tool should achieve this in a structured way.

The author's experience of appraising test-treatment RCTs and developing the theoretical framework highlighted potential approaches that could satisfy these requirements. As part of the deviant case analysis each test-treatment trial was examined by mapping out its central processes in what was essentially a causal diagram. This lent great clarity to the author's understanding of test-treatment comparisons, that not only tended to be poorly described but were also often highly complicated. Causal diagrams provide a simple means of summarising and ordering complicated information, particularly when its components are likely to interact both independently and synergistically – as in complex interventions[279]. In epidemiology they have been used for numerous ends, including to conceptualise measurement bias[280], to model for possible confounding and other forms of bias in retrospective observational studies[281] and to model links between cause and effect[282]. Within the context of test evaluation research, causal diagrams have also been found useful for defining the scope of evidence–synthesis reviews into the effectiveness of screening tests[78].

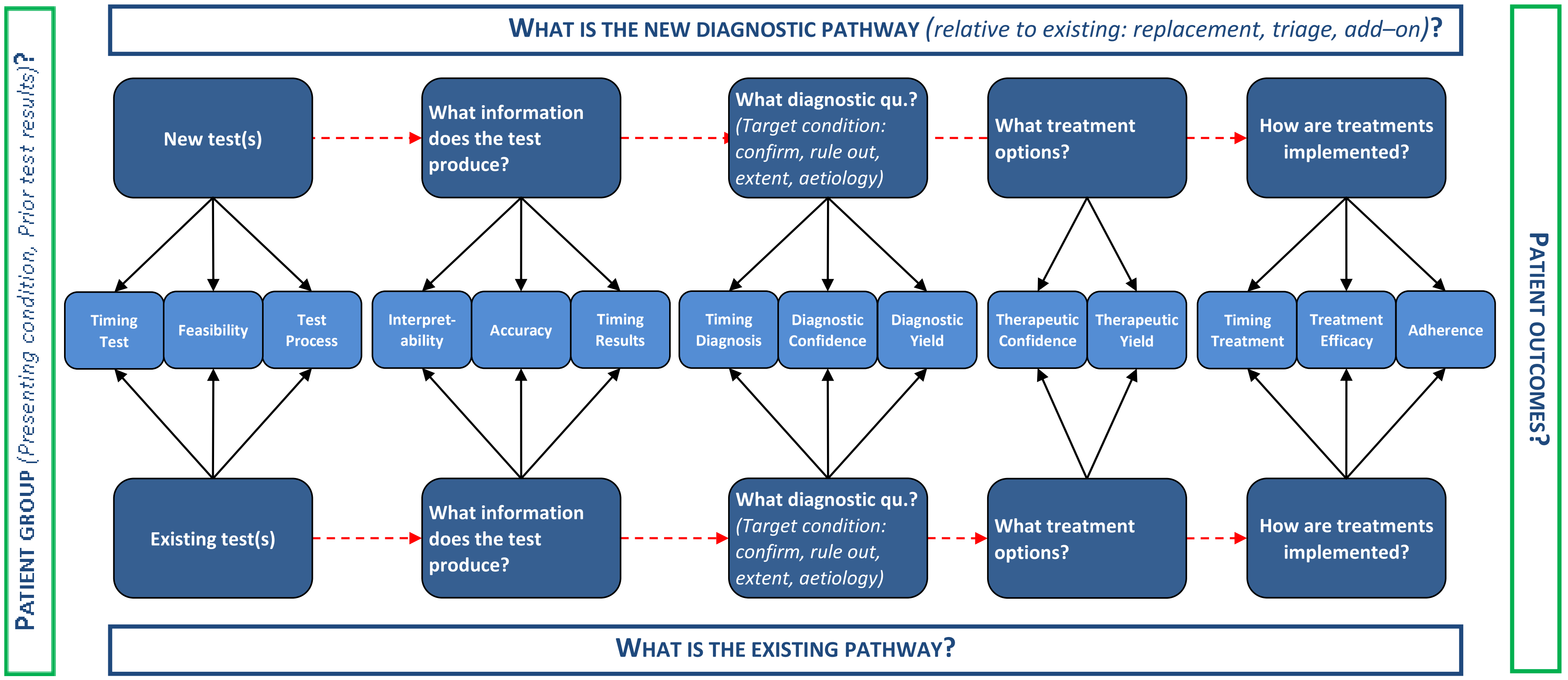
A common approach for achieving a structured consideration of methodological issues in evidence–based medicine has been the application of checklists. Used to appraise the quality of primary studies[71–72,101,283] and evidence–syntheses[284], checklists have proven to be effective tools for systematising the design and appraisal of epidemiological studies since they promote a standardised approach whilst ensuring that important issues or features are not missed. These qualities are particularly important to the task in hand due to the complexity of interactions expected to occur along the test–treat pathway.

For these reasons the theoretical framework was developed into two practical elements, a diagram and a checklist.

8.1.1 Graphic schema

The first suggested task is the completion of a diagram that requires the user to map out important features of the diagnostic comparison, specifying what is being done, to whom and when (Figure 8.1). The schema depicts two comparative ‘patient care pathways’, the components and mechanisms of each illustrated in direct opposition to one another. Each set of five pathway components must be defined, using the questions provided to guide the insertion of key information.

Figure 8.1: Graphic schema to map out comparative interventions.


PATIENT GROUP (Presenting condition, Prior test results)?
WHAT IS THE NEW DIAGNOSTIC PATHWAY (relative to existing: replacement, triage, add–on)?
New test(s)
What information does the test produce?
What diagnostic qu.? (Target condition: confirm, rule out, extent, aetiology)
What treatment options?
How are treatments implemented?
Timing Test
Feasibility
Test Process
Interpret-ability
Accuracy
Timing Results
Timing Diagnosis
Diagnostic Confidence
Diagnostic Yield
Therapeutic Confidence
Therapeutic Yield
Timing Treatment
Treatment Efficacy
Adherence
Existing test(s)
What information does the test produce?
What diagnostic qu.? (Target condition: confirm, rule out, extent, aetiology)
What treatment options?
How are treatments implemented?
WHAT IS THE EXISTING PATHWAY?
PATIENT OUTCOMES?

Also selected for inclusion in the diagram were two aspects of the healthcare setting that, during the process of refining the conceptual model using the cohort of published trials, were considered by the author to strongly influence the care–pathway: the type of comparison (triage, add–on, replacement) and the patient group.

By encouraging the user to think about how each sequential component of the new pathway could be beneficial (relative to the existing strategy), this process is designed to prompt an initial consideration of which mechanisms might be relevant to the comparison.

8.1.2 Checklist

The second element is a checklist of questions designed to give more targeted thought as to which mechanisms might influence patient health in the comparison (Table 8.1). Each item asks the user whether there is likely to be a difference in how one mechanism operates between the two strategies. Since patient experience and perspective relates to the whole pathway, several additional questions have been introduced to try to capture points at which these might also be of influence.

If the mechanism is not considered relevant, the item is assigned a negative ('no'). Questions are given positive answers ('yes') if primary evidence that the statement is true already exists, while queries ('?') are denoted when the item may be true but no primary evidence exists to confirm this supposition. The rationale for how the new strategy can be expected to change patient health is guided by selecting all mechanisms whose questions have received positive or unknown answers, after which a consideration of the four causal pathways (direct, decisional, temporal, perceptual) can be used to refine the process of thought.

Use of this tool is illustrated below by working through a test-treatment comparison contained in the project cohort.

Table 8.1 Checklist derived from the thesis framework *(continued overleaf).*

Might there be an important difference between the existing and new test strategies in:					**Y/N/?**	**Notes**	**Outcome to capture difference**
Test Delivery	Timing of test:	a.	Time to test delivery?	Do the diagnostic strategies administer testing within comparable timeframes, e.g. does the new strategy administer a diagnostic test considerably earlier than its comparator?			
	Feasibility:	b.	Acceptability?	Is one test likely to be more/less acceptable to patients than the other test, e.g. does one test carry a significantly increased risk of harm?			
		c.	Clinical contra-indications?	Is one test likely to be suitable to different proportions of the relevant patient group, e.g. might one test be contraindicated in additional/fewer patients?			
		d.	Technical failure rates?	Do the two test processes produce different proportions of failed procedures, e.g. does the process of one test tend to fail more frequently than the other?			
	Test Process:	e.	Procedural harms or benefits?	Do the two tests differ in how they affect patients during their application both physically or psychologically, e.g. is one test more intrusive than the other, does one test have a higher procedural-related morbidity than the other?			
		f.	Placebo effect?	Could one diagnostic strategy give patients a different perspective on being investigated than the other, e.g. might one test give greater encouragement to patients as to the thoroughness of their investigation?			
Test Result	Interpretability:	g.	Ease of interpretation?	Do the two test processes produce different frequencies of clearly interpretable test results, e.g. once the test has been completed successfully, does one test tend to produce a higher frequency of indeterminate or unreadable results?			
	Accuracy:	h.	Accuracy?	Do the tests correctly identify the target condition in different patients, e.g. does one test have a proven or hypothesised ability to identify a higher proportion of diseased &/or non-diseased patients than the other?			
	Timing of results:	i.	Time to produce a result?	Does the speed with which test data are processed differ between tests, e.g. is the turn-around-time between administration of test and production of results considerably different between tests?			

Table 8.1 Checklist derived from the thesis framework *(cont.)*.

	Might there be an important difference between the existing and new test strategies in:				Y/N/?	Notes	Outcome to capture difference
Diagnostic Decision	Timing of diagnosis:	j.	Speed of diagnosis?	Do the diagnostic strategies produce diagnoses in comparable timeframes, e.g. do patients managed in one strategy receive a diagnosis more quickly than the other?			
	Diagnostic yield:	k.	Diagnoses made?	Do the tests contribute to patient diagnosis to differing degrees, e.g. do the results of one test tend to be given more weight than the other?			
	Diagnostic confidence:	l.	Clinician confidence in the diagnosis?	Does the degree of confidence clinicians have in the validity or applicability of a test result vary with that of its comparator, e.g. does one test provide greater reassurance to physicians, or are its results considered less reliable by physicians?			
Treatment Decision		m.	Patient confidence in the diagnosis?	Is the degree of confidence a patient has in the diagnostic process, or the diagnosis itself, likely to vary between strategies, e.g. does one test provide greater reassurance to patients through increased physician confidence, better testing experience or clearer understanding of test results?			
	Therapeutic yield:	n.	Treatment choices?	Do the comparative tests contribute to the formulation of a management plan to similar degrees, e.g. does one test lead to more patients receiving appropriate treatment than another?			
	Therapeutic confidence:	o.	Confidence in the treatment choice?	Do clinicians have similar confidence in pursuing a treatment plan between intervention arms, e.g. does a test improve treatment success?			
Treatment Implementation		p.	Patient confidence in the treatment choice?	Do patients have diverging confidence in treatment plans due to diagnostic testing, e.g. does the new test encourage patient understanding of the choice in management?			
	Timing of treatment	q.	Time to treatment?	Do the diagnostic strategies lead patients to receive treatment within comparable timeframes, e.g. do patients who receive one test receive treatment earlier than their comparators?			
	Treatment efficacy	r.	Efficacy of treatment?	Does the use of the intervention in those identified as diseased by the more sensitive test lead to improvements in patient outcomes? Or does the avoidance of the intervention in those identified as non-diseased by the more specific test lead to reductions in adverse effects?			
	Adherence:	s.	Adherence to treatment	Are patients more likely to adhere to treatment plans in one arm compared to the other, e.g. does one strategy lead to more refusals of treatment?			

8.2 Worked example: management of suspected acute appendicitis

In order to provide as clear an example as possible, a trial was selected primarily for the simplicity of its comparison; namely that it compared two strategies containing a restricted number of tests and a narrow array of treatment choices for a target condition with a well–understood natural history.

The selected study set out to evaluate the benefits of two diagnostic strategies for confirming or ruling out suspected acute appendicitis in adults presenting to the emergency department with right lower quadrant abdominal pain(T60). By comparison to the routine approach of scanning all such patients with CT, the new strategy aimed to be more selective by ordering CT only upon the presence of specific signs and symptoms; hence this can be considered as a triage comparison.

In order to develop the rationale for how the new selective strategy could influence patient health, the first step is to map out the alternative diagnostic and management pathways that were compared in the trial. The standard diagnostic approach was reported to comprise the following elements:

1) Patients were evaluated by the emergency physician and consulting surgeon,
2) Routine laboratory tests were ordered for all patients (bloods, urinalysis, pregnancy test in women, serum chemistry),
3) Other consultations or tests were procured at the discretion of the treating clinician,
4) Abdominal contrast–enhanced CT scan administered and interpreted immediately for all patients.

The new strategy was designed to allow clinicians to be more selective in which patients receive contrast–enhanced CT imaging, thus after following the first three steps clinicians could choose whether to order a CT scan, on the basis of all prior diagnostic information.

In both arms, patients with confirmed appendicitis were treated with either surgical removal of the appendix, or in rare instances antibiotics. Patients without worrying indications (such as non–specific abdominal pain) were discharged, whilst those with alternative diagnoses were treated according to the disease suspected by the diagnostic strategy. Unfortunately, as was common for included trials (chapter 5 p. 138–142), the investigators failed to report clear details of diagnostic or therapeutic decision–making, and also provided no indication of which additional consultations or tests (step 3 above) could be used by clinicians in either arm; thus certain assumptions had to be made in order to proceed with the comparison. These were discussed with a practising general surgeon to ensure their clinical validity:

- Additional tests were likely to be one or more of: expectant observation, ultrasound, X–ray
- Clinicians following the selective strategy would order CT for patients with atypical clinical presentations.
- Patients with clinical signs and symptoms considered to be typical of appendicitis would therefore not receive CT before surgery
- The ‘immediate interpretation’ of CT mandated in the control strategy (step 4 above) was also implemented for experimental patients receiving a CT scan.

It is worth noting that this trial was conducted in the United States (US) during the early 2000s (possibly earlier, the recruitment dates are not reported), so although there are notable contrasts to the style of clinical management in the UK (for example CT is not routine for suspected appendicitis in the UK), the rationale below has been developed to reflect the setting in which the trial was undertaken.

8.2.1 Developing a rationale for how selective CT may improve patient health

Figure 8.2 maps out this comparison using a completed care–pathway schema. Guided by the checklist, the next step is to note where these components might differ so as to identify the mechanisms likely to be driving change to patient health. This process was again discussed with the same general surgeon in order to confirm clinical details and ensure that causal hypotheses are clinically valid.

Timing of test (a. Time to test delivery)

The nature of the comparison could create a differential in the time to test delivery. While CT is mandatory in control patients, clinicians managing experimental arm patients must evaluate all prior test results and information before deciding whether a CT is necessary. Within the context of a busy emergency department this could result in a delay to receiving CT for experimental arm patients with atypical clinical signs and symptoms.

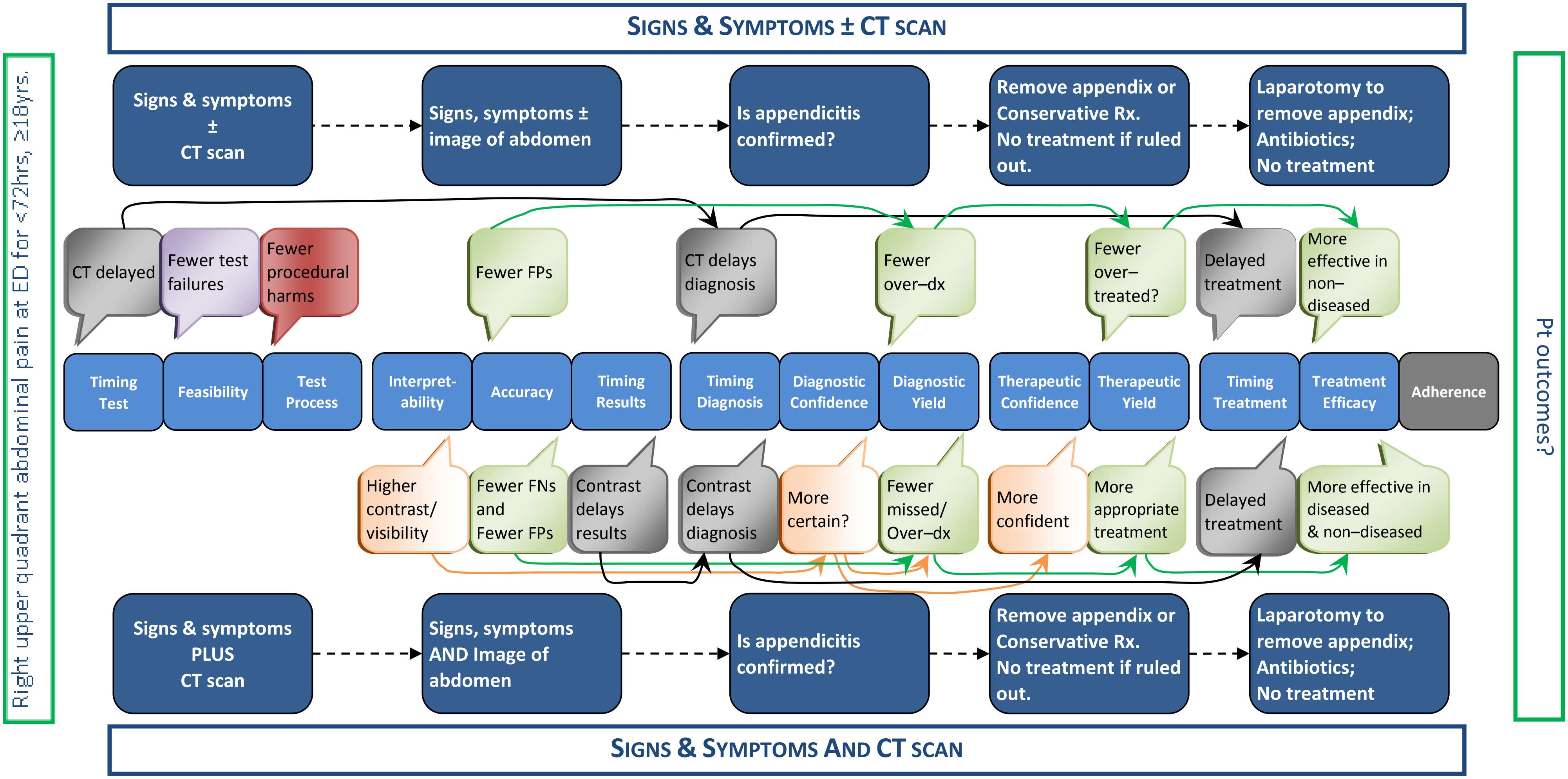
Feasibility (b. Acceptability; c. Clinical contraindication; d. Technical failure rates)

Although the consumption or injection of contrast can be unpleasant for patients, any decreased patient acceptability associated with undergoing CT would be unlikely to result in patients refusing this test.

There are several contraindications to high doses of radiation or the use of contrast agents, however in this example all such patients were excluded (pregnancy, renal insufficiency ascertained by high serum creatinine level, or history of contrast allergy) and so this item no longer needs to be considered for the comparison.

Figure 8.2: Completed schema illustrating a triage test–treatment comparison for confirming suspected appendicitis in an unselected population of emergency department attendees presenting with right lower quadrant abdominal pain[(T60)].

FPs – false positives; FNs – false negatives; over–dx – over–diagnosed


Right upper quadrant abdominal pain at ED for <72hrs, ≥18yrs.
Signs & Symptoms ± CT scan
Signs & symptoms ± CT scan
Signs, symptoms ± image of abdomen
Is appendicitis confirmed?
Remove appendix or Conservative Rx. No treatment if ruled out.
Laparotomy to remove appendix; Antibiotics; No treatment
CT delayed
Fewer test failures
Fewer procedural harms
Fewer FPs
CT delays diagnosis
Fewer over–dx
Fewer over–treated?
Delayed treatment
More effective in non–diseased
Timing Test
Feasibility
Test Process
Interpret-ability
Accuracy
Timing Results
Timing Diagnosis
Diagnostic Confidence
Diagnostic Yield
Therapeutic Confidence
Therapeutic Yield
Timing Treatment
Treatment Efficacy
Adherence
Higher contrast/ visibility
Fewer FNs and Fewer FPs
Contrast delays results
Contrast delays diagnosis
More certain?
Fewer missed/ Over–dx
More confident
More appropriate treatment
Delayed treatment
More effective in diseased & non–diseased
Signs & symptoms PLUS CT scan
Signs, symptoms AND Image of abdomen
Is appendicitis confirmed?
Remove appendix or Conservative Rx. No treatment if ruled out.
Laparotomy to remove appendix; Antibiotics; No treatment
Signs & Symptoms And CT scan
Pt outcomes?

Nonetheless, oral contrast is still likely to be tolerated poorly by patients with usual symptoms of appendicitis (abdominal pain and vomiting), and as a consequence its use may be associated with higher rates of failed procedures where the contrast fails to reach the appendix[285–286]. This could lead to referral for additional tests. Assuming that fewer patients undergo CT in the experimental strategy, the technical failure rates might also be lower in this group.

Test Process (e. Procedural harms/benefits; f. Placebo effect)

Contrast–enhanced CT is more intrusive than any of the alternative imaging tests likely to be used in this setting. If prior test results and alternative non–invasive tests can obviate the need for CT, fewer patients in the experimental arm will be exposed to the low but serious risks of allergic reaction to contrast media[287] and radiation exposure[288].

Since the diagnostic setting seeks to detect and treat an acute condition, patients' perspectives of the thoroughness of their investigations are unlikely to be relevant.

Interpretability (g. Ease of interpretation)

The effects of interpretability are relevant, though their direction of influence is unclear. On the one hand CT produces high–contrast images of all internal structures and so is more likely to produce definitive indications of disease than either the use of physical examination alone, or in conjunction with other modalities that tend to produce more opaque images (e.g. ultrasound or X–ray). As a consequence more indeterminate test results might be expected in the experimental arm, leading to more frequent referral for additional diagnostic procedures, but also potentially leading to different treatment decisions[289]. On the other hand these differences could be reduced if problems with contrast absorption reduce the clarity of CT images and also result in the production of indeterminate scans.

Accuracy (h. Accuracy)

Differences in the accuracy of decision–making between the two strategies are likely to drive changes to the health of patients with suspected appendicitis, although the existing evidence–base is indirect and contradictory making the magnitude and direction of such effects uncertain. Primary evaluations of diagnostic accuracy vary widely in the CT technique being evaluated (contrast techniques and scanner technology), the expertise of diagnosticians (use of surgeons or general emergency physicians), the target patient population (degree of selection on the basis of clinical signs and symptoms) as well as the study design used[290]. Critically, most do not directly compare the performance of contrast–enhanced CT with clinical judgement alone[285,291–295], while one that does has produced tentative evidence that clinical judgement may be at least as accurate as treatment decisions based also on CT imaged[296].

Certain studies show CT to be more sensitive than other imaging modalities, for example ultrasound[297]. If CT is also more sensitive than clinical judgement alone for typical presentations, we would expect fewer true cases of appendicitis to be detected by the experimental strategy. CT could be more accurate in the identification of more serious differential diagnoses, such as bowel obstruction or diverticulitis[298]; if these individuals are mistaken for typical appendicitis in the experimental arm they would not be referred for CT and hence constitute missed true diagnoses, potentially with more severe consequences.

If CT proves to be more specific than the alternative approach[297], we would expect fewer false–positive diagnoses in the control arm. In contrast, other studies suggest that CT may detect milder and more incipient inflammation of the appendix[299] which can overestimate the severity of findings[298] leading to more false–positive detections in the control arm.

Timing of results (i. Time to produce a result)

Oral contrast requires time to reach the lower bowel[285], hence control patients may experience a delay in the production of test results when compared to experimental patients who are spared a CT. Conversely, experimental patients with atypical indications who are delayed in undergoing CT, as it is no longer a routine procedure, may experience a delay in the procurement of test results when compared to controls.

Timing of diagnosis (j. Speed of diagnosis)

Differences in the timing of test results could translate into delays for the production of diagnoses, as iterated for item **(i.)** above.

Diagnostic yield (k. Diagnoses made)

Differences in the accuracy of the two strategies should impact on the diagnoses made for patients: if CT is less specific, more experimental patients could receive negative diagnoses, of which a higher proportion would be false–negatives (particularly those with more serious conditions) and true-negatives; though if CT is more specific then the control arm may demonstrate more equivalence in the number of negative diagnoses, of which a higher proportion would be true–negatives.

Potential increases in diagnostic yield afforded by the higher sensitivity of routine CT could be reduced if any additional tests ordered as a result of technical failures (d.) or indeterminate scans (g.) are less accurate than CT. However this would also depend on the frequency with which this occurs in the experimental arm.

Diagnostic confidence (l. Clinician confidence; m. Patient confidence)

Diagnostic certainty is key to this comparison. CT has been in use for a long time and is the current standard investigation for emergency patients with acute abdomen in the US (the study country)[(T60)]; clinicians are therefore likely to find CT useful in reducing clinical uncertainty. Physicians operating within the selective strategy may lack the confidence to make treatment decisions solely by clinical indication, fearing the consequences of missed diagnoses and unnecessary surgery, which can result in malpractice claims[300]. If confidence is reduced substantially, additional tests may be ordered for patients who – according to the protocol – should not receive further testing. If CT is relied upon, all potential direct and indirect benefits of selective imaging will effectively be reduced to zero. If other less accurate tests are used then diagnostic yield could be reduced.

Again, patient confidence in the diagnosis is unlikely to be relevant in this comparison due to the acute and emergent nature of the clinical setting.

Therapeutic yield (n. Treatment choices)

Differences in diagnostic yield should translate into differences in treatment selection, though again the direction of effect is uncertain and depends on the true accuracy of CT in this population. Fewer experimental patients may be referred for treatment, of which a higher proportion would in reality need treatment (false–negatives).If less specific, more control patients would receive inappropriate referrals for treatment due to the more frequent detection of mild inflammation by CT; though if more specific, treatment referrals would be fewer and more often be appropriate. Again these differences would be mitigated by the quality of diagnostic information provided by additional tests ordered in either arm, as well as the extent to which clinicians base their treatment choices on these findings.

Additional consideration must be given to the choice between antibiotics and surgery to treat appendicitis. Testing strategies could differ in the proportions of patients for whom these treatments are indicated; for example surgeons managing experimental arm patients may demonstrate a lower threshold for selecting surgery as a result of reduced diagnostic certainty.

Therapeutic confidence (o. Clinician confidence; p. Patient confidence)

Surgeons are traditionally more confident in recognising the signs and symptoms of acute appendicitis than general emergency physicians[294], however surgeons may find CT increases therapeutic confidence since in the US these images are also used as a 'road–map' to guide treatment in patients with complicated appendicitis[299].

Patient confidence in the treatment choice is unlikely to be relevant, though could conceivably be reduced in disease–free individuals who do not receive an imaging test, perhaps leading to an increase in reattendance and resource expenditure in the experimental arm.

Timing of treatment (q. Time to treatment)

Differences in the speed with which treatment is administered could occur in three ways.

Firstly, any delays in diagnosis due to the requirement to confirm initial suspicions with a CT scan could mean that treatment is delayed in control patients, when compared to experimental patients who are spared a CT. However such delays could be longer in the subgroup of experimental patients who are referred for CT.

Second, treatment can be delayed as a consequence of missed diagnoses (false–negative appendicitis cases), which are likely to be more numerous in the study arm. Previous studies clearly document how a delay in treating acute disease can cause the appendix to rupture (or perforate)[301], an event associated with a five–fold increased risk of death (from 1% to >5% case–fatality rate in non–perforated compared to perforated appendices respectively), or inflammation to spread to the peritoneum which is associated with other severe health complications[289,299].

Third, delays in receiving appropriate treatment could also occur in patients falsely diagnosed with appendicitis (false–positive appendicitis cases) who are not disease–free but have alternate diagnoses. There is little existing evidence to indicate which diagnostic strategy would more commonly identify such individuals correctly, however CT is thought to be valuable when differentiating acute appendix from gynaecological conditions[289]. If CT is superior in all such cases then we might expect the risks of further health deterioration, and the development of other morbidities associated with a missed diagnosis, to be more common in the experimental arm.

Treatment efficacy (r. Efficacy of treatment)

The efficacy of treatment will depend on three factors: whether the treatment is appropriate for the patient's true condition, whether treatment is delayed, and the quality of treatment. All three are argued to differ between the diagnostic strategies under comparison.

Patients who receive incorrect diagnoses will receive treatments that are unlikely to be effective. If therapeutic yield is expected to be lower when using the selective strategy due to reduced sensitivity but higher specificity, fewer patients randomised to this arm will receive the treatment they need due to a false–negative diagnosis. Although appendicitis may resolve spontaneously in some of these[302], discharge is unlikely to be effective in all, and we might therefore expect higher rates of 'recurrence', or re–attendance for appendicitis symptoms, in the experimental arm.

In the control arm, higher rates of CT–led over–diagnosis will lead to inappropriate treatment in patients without true acute appendicitis; thus the new strategy could be of benefit by reducing the risks of unnecessary surgery[303]. However this is not guaranteed since a lower treatment threshold in experimental patients due to decreased confidence, or a lower diagnostic specificity compared to CT, could also increase the number over–treated in this group.

Any test–related differences in the types of treatment selected (antibiotics or surgical removal) could also influence the success of treatment: fewer indications for surgery would reduce the rate of associated harms, while higher indications for conservative treatment could increase recurrences due to initial treatment failure (although subsequent appendectomy is shown to carry a very low risk of complications)[304]. Rates of surgery could be higher in the control arm if those with indeterminate scans are sent for surgery due to fears of reprisals for missed diagnoses.

On the other hand, if CT might improve treatment success in complicated appendicitis cases by avoiding the need for surgery (e.g. percutaneous drainage of abscess) due to the increased therapeutic confidence of surgeons.

Delays in receiving appropriate treatment, whether as a result of missed diagnoses or the need to wait for confirmatory imaging, will cause some cases of disease to progress risking perforation, abscess and peritonitis in true–appendicitis cases and other morbidities in those with differential diagnoses. Following the rationale above, both arms will be exposed to these risks, the control arm due to potential delays in receiving a diagnosis and the experimental arm due to higher rates of missed disease.

Adherence (s. Adherence to treatment)

The concept of adherence is not relevant to surgery, though could influence the health of patients prescribed antibiotics. However it is unclear what effects receiving a CT scan might have on this mechanism.

Summary of rationale

These considerations can be synthesised to outline the four causal pathways by which the selective strategy might be expected to impact on patient health:

Direct impact: the new strategy is expected to be of clear benefit to patients by reducing exposure to the procedural harms of undergoing contrast–enhanced CT. The degree of improvement observed will depend on clinicians’ confidence in proceeding with treatment decisions without the aid of CT.

Decisional impact: the new strategy is likely to detect fewer true–positives, leading to lower rates of treatment success. This disadvantage could be reduced if the selective strategy benefits from lower rates of failed procedures (due to oral contrast intolerance) and indeterminate results. The new strategy may also detect fewer false–positives, leading to lower rates of treatment harm. However this advantage may be mitigated by low diagnostic and therapeutic confidence that results in more CTs being ordered than necessary, or it may not materialise at all if CT is also more specific causing rates of treatment harm to perhaps be higher.

Temporal impact: the new strategy may be more expedient in treating typical presentations, leading to lower morbidity caused by progressing disease. However treatment may be less expedient in atypical cases which could mitigate the overall benefits observed.

Perceptive impact: no impact expected.

8.2.2 Choosing outcomes to evaluate selective CT

These considerations serve as prompts to select meaningful measures of patient health, but also to identify intermediate outcomes that allow us to evaluate whether hypothesised mechanisms do affect patient management processes as intended.

So, direct health impacts are captured by recording the number of patients receiving a CT scan, the per–person exposure to radiation and the frequency of contrast–induced morbidity (Table 8.2). It is worth noting here that due to their very long–term nature, the consequences of radiation will not be quantifiable in an RCT. The above rationale concludes that downstream health is likely to be affected by differences in how and when diagnostic decisions are made. Thinking through the decisional pathway highlighted three groups of patients who could be identified to differing degrees of accuracy: 1) acute appendix cases, 2) disease–free cases and 3) diseased individuals with a differential

Causal Pathway	Intermediate outcome	Patient outcome
Direct:		
Test Process	Number of patients receiving CT	Mean per patient exposure to radiation % procedural morbidity
Decisional:		
Feasibility	% tests not completed due to contraindication, contrast failure or other reasons	% symptom recurrence % resolution of presenting condition % complicated disease (of all diagnosed conditions, including complicated appendicitis [perforated appendix, peritonitis, abscess]) % therapeutic complications % death Time–to–recovery
Interpretab–ility	% of indeterminate test results	
Accuracy	% diagnosed following initial discharge (FN rate) % diagnosed with appendicitis during follow–up (FN–target condition) % diagnosed with non–appendiceal disease during follow–up (FN–other diagnoses) % negative appendectomies (FP rate)	
Diagnostic Yield	Rates of all diagnoses per test Rates of final diagnoses % diagnosed with acute appendicitis/no disease/differential disease	
Therapeutic Yield	Rates of each prescribed treatment	
Treatment efficacy	% readmitted for failed treatment (e.g. surgery for failed antibiotics)	
Diagnostic confidence	% cases for whom additional tests ordered % with a change in diagnosis as a result of additional testing Reason for ordering additional tests % treated against indication of test results Reason for treating against indication	
Therapeutic confidence	% cases where CT used to guide treatment approach	
Temporality:		
Timing test	Time–to–CT	
Timing diagnosis	Time–to–definitive diagnosis	
Timing treatment	Time–to–treatment Length of stay	

Table 8.2 Outcome selection for the putative trial comparing standard routine CT with a more selective strategy for patients presenting with suspected appendicitis[T60].

diagnosis (appendicitis–free). Consequently the primary useful measure of downstream health should evaluate the degree to which their initial complaint has been resolved, perhaps by measuring symptom resolution or recurrence at the end of an appropriate period of follow–up. To capture mis–diagnosis, rates of perforated appendix, abscess formation and wound infection should be measured, as should a more general marker such as deterioration in health, in order to capture the effects of misdiagnosis in individuals without appendicitis who have differential diagnoses. Of course, these harms may also be caused by delays in arriving at a diagnosis. These two causes could be differentiated by measuring relevant mechanisms as intermediate outcomes. Accordingly, the impact of decision–making could be assessed by recording the number and character of diagnoses (diagnostic yield) and treatment decisions (therapeutic yield) made by each strategy. The appropriateness (accuracy) of these decisions could be assessed by comparing the proportion of patients with recurrent symptoms (total false–negative rate), the proportion treated for appendicitis during follow–up (appendicitis–specific false–negative rate), as well as the proportion of patients with histologically–normal appendices removed during treatment (false–positive rate)[*]. The impact of other influences on the accuracy and appropriateness of decision–making could be assessed by measuring complementary aspects of decision–making behaviour, such as referrals for additional tests and the resulting changes in diagnosis and treatment selection.

The extent to which differences in timing might be responsible for downstream patient outcomes could subsequently be inferred by looking at time–to–CT completion, time–to–definitive diagnosis, time–to–treatment and of course time–to–recovery.

[*] It is important to note that while these processes may indicate accuracy, they are not true measures of accuracy since a reference diagnosis cannot be administered to all patients.

8.3 Applications of the tool

This chapter has presented a practical tool to guide researchers in formulating a clear scientific rationale for the intended effects of a putative test-treatment intervention, by comparison to an existing strategy. The diagram initially focuses the user to give detailed consideration to the composition of test-treatment strategies by requiring each of the five components to be defined, as well as to other key aspects of comparative diagnostic setting. The checklist builds on basic principles by providing a vehicle for elucidating the likely processes of change within the comparison, allowing the rationale of the new strategy's effectiveness to be made explicit through a structures and replicable consideration of mechanisms. This thesis argues that in so doing the tool provides a strong potential to further the evaluation of a test's clinical utility in four key ways.

8.3.1 Designing test-treatment trials

Use of the tool can enhance the design of test-treatment RCTs by helping to ensure that studies fully evaluate competing strategies. Due to the complexity of interactions that can influence health measures, determining which trade–offs are occurring, and estimating how they will ultimately exert their effect, is a task beset by convolution and complexity. The worked example above illustrates this well; though a seemingly straightforward evaluation, the comparison of selective CT vs. routine CT for suspected appendicitis contained numerous potential trade–offs which could pull treatment effects to favour either strategy. For example, appropriate avoidance of a CT scan in experimental arm patients could improve recovery by expediting treatment, yet the new strategy may also delay appropriate treatment in others as a result of reduced sensitivity. Establishing the nature of these often subtle interactions is difficult without a structured frame of reference, risking the incomplete delineation of cause and effect.

Based on a generalised theory of how tests impact on patient health, the tool provides the necessary frame of reference – a structure for thinking through these issues, allowing causal assumptions to be made explicit and clarified, while drawing attention to the potential hidden benefits of a competing technology or strategy. This focuses the consideration of which endpoints should be measured in order to capture the full range of impacts to patient health, while the measurement of relevant mechanisms allows a nested evaluation of process outcomes to clarify whether causal pathways are operating as hypothesised.

8.3.2 Establishing the need for an RCT

Working through the checklist could also assist in ascertaining whether an RCT is necessary to demonstrate a test's clinical utility. Answering each question demands consultation of the existing knowledge base, so a completed checklist provides a summary of evidence regarding how mechanisms are already known to differ between tests.

Comparisons where evidence for all relevant mechanisms has already been documented are unlikely to require RCTs that measure long-term outcomes, since the causal pathways can be pieced together from existing evidence. This is similar to the approach used by the US Preventative Task Force (USPTF), in which causal linkages between a screening programme and desired health outcome are conceptualised and used to target the evidence needed to produce an effectiveness review[78]. Lord and colleagues have also discussed how existing evidence should be linked in comparisons driven by changes in accuracy[19,21,93], although the thesis framework suggests these linkages are likely to be more numerous and more complicated than previously considered, as discussed in Chapter 7.

In agreement with these authors, however, the framework tool may be useful in constructing the case for an RCT when trade–offs occurring between mechanisms cast uncertainty regarding their ultimate impact on patient health. When examining the potential effects of the competing diagnostic strategies in suspected appendicitis patients, the rationale above highlighted the uncertainty as to whether the selective approach would reduce the inappropriate treatment rate through its potentially superior specificity, or increase it due to lower diagnostic confidence. The framework highlights the difficulties of approximating diagnostic settings with highly complex pathways of causal change to patient outcomes. Comparisons that involve multiple causal pathways are unlikely to be appropriate for evaluation by decision–modelling, unless sufficient information already exists to demonstrate the likely impact to patient health, and in these situations RCTs may be the only design that can establish how patients will respond to the entire test-treat strategy.

Yet if adequate evidence of all relevant mechanisms does exist, the wisdom of conducting an RCT becomes questionable. An example is found in the project cohort of trials. Liu and co-workers set out to evaluate whether triage with endoscopic ultrasound (EUS) in patients with suspected biliary pancreatitis reduced morbidity by avoiding endoscopic retrograde cholangiopancreatography (ERCP), a more risky and invasive procedure, in test-negative patients[(T9)]. Timing here is not at stake since the tests are conducted in series during the same anaesthetic. Existing research demonstrates EUS has a lower failure rate than ERCP[305,306], causes significantly less procedural harm[307], is at least equivalent in accuracy with higher sensitivity for detecting small stones[306,308–311], increases diagnostic confidence[312] and improves treatment decisions[313]. Since subsequent mechanisms are all positively affected by these attributes no further trade-offs are involved, and since those mechanisms that are relevant have already been documented, we can conclude we already have enough information without conducting this trial. In such a situation applying the framework tool could save significant resources by highlighting that a decision–model is the more appropriate choice of study design.

This example promotes the framework tool's utility to funders: had the trialists submitted their trial design with a completed checklist, it would have required them to specify what was already known, what they would measure and therefore the nature of additional knowledge the trial would provide. Furnished with this information, it is unlikely the above trial would have been funded.

8.3.3 Interpreting and implementing trial results

Failure to demonstrate effectiveness in an RCT demands cautious interpretation, and should be judiciously distinguished from clear evidence of *in*effectiveness[314]. Findings of no effect are all too often interpreted as 'evidence of absence', when in reality studies rarely make provision for being able to attribute negative results to a truly ineffective diagnostic intervention, a methodologically flawed study design or, importantly, a poorly implemented test-treat strategy. The framework tool confronts these problems directly by providing guidance as to which processes are relevant and need to be measured.

All mechanisms in the framework can be measured as process outcomes, and as attributes that characterise the workings of a given care pathway their measurement provides a critical account of *how* a test-treat strategy performs. Weak links in the causal chain can be identified and strengthened to enhance effectiveness. For example, tests which provide earlier information will only yield benefits when information systems exist to deliver those results in a timely way to allow clinicians to initiate treatment earlier. Assessing relevant mechanisms (e.g. time-to-diagnosis, time-to-treatment) will evaluate these processes. Not only would this approach facilitate the identification of which components in a new intervention might have failed, thus enabling the adjustment of specific mechanisms rather than the entire management strategy, but it also proffers an insight into the mechanisms by which an *effective* intervention has succeeded. This information can be harnessed to organise the implementation of successful strategies, since policy makers can more easily identify the causative ‘ingredients’ of test-treatment interventions and formulate clinical policy guidelines accordingly.

Identifying and measuring mechanistic outcomes are also of use to monitoring discrepancies between the course of management *intended* with that actually *conducted* during the course of the trial, facilitating a subsequent discourse regarding why such differences may have emerged and how they may have affected effectiveness.

The principle of linking process and patient outcomes has been championed by researchers of complex interventions in both healthcare[83,315] and social economics[316–317], both groups of whom have embraced a ‘theory–based approach’ which emphasises the importance of modelling interventions prior to their evaluation so as to enrich the interpretation of subsequent trial results. These discussions have yet to influence diagnostic research, however, where the multiplicity of effects a test can exert on the treatment effect is generally not well articulated.

8.3.4 Appraising trial quality

Lastly, the checklist could assist appraisals of whether test-treatment trials have measured all important outcomes. To illustrate let us return briefly to the worked example and compare the outcomes highlighted by the framework with those actually measured in the trial.

In their introduction the investigators claim to evaluate the hypothesis that “selective CT imaging would reduce the use of CT without increasing the negative appendectomy and perforated appendix rates in comparison with mandatory imaging” [(T60)]. To achieve this they measured the negative appendectomy rate as the primary endpoint, and also compared the frequency of CT scans, diagnoses of acute appendicitis, frequency of surgical treatment, mean time to surgical treatment, the rate of perforated appendix and mortality.

Randomising 152 patients (80 to selective, 72 to control) the authors found no significant differences in the negative appendectomy rate (6/43 vs. 1/39, 11.3% absolute difference [95%CI: –3.5%, 26.3%][†]) or secondary outcomes excepting the number of CT scans which were significantly fewer in the selective strategy (70/72 vs. 54/80, 29.7% absolute difference [95%CI: 18.2%, 40.8%]). Based on these findings the trialists conclude that while a selective strategy will significantly reduce CT use it may also associated with a trend to increased rates of negative appendectomy.

However when this is compared to the framework's rationale we can see that this conclusion is incomplete and not particularly useful as evidence to indicate the utility of either strategy.

How appropriate are patient outcomes?

By comparison to the more extensive list of proposed outcomes developed above it is immediately clear that trialists measured (or reported) far fewer outcomes than were indicated by the framework (Table 8.3). Significantly, the trial does not appear to have fully evaluated the impact of the intervention on all relevant patients. Firstly, trialists do not measure any downstream health benefits such as symptom recurrence or resolution, thus the trial cannot provide a proper indication of whether patients are better off as a result of receiving either strategy. Potential harms are also inadequately assessed; the frequency of perforated appendix wisely captures an aspect of harm, however its focus on the 'target condition' fails to take into account the health consequences of patients with other conditions who may have suffered harms as a result of undergoing inappropriate or delayed diagnosis and treatment. We need to know about the total rate of adverse outcomes, in all patients who will undergo the strategy, to be sure that the intervention is doing more good than harm.

[†] Figures presented as reported; note the wrong denominator is used as discussed in chapter 6 (p.178–180)

Causal Pathway	Intermediate outcome	Patient outcome
Direct:		
Test Process	Number of patients receiving CT	Mean per patient exposure to radiation % procedural morbidity *% procedural complications of CT*
Decisional:		
Feasibility	% tests not completed due to contraindication, contrast failure or other reasons	% symptom recurrence % resolution of presenting condition % complicated disease (of all diagnosed conditions, including complicated appendicitis [perforated appendix, peritonitis, abscess]) % therapeutic complications % death Time–to–recovery
Interpretab–ility	% of indeterminate test results	
Accuracy	% diagnosed following initial discharge (FN rate) % diagnosed with appendicitis during follow–up (FN–target condition) % diagnosed with non–appendiceal disease during follow–up (FN–other diagnoses) % negative appendectomies (FP rate)	
Diagnostic Yield	Rates of all diagnoses per test Rates of final diagnoses % diagnosed with acute appendicitis/no disease/differential disease	
Therapeutic Yield	Rates of each prescribed treatment	
Treatment efficacy	% readmitted for failed treatment (e.g. surgery for failed antibiotics)	
Diagnostic confidence	% cases for whom additional tests ordered % with a change in diagnosis as a result of additional testing Reason for ordering additional tests % treated against indication of test results Reason for treating against indication	
Therapeutic confidence	% cases where CT used to guide treatment approach	
Temporality:		
Timing test	Time–to–CT	
Timing diagnosis	Time–to–definitive diagnosis	
Timing treatment	Time–to–treatment Length of stay	

Table 8.3 Outcomes measured by the selective vs. routine CT for appendicitis trial(T60) (blue) compared to those identified by the framework but not measured (grey) and those partially measured (red).

How appropriate is the primary outcome?

The primary outcome (negative appendectomy rate) evaluates the false–positive diagnosis rate. The framework perspective highlights that this surrogate outcome only measures accuracy and diagnostic yield, without assessing the impact these two mechanisms have on patient health. Since we have determined that several other mechanisms may mitigate these effects, on its own the false–positive rate is an inadequate surrogate for patient health.

One could argue that a more appropriate study hypothesis would be to assess whether selective CT *'can reduce the overall yield of procedural harms without increasing the risk of causing downstream harm to patient health resulting from inappropriate or delayed treatment'*. In addition to mortality, Table 8.2 listed two adverse health outcomes that would capture the effects of the whole care pathway, and hence could constitute an appropriate primary outcome. This would of course require a much larger sample size to ensure the trial were powered to detect these health effects.

Scientific rationale and intermediate outcomes

There can be little doubt that this incomplete selection of patient outcomes reflects the less than comprehensive attention given to the rationalising how the new strategy was expected to create differences in both clinical processes and patient health. Authors initiated their reasoning appropriately in introductory paragraphs by identifying the potential shortcomings of standard care: CT radiation and contrast material may be harmful; as a less than perfect test CT may miss or falsely identify disease; inaccurate diagnoses result in unnecessary surgery and delay of appropriate treatment, both of which can cause morbidity. So far so good. At no point, however, did the investigators frame these contributing factors within a comparative context: they failed to make a case for how selective CT might *differ* in these respects to routine use of CT. The only clue is furnished in their study hypothesis which expects no difference in decision–making at the benefit of reducing the use of CT. Perhaps as a consequence the investigators have neglected to consider other factors that may mediate how a selective approach changes decision–making and downstream patient health. As a result the actions of several important mechanisms, particularly feasibility, diagnostic confidence and therapeutic confidence, were not evaluated. What is more, some mechanisms that were identified were not measured as comparative indicators of performance; for example the rates of final diagnoses were not examined per arm, thus preventing an appraisal of diagnostic yield.

As a consequence we are unable to fully interpret and therefore use the results of this trial. While the selective strategy did significantly reduce the use of CT, no other differences were found. Why were there no differences in negative appendectomy or perforated appendix rates?[‡] Is it because the two strategies produced equivalent diagnostic and treatment yields within comparable timeframes? Did clinicians in either arm resort to additional diagnostic procedures to achieve this equivalence? Did clinical judgement supersede CT results or *vice versa*? Could important adverse events have occurred after the end of study follow–up? Did this equivalence come at the cost of harming patients with diagnoses other than appendicitis? None of these questions can be answered on the basis of published results. Since we cannot be sure of how these results were achieved, we cannot be sure of the impact that implementing a selective CT policy would have on patients presenting to emergency departments with acute lower right quadrant abdominal pain. In this instance the tool has provided a structure for exploring the scientific rationale

[‡] This trial suffers from several other methodological issues that lead the appraiser to question the validity of its conclusions, not least of which is likelihood of type II error; the power calculation did not account for negative appendectomies only being detectable in patients undergoing an appendectomy, reported as approximately 50% in both arms. To account of this the sample size should have inflated its estimate (n=140) by this figure giving 280 (140 x 1/50%). Interestingly, if the same rates of negative appendectomy had been observed in this larger population the finding would have been significant (Chi sq =5.638, p=0.0176). What is more, the trialists inappropriately used a superiority analysis to evaluate their hypothesis that was clearly framed as a question of *non–inferiority* (that negative appendectomy *would be no worse*); this may also have necessitated a larger sample size[145].

of a trial, unpicking its aims and appraising not only whether it has achieved its aims but whether investigators could have evaluated a more useful question.

8.3.5 Study limitations

Due to poor reporting, the author had to make certain assumptions regarding how diagnostic and therapeutic decision–making might be expected to proceed. These were checked with a practicing general surgeon, however since practice is highly variable these may well differ from the decisions intended by the trial investigators. The primary aim was not to be entirely clinically accurate, but to illustrate how a comparison might be worked through to elucidate the workings of comparative test-treatment pathways. It is worth noting, however, that use of the framework tool to appraise trials will be hindered if reporting of the test-treatment protocol is particularly poor.

Critically, the future success of the framework will rely on the ability of users to distinguish the competing clinical pathways and identify which mechanisms are likely to influence patient health trajectories. Whether for appraisal or design clinical expertise in the relevant setting is essential to construct valid and reliable care management pathways. The framework underlines the importance of incorporating patient and organisational perspectives, thus input from all key stakeholders will maximise the reliability of subsequent analysis.

Lastly, the tool is presented as a draft attempt, hence the true extent of the tool's utility will not be apparent until interested parties have trialled it. It is possible that certain aspects may not meet with the demands of those for whom it was intended; two aspects that may need to be developed in future are the absence of specific guidance on how to draw checklist items together into causal pathways, and the tool's restriction to comparing two strategies. In order to develop the tool into a valid resource future work will entail piloting it across diagnostic technologies and improving its utility through engagement with interested stakeholders.

8.4 Conclusions

Expectations that a particular test will improve clinical practice inevitably involve some assumption regarding how the intervention will produce change. As demonstrated in earlier chapters, these assumptions are often poorly articulated and, in all likelihood, incompletely grasped due to both the complexity of how tests impact on heath, and the absence of a conceptual model that offers a template for the theory of change. Accordingly, this thesis has provided a new theoretical framework, and from it developed a methodological tool aimed at identifying the assumptions of causality.

The discussion above argues that the earlier this can be done in the evaluative process, the more reliable and efficient our summative assessments of a test's clinical utility are likely to be. We may avoid unnecessary and expensive trials by recognising we already have all the necessary primary data, or alternatively make a strong case for the need to conduct an RCT and be in a better position to identify all important outcomes and as a consequence provide more reliable trial results that are interpreted comprehensively.

It is clear from table 8.2 that test-treatment trials will need to measure considerably more endpoints than is common in standard treatment trials in order to provide comprehensive evaluations that can be interpreted fully. Limiting patient outcomes to the effects experienced by patients with the target condition, as is done for trials of treatments, can only provide a partial investigation since tests are generally used to differentiate between several diseases with contrasting consequences. Moreover, trials will need to incorporate the measurement of many process outcomes in order to determine why particular effects have been observed. While this presents a considerable addition to the resources needed to undertake a successful test-treatment RCT, measuring these processes could also improve the feasibility of conducting these studies. Changes in intermediate outcomes are likely to be larger and detected with greater power than patient outcomes, which typically only occur in a subgroup of the sample defined by disease status and test results, often a small fraction of the total sample.

Establishing benefit to patient health must remain the priority of diagnostic evaluations, however this thesis argues the importance of engaging in the ‘theory–based approach’ championed by other disciplines[83,316–317]: it is not *sufficient* to measure endpoints, but it is essential to understand how these outputs are created by conducting tangible analyses of their workings. The application of the framework to proposed or completed test-treat evaluations facilitates this approach and encourages a comprehensive understanding of the intervention, features which will be of use methodologists, trialists, reviewers, guideline developers and funders of clinical effectiveness.

9

Discussion & Conclusions:

The utility of test-treatment RCTs

The introductory paragraphs to this thesis highlighted the ever growing need for a high–quality evidence–base that can inform clinicians, reviewers and policy–makers on which diagnostic tests result in better patient health. Increases in diagnostic test research over the last three decades have been dominated by demonstrating the accuracy of these tests. Despite agreement that such evidence is insufficient to warrant the uptake of new technologies, many diagnostic effectiveness reviews currently fail to find direct evidence of the test's impact on downstream patient health. By analogy with the evidence–based evaluation of treatments, RCTs are recommended as the 'gold–standard' approach. Yet these 'test-treatment' RCTs have attracted criticism that lead one to question how useful they are likely to be in providing the high–quality evidence that is needed. In the absence of substantial systematic exploration of these criticisms, the author set out to begin to answer this question by finding and then analysing all identifiable test-treatment RCTs published between 2004 and 2007. Four aims were used to address the central research question, each reflecting a challenge that these RCTs are hypothesised to face. The analyses performed have begun to provide evidence regarding how useful test-treatment RCTs are.

This final chapter first provides a summary and interpretation of main findings from chapters 2–8, and discuss the extent to which they answer the four thesis aims set out in chapter 1. After considering the limitations to the thesis, implications for practice and proposals for future research are presented. The thesis concludes by addressing the central research question: how useful are test-treatment RCTs?

9.1 Aim 1: Are test-treatment RCTs feasible?

In response to concerns that test-treatment RCTs are rarely found when conducting effectiveness reviews, the first aim was designed to investigate whether these studies can be completed successfully. Two studies were performed to address this question.

Chapter 3 presented a capture–recapture analysis in which the project search was compared to a second, different search of CENTRAL in order to estimate the number of relevant RCTs missed by both strategies. By allowing the total number of test-treatment trials published (2004–2007) to be estimated, this approach also allowed verification that the project search had not inadvertently missed many trials due to the potential difficulties in identifying them. Despite using a search strategy that specifically targeted certain tests, the analysis confirmed trial ascertainment to be very resource-intensive, requiring many thousands of records to be checked for yields of 0.9–1%. Approximately 145–146 test-treatment trials were indexed in CENTRAL over the search period, suggesting that only 36–37 test-treatment RCTs were published per year between 2004 and 2007. This is just a tiny fraction of the total number of RCTs that are published in medical journals every year. These findings confirm existing opinion that test-treatment RCTs are rare. Whilst the very low yields of relevant trials found by both searches also attests to the great difficulties in locating these studies, the relatively low estimate of missed trials (n=12) suggests the rarity is not artefactual.

That RCTs are not attempted could signal that these designs are often considered too difficult to evaluate diagnostic tests. Indeed, this explanation is expounded by individuals and research organisations alike, who commonly cite logistical difficulties and resulting expense as key barriers to the feasibility of conducting test-treatment RCTs[15–16,51,95]. Hunink and Krestin, researchers with extensive experience of conducting test-treatment

RCTs, argue that trial results may often not justify the associated ‘price tag’ [6]. Discussing the conflict between the need to keep up with the rapid development of test technologies and the need to perform thorough evaluations of effectiveness, they argue that ‘new’ diagnostic tests often either present incremental improvements to existing technologies, or are added to existing strategies and thus many trials must recruit unrealistic sample sizes in order to capture effects that will generally be very small.

The detailed characterisation of included trials presented in **chapter 4** provides further insight into potential reasons for the scarcity of test-treatment trials. Aiming to distinguish which diagnostic settings were successfully evaluated to completion, this analysis described the cohort of trials identified in chapter 2 and provided an overview of the diagnostic questions they evaluated. Despite considerable overall heterogeneity, there was a clear predominance of evaluations of imaging and to a lesser extent biochemical tests, tests used in cardiovascular medicine, secondary/tertiary care settings, and replacement comparisons. The latter finding corroborates the ‘feasibility’ argument, since it could indicate that trials are more likely to be attempted or successfully completed when effect sizes are largest.

On the other hand, trials most commonly evaluated questions posed by disciplines that also have the strongest traditions of academic research, namely imaging, biochemistry and cardiovascular disease. This warrants consideration of an alternative explanation, that perhaps trials are not carried out in other disciplines because the need for effectiveness research is not as well articulated, and possibly not as well understood. Within the wider clinical and research communities, test accuracy studies are commonly misconstrued as effectiveness studies – as put eloquently by Patrick Bossuyt, the value of tests are often thought to lay in the *truth* of their results rather than in the *consequences* of their results[318]. Though the dissemination of rigorous methods for undertaking accuracy studies and reviews has had many benefits, not least by highlighting the need for evaluating tests[58], the focus on such methods may also have unwittingly generated the perception that trials of tests are unnecessary when evidence of their accuracy already exists. This is somewhat supported by the observation that the complexity of designs *per se* did not appear to be a barrier to performing test-treatment RCTs, with many included studies comparing very complicated strategies evaluating multiple phases of testing. Increasing awareness that effectiveness evaluations are needed[12,44,75,77], as well as providing guidance on how to conduct these trials, could therefore increase the number published quite substantially, and thus increase the availability of evidence that is needed to discern whether tests do more good than harm to patients.

9.2 Aim 2: How informative are test-treatment RCTs?

A second key requirement for trials to be useful is that they produce evidence that can be interpreted and translated into practice. In response to claims that achieving these two requirements could encounter particular difficulties due to the complexities of test-treatment interventions, the author set out to evaluate the extent to which published trials produce evidence that is informative by evaluating the reporting quality of included trials. The review examined the extent to which the reader could satisfactorily glean what was done, to whom, when and why. This was achieved by appraising the adequacy of descriptions of test-treatment interventions, complete reporting of participant flow and definition and documentation of primary outcome assessment.

Observed reporting quality was very poor for each of the three items assessed. Incomplete accounting of participant flow was common, hampering a full interpretation of the meaning of observed test-treatment effects. Although primary outcomes were defined by almost all trials, methods and timing of measurement were beleaguered by incomplete reporting, a barrier to the replication of measurements and highly likely to lead to problems when seeking to compare effects between trials. The most important finding was the meagre documentation of what interventions were used, how they were used, and why. Only three trials provided at least some detail of tests and treatments and decision–making for all comparative interventions, and generally several components were missing. Control interventions were particularly poorly outlined, as were the decision–making processes of all interventions. As a result the nature of what was being compared was frequently unclear. This was argued to pose two limitations to using RCT evidence. First, without clear and full descriptions of the tests given to patients, the operational criteria for diagnoses, and how these should lead to the selection of treatments, it will be almost impossible to interpret how observed results were created.

Second, poor descriptions were likely to inhibit the ability to reproduce apparently beneficial interventions in a manner that is not only safe but that will be able to replicate the desired effects. On this basis the review of reporting quality concluded that published test-treatment trials do not produce sufficient detail for users to interpret trial results, nor to translate test-treatment interventions into practice.

The poor level of reporting was also supported by evidence from other chapters. Difficulties were encountered when attempting to characterise the diagnostic questions evaluated by included test-treatment trials in chapter 4, particularly in trying to discern what these trials had evaluated. The review of methodological quality in chapter 6 also revealed that important aspects of trial conduct were very poorly reported, at times curtailing the ability to assess the true quality of methods.

Concerns that the complex nature of these decision–making, multiple–component interventions could create increased difficulties in full reporting appeared to be confirmed when these findings were compared to similar reviews of treatment RCTs and complex intervention RCTs. While failure to report trial methods was partly explained by the suboptimal reporting quality found in all RCTs, the considerably inferior quality of intervention documentation highlighted an issue of particular concern to the production of informative test-treatment trials.

As has been found with complex interventions[156,319], difficulties in standardising diagnostic and therapeutic decision–making processes could account for the common absence of decision–making criteria. Though certainly more challenging to describe, this issue does not in itself prevent adequate documentation.

Reporting issues were all in essence surmountable, though barriers to informative description were also found to be 'conceptual'; by focusing on the experimental test method to the common omission of the three other components, the reporting style evoked a sense of being trapped in the mindset of reporting standard treatment trials rather than adopting a complex intervention approach which is far closer in nature to test-treatment strategies. Indeed, trials did not appear to identify themselves as evaluations of complex interventions.

This would suggest that while test-treatment trials could in future be informative, notions of how this is to be achieved must be addressed and disseminated.

9.3 Aim 3: Are test-treatment RCTs internally valid?

If we are to rely on the results of test-treatment RCTs, these studies must offer the opportunity to provide internally valid evaluations that are free of bias. Yet researchers have claimed that the complex composition of test-treatment interventions makes the methods necessary to minimise bias and error difficult, if not impossible, to implement. The thesis set out to evaluate whether these claims are justified in **chapter 6** by critically reviewing the methodological quality of included trials with full results (n=103), and by examining whether the issues encountered were likely to constitute particular challenges to test-treatment RCTs by comparing these findings to the internal validity of treatment intervention RCTs.

Due to the lack of an existing tool to appraise test-treatment RCTs, the author reviewed the adequacy of the five main threats to the internal validity of RCTs using items contained in three validated, internationally accepted standards for the optimal conduct and reporting of RCTs: selection bias, performance bias, ascertainment bias, attrition bias and type II error. The analysis revealed considerable challenges to the validity of published test-treatment RCTs due to the suboptimal implementation of all five methodological safeguards, raising the distinct possibility that their results reflect artefacts of study design and subjective expectations rather than true clinical effectiveness. Comparison to the methodological quality of treatment RCTs demonstrated that some of these inadequacies could be explained by the suboptimal performances generally found in all RCTs; nonetheless test-treatment designs were characterised by higher rates of patient exclusion and drop–out, smaller sample sizes and a greatly reduced propensity to blind patients, clinicians and to a lesser extent outcome assessors.

9.3.1 Practical barriers to internal validity: attrition and lack of power

A key finding of the review is the empirical confirmation that test-treatment trials are particularly susceptible to attrition and lack of power. Although excluding randomised participants is generally inappropriate in any RCT[320], the loss of data through patient drop–out could be amplified by the practical difficulties of maintaining patient compliance during trials that involve numerous interventions, and are thus characterised by longer study periods and potentially more intensive follow–up regimes.

These difficulties become particularly acute when trials must recruit much larger study populations in order to stand an acceptable chance of detecting a clinically meaningful difference in treatment effect. Failures to attain target sample sizes, accompanied by some reports of problems in recruitment, attest to the practical difficulties these trials are likely to face in order to address study objectives adequately. Of course, trials that do power adequately would need to recruit considerably higher numbers of patients, which as Deeks points out could theoretically prove prohibitive if sufficient patient numbers do not exist[176]. Even if they do, attaining large sample sizes will require engagement with more study centres, longer recruitment periods, and hence more expense[6].

Nevertheless, as practical issues these difficulties can be overcome: there is no theoretical reason why test-treatment RCTs cannot minimise the risks of attrition bias and type II error to a similar degree as standard treatment RCTs.

9.3.2 Methodological barriers to internal validity: performance and ascertainment bias

In contrast, blinding was shown to pose a considerable challenge to the validity of existing and future test-treatment RCTs. Very few published trials performed blinding, but critically 'double–blinding' (defined as masking 2 or more of the patient, clinician, outcome assessor) is likely to be impossible in virtually all test-treatment RCTs due to the considerable practical difficulties involved in performing sham diagnostic procedures, masking real test results and producing standardised diagnostic reports. As argued in chapter 6, in view of strong meta–epidemiological evidence associating the lack of double–blinding with bias, these findings can be interpreted to indicate that most test-treatment RCTs are at risk of producing distorted treatment effects; and what is more, most future trials will not be able to avoid the risk of these biases.

This inference assumes that test–treat effects are distorted in the same way as effects produced by treatment trials, yet how successfully the concept of performance bias is transposed to test-treatment designs presents an intriguing dilemma. Blinding was designed to isolate the effects of pharmaceutical treatments from any other influences that are extraneous to the subject of evaluation[321]. These factors constitute any behaviour that could influence the causal link between the treatment a patient receives, and its effect on health. Clinician behaviour is controlled to avoid knowledge of the allocated treatment triggering the provision of additional/differential care. Patient behaviour is controlled to avoid knowledge of which treatment has been received influencing adherence, consumption of additional medicines or therapies, as well as the treatment response itself. Outcome assessor behaviour is controlled to prevent knowledge of the allocated treatment from unduly influencing the measurement of study endpoints. When all three controls are implemented, this allows us to be more certain that the only factor causing observed differences is the treatment[322]. Thus blinding strengthens our certainty in causality.

This argument is not quite so straightforward when we consider test-treatment RCTs; since these interventions aim to cause effects by changing clinical behaviour it is far more difficult to determine which aspects of the intervention are extraneous, and therefore to isolate their effects. The finding that blinding patients or care–providers may alter or even eliminate desired effects suggests that attempting to control for performance bias is not appropriate for all test-treatment RCTs. While further work is urgently required to determine how blinding (or the lack of it) influences test-treat effects, this observation raises the possibility that the inability to blind clinicians may not always pose a serious threat to the internal validity of test-treatment RCTs, and therefore should not automatically be considered as an obstacle to achieving reliable results.

On the other hand, the risk of ascertainment bias is likely to remain when outcomes are not assessed in blinded manner. Though blinding outcome assessors was more often feasible, it was still judged to be impossible to achieve in a third of trials. This finding presents a more pessimistic situation than forecast by previous researchers, who had projected that it would generally be possible to blind outcome assessors[19,44,174]. Most worrying was the discovery that it appears virtually impossible to blind assessors (generally the patient) when the outcome being measured is subjective. Since distortions in treatment effects from lack of blinding are reportedly higher in trials measuring subjective outcomes[107–108], one can therefore conclude that the risk of ascertainment bias is likely to be considerable in test-treatment trials that seek to measure such effects.

9.4 Aim 4: Do test-treatment RCTs fully evaluate their interventions?

The final challenge levelled at test-treatment RCTs speculates that identifying and measuring all important effects may prove demanding due to the indirect relationship between tests and downstream health outcomes. As an initial step to tackling this issue, the thesis undertook to develop a solid theoretical understanding of how test-treatment interventions cause effects, which could act as the necessary framework of reference for future assessments of whether test-treatment RCTs have fully evaluated their interventions. Two studies were conducted to achieve these objectives.

Chapter 7 was designed to develop the theoretical framework that conceptualises all the ways in which tests influence health outcomes. It achieved this by synthesising existing theories, and used them to generate a preliminary explanatory model. This model was tested, refined and explained by examination of the project cohort of published test–treatment RCTs using analytic inductive methods. The resulting conceptual framework presented 14 mechanisms that interact to influence health outcomes in four ways: by direct impact, by altering the decisions made, by altering the timing of the test-treatment process, and by altering the patient's and/or clinician's perception or experience of the test-treatment process. Not only did this framework identify more mechanisms than apparent in the existing literature, but it identified a more complex relationship between mechanisms than is commonly accepted. In particular it posits that these relationships are not linear, as frequently conceived, but that mechanisms share common interdependencies and interact synergistically along multiple causal pathways to alter health outcomes. A key implication was that superior accuracy can no longer be considered as sufficient *or* necessary for improvements to patient health, since these effects may be achieved through other causal pathways – chiefly by expediting the test-treatment process.

Chapter 8 developed these concepts into a practical tool designed to assist users of diagnostic evidence to formulate a comprehensive rationale for how a new testing strategy is expected to impact on health. The tool was presented as a graphic schema, designed to assist researchers to map out the five components of two comparative test-treatment strategies, and an accompanying checklist designed to get users to consider all mechanisms and conceptualise whether each might influence health outcomes.

By reference to several examples derived from the project cohort, the tool was argued to provide added value to four key evaluative tasks:

1. By requiring consultation of the evidence–base to address each checklist question, the tool can assist in establishing whether an RCT is necessary

2. By providing a structure for systematically identifying the many trade–offs that are likely to be operating in even the more seemingly straightforward comparisons, the tool can be used to identify important process and patient outcomes and so could assist in trial design

3. By providing guidance as to which mechanisms are relevant and should be measured, the tool could assist investigators to interpret *how* interventions have caused important health effects

4. By providing a comprehensive frame of reference for all the ways in which test-treatment strategies may differ, the tool could assist to appraise whether existing trials have measured all appropriate outcomes

9.4.1 Have published test-treatment RCTs fully evaluated their interventions?

Although the question of whether trials fully evaluated test-treatment interventions was not assessed directly, the updated theory provided by the new framework, alongside other observations regarding the characteristics and reviews of published trials, begin to provide a tentative insight into this matter.

Firstly, although 50% of trials used health outcomes as their primary measure, half of these were surrogate measures which are not guaranteed to capture all the intended effects of multiple, synergistic causal pathways. Second, during examination of how test-treatment comparisons may be causing their effects for the purposes of developing the theoretical framework, the author found that most trials presented at best partial rationales for how the experimental strategy was expected to benefit patients. Along with the poor documentation of interventions, this may indicate that comprehensive thought was not given as to which aspects of comparative strategies differed and thus were likely to cause an effect. Third, the test-treatment framework revealed more numerous and complex causal pathways than commonly recognised by existing research frameworks. Since the previous perspective did not encompass the increased breadth and complexity of mechanisms at the time these trials were designed, it is a reasonable assumption that many failed to identify all the ways in which new interventions may be causing their effects. If this is the case, then one can deduce that many published trials are unlikely to have fully evaluated their test-treatment interventions.

This of course suggests that existing trials may suffer from not having measured the impact of test-treatment strategies on all individuals who undergo them. An equally important adverse consequence is that by not measuring all important intermediate processes, their results are unlikely to furnish readers with sufficient information regarding why health effects were, or were not, created.

This has also been a major critique of complex interventions. Early RCT evaluations were criticised for selecting a desired health benefit of an intervention and concentrating on whether this effect was statistically significant, to the detriment of how understanding how it was created[317]. Although this 'black box' approach is adequate for pharmaceutical RCTs, which evaluate a short and direct causal chain, it is considered insufficient for complex interventions which are characterised by multiple, interacting components that influence health outcomes in a compound, and often unexpected, way[323–324]. Recent guidance for evaluating[83] and reporting[147] these evaluations emphasise the depth of attention that must be paid to the internal workings of these black boxes in order to perform comprehensive evaluations that can also reveal *why* particular effects were achieved. Process evaluations, in which the causal mechanisms of interventions are identified and measured, are now promoted as a gold–standard adjunct to complex intervention RCTs[83,315] in order to isolate the 'active ingredient' of the intervention. While this point lays at the centre of the MRC's framework that guides the development and evaluation of complex interventions[83], the framework does not provide guidance as to which processes to measure.

The test-treatment framework sits comfortably within the complex intervention structure, and as illustrated in chapter 8 it can be applied to each of the recommended five stages of evaluation: it enables researchers to construct a coherent theoretical argument for how desired improvements will occur; it draws attention to areas of uncertain causality which can be piloted during pre–trial planning; it informs the choice of which outcomes to measure during the trial; it can aid documentation by making the composition of test-treatment interventions clear and the differences between them explicit; and finally it can be used to demonstrate which active ingredients have caused change to facilitate the implementation of successful strategies. The MRC framework enriches the test-treatment framework by iterating how one stage of evaluation should proceed to the next (the MRC's 'development–evaluation–implementation' process)[83], whilst its focus on the impact of organisational differences serves to 'ground' the test-treatment framework which is more conceptual and systems–driven.

Arguably, however, the framework presented by this thesis goes further than the complex intervention guidance. By setting out a comprehensive list of all possible causal mechanisms, it provides explicit guidance on which processes should be identified and measured in order to perform a full process evaluation within a test-treatment RCT. Hence a key conclusion is that trials will be able to conduct full evaluations if they follow a similar process as that advocated by the framework tool. An important caveat is that such efforts are likely to highlight the need to measure many patient and process outcomes, which could prove both costly and challenging.

9.5 Limitations

Limitations to the methods used for individual analyses were discussed in each of the relevant chapters, however it is also appropriate to consider whether the general approach taken may have failed to answer the main thesis question comprehensively.

9.5.1 Indirect methods for evaluating feasibility

The first shortcoming concerns the indirect approach used for investigating how feasible it is to perform test-treatment RCTs. By limiting the analysis to completed and published trials, this research may have missed or underemphasised key practical obstacles to the ability to perform these trials. This potential form of 'publication bias' would have required extensive searching of trial registration databases to quantify, and possibly interviews with trialists to try to capture practical issues more directly. While this was not possible to perform due to time constraints, the author knows of at least two trials originally funded by the NETSCC HTA that were discontinued, and a further 3 that were not accepted for the publication in the associated journal (of which one was published in another journal within the searching timeframe(T31)). Establishing the reasons for discontinuation could provide very interesting insights into the interplay between the practical needs of these difficult studies (particularly recruitment) and the expectations of funders.

9.5.2 Confounding of true methodological quality

An issue already raised in chapters 5 and 6, the thesis established the relative performance of test-treatment designs by comparing frequencies of observed quality items to those found by previous reviews of treatment RCTs and complex intervention RCTs. This is a somewhat crude measure, since it assumes that methodological quality was assessed in the same way by all investigators, and also that reporting reflects methodological quality to an equal extent across all types of RCT. If this were not the case then the interpretation deduced here may not be entirely valid. However the author could find no overt suggestions that reporting of adequate methods was worse, since several methods of trial conduct were found to have been reported with similar frequencies, and in some cases more frequently, than standard RCTs. Additionally, measures were taken to ensure that quality appraisal was conducted to the most rigorous standards, and while some judgements were independently verified to a high degree of inter–rater agreement by an experienced reviewer for a sample of included trials, the findings would benefit from a complete independent appraisal.

9.5.3 Generalisability of test-treatment trials

This thesis did not explore the extent to which published test-treatment RCTs were generalisable, largely because the review's inclusion criteria necessitated the selection of a wide range of clinical settings. Since generalisability is a highly contextualised judgement[325], its assessment in this study would have presented a complex and time–consuming task involving consultation with a wide range of clinical specialists.

However this issue has proven to be a key difficulty for complex interventions, for reasons that are also likely to apply to test-treatment interventions. Due to their many interacting components, complex interventions tend to be characterised by poor fidelity if over–standardised. Therefore there is a tension between designing methodologically rigorous RCTs, which requires standardisation in order to be reproducible, and achieving generalisable results. As discussed above, the properties of test-treatment interventions are likely to be very similar in this regard, thus the threat to the generalisability of test-treatment results could prove to be a crucial determinant of whether these studies are ultimately considered to be useful.

This view is not accepted by within the complex intervention community, however, and recent discussions place the 'exploratory versus pragmatic' argument as a relic of the traditions established within the pharmaceutical RCT paradigm[182,326]. Rather, carrying out embedded process evaluations that measure how interventions are actually administered has become a critical part of the evaluation itself [317]. Some have put forward the need to reconceptualise the notion of 'fidelity to complex interventions', whereby rather than seeking to standardise how the physical intervention is administered, the intended *function* is standardised instead[182]. Following this reasoning, the intervention is designed to achieve common, pre–defined goals though can be modified at a local level to suit organisational differences. Although diagnostic tests differ from the public health interventions for which this approach was conceived, it may turn out to be an interesting proposition when seeking to evaluate ‘unstandardisable’ tests, in particular clinical examinations.

9.5.4 Indirect appraisal of appropriate outcome measures

The thesis did not directly evaluate the extent to which each included trial had measured all the outcomes necessary to have fully evaluated its interventions. The possibility therefore remains that trials performed better in this respect than has been deduced. Performing a reliable appraisal would have required extensive consultation with clinical experts so as to identify the most important outcomes for the full range of diagnostic settings included. Nonetheless, included trials were afforded a highly detailed examination, through appraisal of their methods but also to scrutinise how health effects were created whilst developing the theoretical framework. The author’s experience of this thorough process strongly intimates that the greater majority of trials failed to identify and measure all potential benefits, however these suspicions must be investigated further.

9.6 A comment on the need for trials

This thesis has examined whether the RCT is fit for purpose for evaluating the clinical effectiveness of diagnostic tests. To the author’s knowledge there is no existing research that has sought to tackle this issue directly. However there is a larger body of research that has examined the closely related question of *when* RCTs might be needed to evaluate diagnostic tests. Though the present work did not intend to address this directly, it is the author’s belief that some findings may contribute to this discourse and thus merit brief discussion.

Working under the premise that test-treatment RCTs are unlikely to be feasible, and are therefore unavailable to provide direct evidence of health impact, several researchers have examined what evidence should be sought to estimate health effects without needing to resort to an RCT. Most of these works focus on the ability to use existing evidence of test performance and treatment efficacy to establish the nature of trade–offs between the health consequences of comparative sensitivity, specificity and the procedural harms of undergoing testing and treatment[21,91,256]. Attention is drawn in particular to the meticulous and thought–provoking research published by Sally Lord and colleagues[19,21,93]. The authors argue that assumptions linking changes in accuracy to health outcomes can be confirmed by existing evidence of therapeutic benefit in individuals who would receive discrepant diagnoses as a result of receiving the new test. The authors present a framework[19] for setting out these assumptions, and identifying the discrepant groups so that all possible health consequences of reclassification can subsequently be iterated. For example, they deduce that if the new test has a higher sensitivity but lower specificity than its comparator, then the health consequences of more patients receiving appropriate treatment must be balanced against those of the higher proportion of patients being over–treated (or receiving inappropriate treatment). They argue that discrepancy can also be created when new tests diagnose cases that represent a different spectrum of disease, even though the number of true– and false–positives are equivalent. Key to establishing the need for an RCT is identifying the expected benefits of a new test, which they propose doing by examining the trade–offs occurring within the decisional causal pathway to identify how the potential value to patient health is generated. The authors conclude that comparisons will only require test-treatment RCTs when the treatment response is uncertain in newly identified individuals. Thus when assumptions linking accuracy to health outcomes are not in doubt, lower levels of diagnostic evidence linked with evidence of treatment efficacy will be sufficient.

More recently, Lord et al produced a second framework to assist users in determining what evidence is needed to compare the health impact of tests[19]. This builds on the 2006 framework by presenting the idea of using a hypothetical RCT to identify where differences might arise between two entire test-treatment pathways. While working through the RCT comparison, what the authors refer to as the ‘test evaluation flow diagram’, the authors advocate that the evidence needed to replace an RCT is determined by where differences between arms are likely to occur, and whether their health consequences have been demonstrated. The authors conclude that RCTs are not necessary when all potential consequences of the new test have already been evaluated using studies located earlier in the evaluative hierarchy, which can then be linked together.

As with their earlier framework, the authors concentrate on two potential differences: those defined in this thesis as the ‘direct’ and the ‘decisional’ causal pathways; that is, they conceive that differences in the performance of test-treatment strategies can occur due to the direct impact of tests, and mainly due to the impact of differences in accuracy characteristics.

Although their primary intention was to discusses to what extent accuracy studies and simplified designs could replace RCT designs, the work of Lord and colleagues[19] shares some similarities to the framework presented in chapters 7 and 8. Both share the same premise, that determining the effectiveness of new tests requires the evaluator to map out the competing test-treatment pathways and conceptualise where differences in important processes may occur. Both frameworks also posit that identifying key intermediate processes should drive the rationale for how tests are to be evaluated, since the potential benefits of a new strategy are expressed through these intermediate outcomes. Discussion of the thesis framework tool (chapter 8) considered how the identification of mechanisms (and conceptualisation of how they interact within causal pathways) could be used to target the evidence needed to demonstrate effectiveness, and as a consequence determine whether an RCT is needed; this was of course also the key aim of the Lord framework[19]. Finally, the two works also concur that benefits and harms can be created along multiple causal pathways.

Although these notions are the same, the thesis framework adds to the practical value of Lord’s framework by setting out a complete list of possible mechanisms and causal pathways; this allows the associated tool to offer more specific guidance on how to work through all possible differences between strategies.

However the thesis also augments the conceptual basis of the Lord framework. Finding that the relationship between tests and patient outcomes is far more complex than previously conceived, this thesis reveals a slightly different position to that expounded by Lord and colleagues. The existence of more mechanisms means that in order to avoid an RCT, reviewers will need to find more evidence so as to ensure that potential effects caused by all relevant causal pathways can be quantified. Since mechanisms are far more complex in the ways they interact to cause change, the process of linking this evidence together will therefore require many more assumptions about how interactions between mechanisms will impact on health outcomes. This in turn could portend that uncertain linkages will be more numerous, for example because mechanisms cannot be guaranteed to perform as desired. Not only does this imply that the process of linking evidence together could be more challenging than previously thought, but a logical deduction following the approach advocated by Lord and colleagues would be that RCTs may be needed more often than is currently thought necessary in order to confirm that complicated mechanistic synergies are functioning as hypothesised, or to capture unintended effects which may remain obscure during the development of the scientific rationale.

9.7 Implications for practice

Aside from questions of whether RCTs are useful, the thesis findings suggest several recommendations for how these evaluations should be improved when they are attempted.

9.7.1 Moving towards a full documentation of test-treatment interventions

The reporting of test-treatment trials was very poor and urgently needs improvement if results are to be used to improve diagnostic practice. The framework in chapter 7 suggests that five key components of care must be described for each intervention in order for reports to be useful. Thus even when test-treatment comparisons are relatively simple, adequate reporting is likely to require far more detailed documentation of the study setting than is currently expected for standard treatment trials. Particular attention will need to be paid to interventions that are less amenable to standardisation, since this study suggests they were more poorly reported. Attempts to provide fuller descriptions for these intrinsically more variable strategies are likely to necessitate more lengthy reports.

At the same time, it is important to acknowledge that the space premium is a major issue for journals and although complex multiple interventions require more detailed attention to reporting of interventions and their implementation, they are unlikely to receive preferential increases in word counts. In addition to making use of more recent opportunities to append supplementary documents to journal publication, one approach to accommodate both space restrictions and adequate reporting of interventions may be to present some information graphically. It has been suggested that graphical representation of complex interventions aids reporting, and hence dons clarity to the interpretation of trial results[83,136]. This was certainly confirmed in the current study, where the complexity of test-treat strategies and extensive variation in clinical settings compounded the disadvantages of opaque reporting. Perera and colleagues[327] have proposed a standardised schema to accompany written descriptions of non–pharmaceutical interventions, developed to elucidate all components of a complex intervention, their timing, and how they contrast with the composition of their comparators. Since test-treatment interventions differ in terms of the key components that influence change, the graphic tool presented in chapter 8 could present a suitable alternative.

Further guidance may be needed to ensure that test-treatment trials are reported consistently and comprehensively. Much of the CONSORT statement, particularly the extension for non–pharmacologic interventions[147] remains relevant to test-treat RCTs and should be followed as is recommended for all trials. However, the conceptual barriers identified would indicate that in order to improve, specific guidance will be needed to disseminate the notion that test-treatment RCTs must evaluate whole management pathways. The present analysis therefore suggests that the requirements for full reporting of test-treatment RCTs are sufficiently different from those of other designs to warrant their own CONSORT extension.

9.7.2 Defining all intended benefits of a new intervention

A key output from this thesis concerns the attention drawn to the multiple and complex ways in which tests exert their effects on patient health. In order to evaluate test-treatment interventions fully, trials are likely to need to measure many more outcomes than is common in standard trials. More patient outcomes need to be measured to ensure that the impact of the strategy on all randomised individuals is assessed, while more process outcomes need to be measured to ensure that the resulting health effects can be interpreted and translated into clinical practice.

As argued in chapter 8, new trials are likely to benefit greatly by carefully developing a comprehensive rationale for how test-treatment interventions are expected to impact on patient health prior to evaluation. This will require extensive consultation between clinical experts, researchers and at times also patient representatives. Although not yet fully validated, this thesis would recommend that the theoretical framework and its associated tool are used as an aid to think through these processes.

9.7.3 Improving methodological quality

The methodological quality of published trials was found to be poor and in urgent need of improvement. While advances to validity can be made by promoting standard trial methods, the thesis highlights that certain methods differ to those needed for treatment trials. Chief amongst these are the requirements to inflate sample sizes by the reclassified fraction in order to properly power for health effects, and the need to consider whether blinding may actually eliminate the desired effects. These points should be disseminated to the research community by the production of methodological guidelines designed specifically for test-treatment trials.

9.7.4 Offsetting the challenges of blinding with process evaluation

Blinding is unlikely to be feasible or appropriate in many instances, however trialists may be able to offset the practical and methodological challenges of blinding by implementing other methodological safeguards. One key solution will involve the close accounting of actual clinical behaviour, including test use, decision–making and treatment use, in order to monitor how actual patient management differs from the intended test-treatment protocol. In this way, investigators may be able to discern between–arm differences due to genuine divergence in diagnostic impact, from those that reflect artefacts of study design. To limit ascertainment bias, the solution will entail ensuring that measurements are standardised so as to minimise systematic differences occurring between arms. Characterising any differences that do occur will also allow the possibility and extent of ascertainment bias to be incorporated into the interpretation of trial findings.

9.8 Future research

Through an analysis of the thesis findings, as well as its limitations, this research reveals several methodological issues that should be researched further.

9.8.1 Methodological research into the role of blinding is urgently needed

The considerable challenges to implementing blinding demand that researchers explore methods to achieve it. Perhaps more urgent, however, is the need to improve our current understanding of the situations in which the absence of blinding truly threatens trial validity. The in–depth examination of trials for the purposes of the feasibility analysis begin to suggest that the consequences of blinding and not blinding are likely to pose a particular dilemma for test-treatment trials. For example, in order to mask treating clinicians we must remove their ability to interpret test results and, depending on the types of tests involved, also remove their ability to make the diagnosis. However, as demonstrated by the framework in chapter 7, differences in clinical behaviour often form part of the causal pathway for test–treatment effects, and so in these situations blinding could serve to remove or alter the treatment effect. These observations begin to suggest that deciding whether or not to blind will require a very careful deliberation by trialists in order to ensure that the correct balance is achieved between minimising the risks to performance and ascertainment bias and ensuring that important intended effects are measured. Research is urgently needed to determine how the lack of blinding impacts on treatment effects, and to establish explicit recommendations for the situations in which blinding is inappropriate.

9.8.2 Methodological development of sample size calculations is needed

The delineation of four separate causal pathways of test-treatment effect may suggest that power calculations might not always require inflation, however this needs to be explored.

The inflation factors proposed by previous authors[176–177,179] assume that health benefits occurring as a result of the test will only occur in the subgroup of patients who will have been managed differently as a result of receiving the new diagnostic intervention. When intended effects are driven by decisional mechanisms, the framework supports this assumption since differences in accuracy and diagnostic/therapeutic yield can only be observed in a small proportion of the study population. The same assumption is contradicted when we consider the remaining 3 causal pathways, in particular temporal or perceptual effects since direct test effects are likely to be compared as trade–offs against the other causal pathways. Most importantly, changes to the timing of test-treatment strategies are likely to be experienced by all randomised participants, therefore it is possible that no correction need be applied. Similarly effects hypothetically driven by changing patient perceptions and experience may only need to be adjusted if the principle cause of effect is perception of the diagnostic category assigned to the patient. These hypotheses need to be evaluated and developed further by statistical analysis of test-treatment RCTs.

9.8.3 Developing the framework tool for trial design and appraisal

As posited in chapter 8, the framework tool could be of use to several aspects of diagnostic test evaluation. However, the version presented in this thesis is a preliminary attempt that must be piloted, discussed more extensively with clinicians, and refined so that it can be useful to those who need to commission, design and appraise test-treatment RCTs.

9.8.4 Tool for appraising the methodological quality of test-treatment RCTs is needed

Since the methods needed to conduct valid trials differ from standard requirements, future appraisals of methodological quality will need to modify the current approach in order to ensure that studies are not inappropriately excluded from syntheses, but also that they are critiqued comprehensively. For example, while the absence of blinding may be considered an exclusion criterion when synthesising treatment trials, this approach is unlikely to be appropriate or sufficient for test-treatment syntheses. These issues require further exploration, perhaps by further empirical analysis of the correlations between treatment effect sizes and failure to implement methodological safeguards, and dissemination to the wider evidence–synthesis community.

9.8.5 Further research must urgently address when RCT evidence is needed and the role of decision models

Due to the current scarcity of RCT evidence, decision–modelling is likely to remain the most common source of diagnostic practice guidelines in the near future. Yet the more complex relationship between tests and health outcomes revealed by the theoretical framework may suggest that our current appreciation of what evidence to link together in decision models is incomplete. The notions of what evidence is required for decision–analysis therefore needs to be revisited in light of the framework. At the same time, it will also be critical to evaluate the validity and reliability of the estimated health effects produced by decision–models.

9.9 Conclusions: How useful are test-treatment RCTs?

This thesis has presented the first empirically–based review of RCT methods for evaluating test-treatment interventions. The examination of this fascinating cohort of published trials provides empirical credence to assertions that test-treatment RCTs are highly complex studies that will be challenging to perform to a reliable standard. The interventions are difficult to capture and translate into protocols; several methodological safeguards traditionally used to eliminate bias are more difficult to implement, and in the case of blinding could be impossible to implement in the majority of comparisons; and the way in which test-treatment strategies impact on patient health are both numerous and highly complicated. Taking all these factors into account, the quality of published trials is certainly very poor and there can be little doubt that they present scarce, uninformative, unreliable and incomplete evidence for the effectiveness of the tests they sought to evaluate. But does this mean that RCTs are not useful to answer such questions? At this stage the answer is not straightforward.

Although most trials generally committed a litany of methodological offenses, many of these threats to their validity and utility are theoretically surmountable given the adequate dissemination of methodological guidance. Interventions can be described with more care; selection and attrition bias can be minimised by improving standard RCT methods; type II errors can be avoided by performing power calculations that correctly adjust for the reclassified fraction, allowing the appropriate number of patients to be recruited; and interventions could be evaluated comprehensive by carefully composing a scientific rationale for how they are expected to impact on patients. The latter point highlights what the author believes is the greatest contribution of this thesis to existing knowledge. Though the conceptual framework and associated tool must now be validated through piloting and discussion amongst the academic and clinical communities, it presents a potential solution to an issue that had been posited as an inherent failure of the test-treatment design: that one cannot disentangle the contribution of the test from that of treatment in observed test–treat effects[59,77,175,187]. Instead, we can now posit that these contributions *can* be distinguished by measuring the workings of the test–treatment strategy – the processes contained within the 'black box'.

Several of the issues highlighted in this thesis could also be rectified by following the MRC's guidance for developing and piloting complex interventions[83], before proceeding to a full trial. More conscientious development of new test-treatment interventions, and more careful mapping out of control interventions, would significantly improve the ability to convey the nature of what is to be evaluated in trial protocols. Piloting these interventions could serve to check that patient numbers are recruitable, provide insight into how much attrition might be expected, whilst also checking whether predicted effect sizes are realistic.

On the other hand, significant obstacles remain that would indicate long–term RCTs will not be the most appropriate method in some diagnostic settings. Chief amongst these is frequent inability to blind outcome assessors to subjective outcomes, a proven source of bias in treatment trials. While further research is needed to establish whether test-treatment interventions risk similar bias, it could be that trials are unlikely to provide reliable answers when patients can't be blinded and subjective outcomes must be measured. Arguably the biggest threat to the utility of the RCT probably lays in the practical ability to recruit the number of patients necessary to avoid type II errors, the willingness of funders to pay for these costs and the enthusiasm of policy–makers to accept that trial evidence will take far longer to produce than is common for treatment RCTs. When the sample sizes required to measure downstream health outcomes are considered impractical, one solution could be to use RCTs as tools to ensure that the workings of test-treatment strategies operate as intended, since by evaluating intermediate outcomes trials will have greater power to detect effects, and use the resulting data to model the impacts to downstream patient health. Then again, models may not be able to incorporate all relevant mechanisms in particularly complex comparisons, nor will they be able to account for unexpected or unpredictable effects. A key advantage of the RCT is that it can reflect the test-treatment process accurately. In reality patient management decisions occur as part of an iterative process between test results and available treatment options, and if trials are designed and conducted well they provide unique tools to capture this process in its entirety and relate it directly to observed changes in health outcomes.

Trials have the potential to be very useful instruments for evaluating whether tests do more good than harm to patients. Though highly challenging, these designs can prove to be both reliable and informative. However, it must be acknowledged that these complex designs will not be suited to all comparisons, thus placing the RCT at the summit of a rigid hierarchy of evidence is also unlikely to be the best way forward.

Bibliographic references

1. Cochrane, AL. The history of the measurement of ill health. Int J Epidemiol 1972;1: 89–92.

2. Verrilli D, Welch HG. The impact of diagnostic testing on therapeutic interventions. JAMA 1996;275:1189–1191.

3. Sattar N, Welsh P, Panarelli M, Forouhi NG. Increasing requests for vitamin D measurement: costly, confusing, and without credibility. Lancet 2012;379:95–96.

4. Pines JM. Trends in the rates of radiography use and important diagnoses in emergency department patients with abdominal pain. Med Care 2009;47:782–786.

5. Miglioretti DL, Smith-Bindman R. Overuse of computed tomography and associated risks. Am Fam Physician 2011;83:1252–1254.

6. Hunink MGM, Krestin GP. Study design for concurrent development, assessment, and implementation of new diagnostic imaging technology. Radiology 2002;222:604-614.

7. Shah BR, Patel MR, Peterson ED, Douglas PS. Defining optimal research study design for cardiovascular imaging using computed tomography angiography as a model. Am J Cardiol 2008;102:943–948.

8. Bruns DE, Boyd JC. Assessing the impact of biomarkers on patient outcome: an obligatory step. Scand J Clin Lab Invest Suppl 2010;242:85–89.

9. Mitka M. Research offers only a limited view of imaging's effect on patient outcomes. JAMA 2010;303:599–600.

10. Stone JH. Incidentalomas – clinical correlation and translational science required. N Engl J Med 2006; 354:2748–2749.

11. Haynes RB, You JJ. The architecture of diagnostic research. In: Knottnerus JA, Buntinx F, editors. The evidence base of clinical diagnosis. 2nd Edition. Oxford: Wiley–Blackwell; 2009. p.20–41.

12. Feinstein AR Misguided efforts and future challenges for research on "diagnostic tests". J Epidemiol Community Health 2002;56:330–332.

13. Price CP, Christenson RH. Evaluating new diagnostic technologies: perspectives in the UK and US. Clin Chem 2008;54(9):1421–1423.

14. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991;11:88–94.

15. Jarvik JG. Fundamentals of clinical research for radiologists. Am J Roentgenol 2001;176:873–877.

16. Agency for Healthcare Research and Quality. Methods guide for medical test reviews. Rockville, MD; 2010. Available at URL: http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?pageaction=displayProduct&productID=454 [Accessed 30th November 2010].

17. Schünemann HJ, Oxman AD, Brozek J, Glasziou P, Jaeschke R, Vist GE, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ 2008;336:1106e10.

18. National Institute for Health and Clinical Excellence. Diagnostics assessment programme manual. 2011. Available at URL: www.nice.org.uk/media/A0B/97/DAPManualFINAL.pdf [Accessed 13th January 2012].

19. Lord SJ, Irwig L, Bossuyt PMM. Evaluating new tests: when can comparative evidence of test accuracy and other intermediate outcomes be used as an alternative to randomized controlled trials. Med Decis Making 2009;29:E1–E12.

20. Valk PE. Randomized controlled trials are not appropriate for imaging technology evaluation. J Nucl Med 2000;41:1125–1126.

21. Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006;144:850–855.

22. Lijmer JG, Bossuyt PMM. Various randomized designs can be used to evaluate medical tests. J Clin Epidemiol 2009;62:364-373.

23. Black WC. Randomized clinical trials for cancer screening: rationale and design considerations for imaging tests. J Clin Oncol 2006;24:3252–3260.

24. Tatsioni A, Zarin DA, Aronson N, Samson DJ, Flamm CR, Schmid C, et al. Challenges in systematic reviews of diagnostic technologies. Ann Intern Med 2005;142:1048–1055.

25. National Institute for Health and Clinical Excellence [Internet]. Available from URL: www.nice.org.uk/aboutnice/whatwedo/aboutdiagnosticassessment/diagnosticassessmentprogramme.jsp [Accessed 30th July 2012].

26. Goodacre S, Bradburn M, Fitzgerald P, Cross E, Collinson P, Gray A, et al. The RATPAC (Randomised Assessment of Treatment using Panel Assay of Cardiac markers) trial: a randomised controlled trial of point-of-care cardiac markers in the emergency department. Health Technol Assess 2011;15:1-102.

27. Delaney BC, Qume M, Moayyedi P, Logan RF, Ford AC, Elliott C, et al. Helicobacter pylori test and treat versus proton pump inhibitor in initial management of dyspepsia in primary care: multicentre randomised controlled trial (MRC-CUBE trial). BMJ 2008;336:651–654.

28. Kerry S, Hilton S, Patel S, Dundas D, Rink E, Lord J. Routine referral for radiography of patients presenting with low back pain: is patients' outcome influenced by GPs' referral for plain radiography? Health Technol Assess 2000;4:1-119.

29. Kendrick D, Fielding K, Bentley E, Miller P, Kerslake R, Pringle M. The role of radiography in primary care patients with low back pain of at least 6 weeks duration: a randomised (unblinded) controlled trial. Health Technol Assess 2001;5:1–69.

30. Liston J, Wilson R, editors. Clinical guidelines for breast cancer screening assessment. 3rd edition. June 2010; NHSBSP Publication No 49. Sheffield: NHS Cancer Screening Programmes. Available from URL: http://www.cancerscreening.nhs.uk/breastscreen/publications/numbered-index.html [Accessed 10th October 2012].

31. Sabhan D, Pylypiw L. Guidelines for Newborn Blood Spot Sampling. Update, February 2012. UK Newborn Screening Programme Centre. Available from URL: www.newbornbloodspot.screening.nhs.uk [Accessed 10th August 2012].

32. Little P, Turner S, Rumsby K, Warner G, Moore M, Lowes JA, et al. Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort and qualitative study. Health Technol Assess 2009;13:1–73.

33. Bryan S, Weatherburn G, Bungay H, Hatrick C, Salas C, Parry D, et al. The cost-effectiveness of magnetic resonance imaging for investigation of the knee joint. Health Technol Assess 2001;5:1–95.

34. Maziak DE, Darling GE, Inculet RI, Gulenchyn KY, Driedger AA, Ung YC, et al. Positron emission tomography in staging early lung cancer: a randomized trial. Ann Intern Med 2009;151:221–8.

35. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ 2009;338:b375.

36. Hingorani A, van der Windt D, Riley RD, Abrams K, Moons K, Steyerberg E, Schroter S, Altman DG, Hemmingway H. Prognosis Research Strategy (PROGRESS) 4: stratified medicine research. BMJ (In Press).

37. Morey SS. American Urological Association issues guidelines on the management of bladder cancer. Am Fam Physician 2000;61:3734–3736.

38. Clar C, Barnard K, Cummins E, Royle P, Waugh N, Aberdeen Health Technology Assessment Group. Self-monitoring of blood glucose in type 2 diabetes: systematic review. Health Technol Assess 2010;14:1–140.

39. Glasziou PP, Irwig L, Aronson JK, editors. Evidence-based medical monitoring: from principles to practice. Oxford: Blackwell Publishing; 2008.

40. Riley RD, Hayden J, Moons KGM, Steyerberg E, Abrams KR, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med (in-press).

41. Irwig L, Houssami N, Armstrong B, Glasziou P. Evaluating new screening tests for breast cancer. BMJ 2006;332:678–679.

42. UK National Screening Committee [internet]. Available from URL: http://www.screening.nhs.uk/criteria [Accessed 25th August 2012].

43. Morrison AS. The effects of early treatment, lead time and length bias on the mortality experienced by cases detected by screening. Int J Epidemiol 1982;11:261–267.

44. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. Can Med Assoc J 1986;134:587–594.

45. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clin Radiol 1995;50:513–518.

46. Pearl WS. A hierarchical outcomes approach to test assessment. Ann Emerg Med 1999;33:77–84.

47. Pepe MS. Evaluating technologies for classification and prediction. Stat Med 2005;24:3687–3696.

48. Kent DL, Larson EB. Disease, level of impact, and quality of research methods. Three dimensions of clinical efficacy assessment applied to magnetic resonance imaging. Invest Radiol 1992;27:245–254.

49. Van den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on technical and diagnostic accuracy, impact on patient outcome and cost-effectiveness is needed. J Clin Epidemiol 2007;60:1116–1122.

50. Silverstein MD, Boland BJ. Conceptual framework for evaluating laboratory tests: case-finding in ambulatory patients. Clin Chem 1994;40:1621–1627.

51. Freedman LS. Evaluating and comparing imaging techniques: a review and classification of study designs. Br J Radiol 1987;60:1071–1081.

52. Taylor CR, Elmore JG, Sun K, Inouye SK. Technology assessment in diagnostic imaging: a proposal for a phased approach to evaluating radiology research. Invest Radiol 1993;28:155–161.

53. van der Schouw YT, Verbeek AL, Ruijs SH. Guidelines for the assessment of new diagnostic tests. Invest Radiol 1995;30:334–340.

54. Zweig MH, Robertson EA. Why we need better test evaluations. Clin Chem 1982;28:1272–1276.

55. Sackett DL, Haynes RB, The architecture of diagnostic research. BMJ 2002;324:539–541.

56. Gluud C, Gluud LL. Evidence based diagnostics. BMJ 2005;330:724–726.

57. Lijmer JG, Leeflang M and Bossuyt PMM. Proposals for a phased evaluation of medical tests. Med Decis Making 2009;29:E13–E21.

58. Whiting P, Harbord R, de Salis I, Egger M, Sterne J. Evidence-based diagnosis. J Health Serv Res Policy 2008;13(suppl 3):57–63.

59. Normand ST. Utilization of diagnostic tests: assessing appropriateness. Acad Radiol 1999;6(suppl 1):S47–S51.

60. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ 2002;324:669–671.

61. Bossuyt PMM. Study design and quality of evidence. In: Price CP, Christenson RH, editors. Evidence–based laboratory medicine from principles to practice. Washington, DC: AACC Press; 2003. p.75–92.

62. Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-Ray techniques: Public Health Rep 1947;62:1432–1449.

63. Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ 2006;332:1089–1092.

64. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ 1994;308:1552.

65. Ransohoff DF. Challenges and opportunities in evaluating diagnostic tests. J Clin Epidemiol 2002;55:1178–1182.

66. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 2004;140:189–202.

67. Goehring C, Perrier A, Morabia A. Spectrum bias: a quantitative and graphical analysis of the variability of medical diagnostic test performance. Stat Med 2004;23:125–135.

68. Lord SJ, Staub LP, Bossuyt PM, Irwig LM. Target practice: choosing target conditions for test accuracy studies that are relevant to clinical practice. BMJ 2011;343:d4684.

69. Rutjes AW, Reitsma JB, Di Nisio M, Smidt N, van Rijn JC, Bossuyt PM. Evidence of bias and variation in diagnostic accuracy studies. CMAJ 2006;174:469–476.

70. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999;282:1061–1066.

71. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529–536.

72. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41–44.

73. Haynes RB, Wilczynski NL. Optimal search strategies for retrieving scientifically strong studies of diagnosis from Medline: analytical survey. BMJ 2004;328:1040.

74. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM, Cochrane Diagnostic Test Accuracy Working Group. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008;149:889–897.

75. Neumann PJ, Tunis SR. Medicare and medical technology--the growing demand for relevant outcomes. N Engl J Med 2010;362:377–379.

76. Mol BW, Lijmer JG, Evers JL, Bossuyt PM. Characteristics of good diagnostic studies. Semin Reprod Med 2003;21:17–25.

77. Revicki DA, Yabroff KR, Shikiar R. Outcomes Research in Radiologic Imaging: Identification of barriers and potential solutions. Acad Radiol 1999;6(suppl 1):S20–S28.

78. Harris RP, Helfand M, Woolf SH, Lohr KN, Mulrow CD, Teutsch SM, et al. Current methods of the US Preventive Services Task Force: a review of the process. Am J Prev Med 2001;20(3 suppl):21-35.

79. Bossuyt PM, Lijmer JG, Mol BW. Randomised comparisons of medical tests: sometimes invalid, not always efficient. Lancet 2000;356:1844–1847.

80. Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7:1–173.

81. Altman DG. Randomisation, essential for reducing bias. BMJ 1991;302:1481–1482.

82. de Bree R, van der Putten L, Hoekstra OS, Kuik DJ, Uyl-de Groot CA, van Tinteren H, et al. A randomized trial of PET scanning to improve diagnostic yield of direct laryngoscopy in patients with suspicion of recurrent laryngeal carcinoma after radiotherapy. Contemp Clin Trials 2007;28:705–712.

83. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: new guidance. Medical Research Council; 2008. Available at URL: www.mrc.ac.uk/complexinterventionsguidance [Accessed 12th December 2009].

84. Hunink MGM. Decision making in the face of uncertainty and resource constraints: examples from trauma imaging. Radiology 2005;235:375–383.

85. Plevritis SK. Decision analysis and simulation modeling for evaluating diagnostic tests on the basis of patient outcomes. Am J Roentgenol 2005;185:581–590.

86. Howard K, Lord SJ, Speer A, Gibson RN, Padbury R, Kearney B. Value of magnetic resonance cholangiopancreatography in the diagnosis of biliary abnormalities in postcholecystectomy patients: a probabilistic cost-effectiveness analysis of diagnostic strategies. Int J Technol Assess Health Care 2006;22:109–118.

87. Trikalinos TA, Siebert U, Lau J. Decision-analytic modeling to evaluate benefits and harms of medical tests: uses and limitations. Med Decis Making 2009;29:E22–E29.

88. Schaafsma JD, van der Graaf Y, Rinkel GJ, Buskens E. Decision analysis to complete diagnostic research by closing the gap between test characteristics and cost-effectiveness. J Clin Epidemiol 2009;62:1248–1252.

89. Gregor JC, Ponich TP, Detsky AS. Should ERCP be routine after an episode of “idiopathic” pancreatitis? A cost-utility analysis. [see comment]. Gastrointest Endosc 1996;44:118–123.

90. Bass EB, Steinberg EP, Pitt HA, et al. Cost-effectiveness of extracorporeal shock-wave lithotripsy versus cholecystectomy for symptomatic gallstones. Gastroenterology 1991;101:189–199.

91. Phelps CE, Mushlin AI. Focusing technology assessment using medical decision theory. Med Decis Making 1988;8:279–289.

92. Sutton AJ, Cooper NJ, Goodacre S, Stevenson M. Integration of meta-analysis and economic decision modeling for evaluating diagnostic tests. Med Decis Making 2008;28:650-67.

93. Staub LP, Lord SJ, Simes RJ, Dyer S, Houssami N, Chen RYM, Irwig L. Using patient management as a surrogate for patient health outcomes in diagnostic test evaluation. BMC Med Res Methodol 2012;12:12–21.

94. van Belle A, Büller HR, Huisman MV, Huisman PM, Kaasjager K, Kamphuisen PW, et al. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. JAMA 2006;295:172–179.

95. Medical Services Advisory Committee. Guidelines for the assessment of diagnostic technologies. MSAC;2005. Available from URL: www.msac.gov.au/ [Accessed 3rd September 2007].

96. Bhopal R. Concepts of epidemiology. 2nd ed. Oxford: Oxford University Press; 2008.

97. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 1995;273:408–412.

98. Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 1998;352:609–13.

99. Jüni P, Tallon D, Egger M. `Garbage in - garbage out'? Assessment of the quality of controlled trials in meta-analyses published in leading journals [abstract]. *Proceedings of the 3rd symposium on systematic reviews: beyond the basics,* St Catherine's College, Oxford. Oxford: Centre for Statistics in Medicine, 2000:19.

100. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of Internal Medicine 2001;135:982-9.

101. Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973–2982.

102. Egger M, Juni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Empirical study. Health Technol Assess 2003;7:1–76.

103. Contopoulos-Ioannidis DG, Gilbody SM, Trikalinos TA, Churchill R, Wahlbeck K, Ioannidis JP. Comparison of large versus smaller randomized trials for mental health-related interventions. Am J Psychiatry 2005;162:578–584.

104. Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Stat Med 2007;26:2745–2758.

105. Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Bürgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ 2009;339:b3244.

106. Jüni P, Altman DG, Egger M. Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 2001;323:42–46.

107. Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336:601–605.

108. Savović J, Jones H, Altman D, Harris R, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess 2012;16:1–82.

109. Sterne JAC, Jüni P, Schulz KF, Altman DG, Bartlett C, Egger M. Statistical methods for assessing the influence of study characteristics on treatment effects in 'meta-epidemiological' research. Stat Med 2002;21:1513–1524.

110. Gluud LL, Thorlund K, Gluud C, Woods L, Harris R, Sterne JA. Correction: reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Annals of Internal Medicine 2008;149:219.

111. Als-Nielsen B, Chen W, Gluud LL, Siersma V, Hilden J, Gluud C. Are trial size and reported methodological quality associated with treatment effects? Observational study of 523 randomised trials [abstract]. 12th Cochrane Colloquium: Bridging the Gaps; 2004 Oct 2-6; Ottawa, Ontario: Canada. Available from URL: http://www.cochrane.org [Accessed 13th February 2011].

112. Altman DG and Schulz KF. Concealing treatment allocation in randomised trials. BMJ 2001;323:446–447.

113. Schulz KF and Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002;359:614–618.

114. Schulz KF. Subverting randomization in controlled trials. JAMA 1995;274:1456–1458.

115. Schulz KF, Altman DG, Moher D. Allocation concealment in clinical trials. JAMA 2002;288:2406–2407.

116. Jüni P, Egger M. Allocation concealment in clinical trials. JAMA 2002;288:2407–2408.

117. Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM. Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. BMJ 2005;330:1057–1058.

118. Hopewell S, Dutton S, Yu L–M, Chan A–W, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ 2010;340:c723.

119. Akl EA, Sunc X, Bussec JW, Johnston BC, Briel M, Mull S et al. Specific instructions for estimating unclearly reported blinding status in randomized trials were reliable and valid. J Clin Epidemiol 2012;65:262–267.

120. Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001;285:2000-2003.

121. Moher D, Hopewell S, Schultz, Montori V, Gøtzsche P, Devereaux PJ, et al. CONSORT 2010 Explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 2010;340:c869.

122. Schulz KF, Grimes DA. Sample size slippages in randomised trials: exclusions and the lost and wayward. Lancet 2002;359:781–785.

123. Higgins JPT, Altman DG, Sterne JAC, Cochrane Statistical Methods Group, Cochrane Bias Methods Group, editors. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. The Cochrane Collaboration; 2011. Version 5.1.0 [updated March 2011]. Available from URL: www.cochrane-handbook.org [Accessed 24th November 2012].

124. Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ. Assessing the impact of attrition in randomized controlled trials. J Clin Epidemiol 2010;63:1264–1270.

125. Tierney JF, Stewart LA. Investigating patient exclusion bias in meta–analysis. Int J Epidemiol 2005;34:79–87.

126. Porta N, Bonet C, Cobo E. Discordance between reported intention-to-treat and per protocol analyses. J Clin Epidemiol 2007;60:663–669.

127. Schulz KF, Grimes DA, Altman DG, Hayes RJ. Blinding and exclusions after allocation in randomised controlled trials: survey of published parallel group trials in obstetrics and gynaecology. BMJ 1996;312:742–744.

128. Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. Lancet 2005;365:1348–1353.

129. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995;311:485.

130. Charles P, Giraudeau B, Dechartres A, Baron G, Ravaud P. Reporting of sample size calculation in randomised controlled trials: review. BMJ 2009;338:b1732.

131. Tsay M, Yang Y. Bibliometric analysis of the literature of randomized controlled trials J Med Libr Assoc 2005;93:450–458.

132. Heneghan C. How many randomized trials are published every year? [internet]. 2010 Available from URL: http://blogs.trusttheevidence.net/carl-heneghan/how-many-randomized-trials-are-published-each-year [Accessed 25th August 2012].

133. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med 2010; 7:e1000326.

134. Huwiler-Müntener K, Jüni P, Junker C, Egger M. Quality of reporting of randomized trials as a measure of methodologic quality JAMA 2002;287:2801–2804.

135. Rothwell, PM. External validity of randomised controlled trials: “To whom do the results of this trial apply?” Lancet 2005;365:82–93.

136. Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ 2008;336:1472–1474.

137. Glasziou P, Chalmers I, Altman DG, Bastian H, Boutron I, Brice A, et al. Taking healthcare interventions from trial to practice. BMJ 2010;341:c3852.

138. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, et al. Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 1996;276:637–639.

139. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990;335:149–153.

140. The CONsolidated Standards Of Reporting Trials [internet]. How CONSORT began. 2012. Available at URL: http://www.consort-statement.org/about-consort/history/ [Accessed 13th August 2012].

141. Moher D, Schulz KF, Altman DG, CONSORT Group (Consolidated Standards of Reporting Trials). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. Ann Intern Med 2001;134:657–662.

142. Schulz KF, Altman DG, Moher D, CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomized trials. Ann Intern Med 2010;152:1–8.

143. The CONsolidated Standards Of Reporting Trials [internet]. CONSORT endorsers. 2012. Available at URL: http://www.consort-statement.org/about-consort/consort-endorsement/consort-endorsers---journals/ [Accessed 13th August 2012].

144. Campbell MK, Elbourne DR, Altman DG, CONSORT Group. CONSORT statement: extension to cluster randomised trials. BMJ 2004;328:702–708.

145. Piaggio G, Elbourne DR, Altman DG, Pocock SJ, Evans SJ, CONSORT Group. Reporting of noninferiority and equivalence randomized trials: an extension of the CONSORT statement. JAMA 2006;295:1152–1160.

146. Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ 2008;337:a2390.

147. Boutron I, Moher D, Altman DG, Schulz KF, Ravaud P, CONSORT Group. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 2008;148:295–309.

148. Gagnier JJ, Boon H, Rochon P, Moher D, Barnes J, Bombardier C, et al. Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. Ann Intern Med 2006;144:364–367.

149. MacPherson H, Altman DG, Hammerschlag R, Youping L, Taixiang W, White A, et al. Revised STandards for Reporting Interventions in Clinical Trials of Acupuncture (STRICTA): extending the CONSORT statement. PLoS Med 2010;7:e1000261

150. Ioannidis JP, Evans SJ, Gøtzsche PC, O'Neill RT, Altman DG, Schulz K, et al. Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 2004;141:781–788.

151. Hopewell S, Clarke M, Moher D, Wager E, Middleton P, Altman DG, et al. CONSORT for reporting randomised trials in journal and conference abstracts. Lancet 2008;371:281–283.

152. Chan A–W, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. Lancet 2005;365:1159–1162.

153. Mills EJ, Wu P, Gagnier J, Devereaux PJ. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. Contemp Clin Trials 2005;26:480–487.

154. Cochrane Library Central Register of Controlled Trials [Internet]. Available from URL: http://www.thecochranelibrary.com [Accessed 24th November 2012].

155. McCulloch P, Taylor I, Sasako M, Lovett B, Griffin D. Randomised trials in surgery: problems and possible solutions. BMJ 2002;324:1448–1451.

156. Delaney A, Angus DC, Bellomo R, Cameron P, Cooper DJ, Finfer S, et al. Bench-to-bedside review: the evaluation of complex interventions in critical care. Crit Care 2008;12:210–219.

157. Walwyn R, Wessely S. RCTs in psychiatry: challenges and the future. Epidemiol Psichiatr Soc 2005;14:127–131.

158. Whiting P, Westwood M, Bojke L, Palmer S, Richardson G, Cooper J, Watt I, Glanville J, Sculpher M, J Kleijnen. Clinical effectiveness and cost-effectiveness of tests for the diagnosis and investigation of urinary tract infection in children: a systematic review and economic model. Health Technol Assess 2006;10:1–154.

159. Collins R, Burch J, Cranny G, Aguiar-Ibanez R, Craig D, Wright K, Berry E, Gough M, Kleijnen J, Westwood M. Duplex ultrasonography, magnetic resonance angiography, and computed tomography angiography for diagnosis and assessment of symptomatic, lower limb peripheral arterial disease: systematic review. BMJ 2007;334:1257–1265.

160. Dyer SM, Levison DB, Chen RY, Lord SJ, Blamey S. Systematic review of the impact of endoscopic ultrasound on the management of patients with esophageal cancer. Int J Technol Assess Health Care 2008;24:25–35.

161. Kuukasjärvi P, Nordhausen K, Malmivaara A. Reanalysis of systematic reviews: The case of invasive strategies for acute coronary syndromes. Int J Technol Assess Health Care 2006;22:484–496.

162. Doan Q, Enarson P, Kissoon N, Klassen TP, Johnson DW. Rapid viral diagnosis for acute febrile respiratory illness in children in the Emergency Department. Cochrane Database of Systematic Reviews 2009;7(4):CD006452.

163. Jørgensen H, Jensen CH, Dirks J. Does prehospital ultrasound improve treatment of the trauma patient? A systematic review. Eur J Emerg Med 2010;17:249–253.

164. Hislop J, Quayyum Z, Flett G, Boachie C, Fraser C, Mowatt G. Systematic review of the clinical effectiveness and cost-effectiveness of rapid point-of-care tests for the detection of genital chlamydia infection in women and men. Health Technol Assess. 2010;14:1–97.

165. McKenna C, Wade R, Faria R, Yang H, Stirk L, Gummerson N, et al. EOS 2D/3D imaging system. 2011. NICE Diagnostic Assessment Report 1. NIHR HTA Project No. 10/67/01. Available from URL: http://guidance.nice.org.uk/DT/Published [Accessed 30th July 2012].

166. Sharma P, Boyers D, Boachie C, Stewart F, Miedzybrodzka Z, Simpson W, et al. Elucigene FH20 and LIPOchip for the diagnosis of familial hypercholesterolaemia. 2011. NICE Diagnostic Assessment Report 2. NIHR HTA Project No. 10/70/01. Available from URL: http://guidance.nice.org.uk/DT/Published [Accessed 30th July 2012].

167. Westwood M, Maiwenn A, Burgers L, Redekop K, Lhachimi S, Armstrong N, et al. Computed tomography (CT) scanners for cardiac imaging – Somatom Definition Flash, Aquilion One, Brilliance iCT and Discovery CT750 HD. 2011. NICE Diagnostic Assessment Report 3. NIHR HTA Project No. 10/107/01. http://guidance.nice.org.uk/DT/Published [Accessed 30th July 2012].

168. Ritchie G, Glanville J, Lefebvre C. Do published search filters to identify diagnostic test accuracy studies perform adequately? Health Info Libr J 2007;24:188–192.

169. Kastner M, Wilczynski NL, McKibbon AK, Garg AX, Haynes RB. Diagnostic test systematic reviews: bibliographic search filters ("Clinical Queries") for diagnostic accuracy studies perform well. J Clin Epidemiol 2009;62:974–981.

170. The Cochrane Collaboration [internet]. Available at URL: http://www.cochrane.org/news/blog/embase-introduces-diagnostic-test-accuracy-study-indexing-term [Accessed 5th May 2012].

171. Doust JA, Pietrzak E, Sanders S, Glasziou PP. Identifying studies for systematic reviews of diagnostic tests was difficult due to the poor sensitivity and precision of methodologic filters and the lack of information in the abstract. J Clin Epidemiol 2005;58:444–449.

172. Bachmann LM, Coray R, Estermann P, Ter Riet G. Identifying Diagnostic Studies in MEDLINE: Reducing the Number Needed to Read. J Am Med Inform Assoc 2002;9:653–658.

173. Leeflang MMG, Scholten RJPM, Rutjes AWS, Reitsma JB, Bossuyt PMM. Use of methodological search filters to identify diagnostic accuracy studies can lead to the omission of relevant studies. J Clin Epidemiol 2006;59:234–240.

174. Bruns DE. Laboratory-related outcomes in healthcare. Clin Chem 2001 Aug;47:1547–1552.

175. Blackmore CC, Black WC, Jarvik JG, Langlotz CP. A Critical Synopsis of the Diagnostic and Screening Radiology Outcomes Literature. Acad Radiol 1999;6(suppl 1):S8–S18.

176. Deeks JJ. Assessing outcomes following tests. In: Price CP, Christenson RH, editors. Evidence-based laboratory medicine: principles, practice and outcomes. 2nd edition. Washington DC: AACC Press; 2007. p.95–111.

177. Pletcher MJ, Pignone M. Evaluating the clinical utility of a biomarker: a review of methods for estimating health impact. Circulation 2011;123:1116–1124.

178. Campbell M, Fitzpatrick R, Haines A, Kinmonth AL, Sandercock P, Spiegelhalter D, et al. Framework for design and evaluation of complex interventions to improve health. BMJ 2000;321:694–696.

179. De Bono M, Fawdry RDS, Lilford RJ. Size of trials for evaluation of antenatal tests of fetal wellbeing in high risk pregnancy. J Perinat Med 1990;18:77–87.

180. Lijmer JG, Bossuyt PM. Chapter 4: Diagnostic testing and prognosis: the randomised controlled trial in diagnostic research. In: Knottnerus JA, editor. The evidence–base of clinical diagnosis. London, UK: BMJ Books; 2002. p.61–80.

181. Vis JY, Wilms FF, Oudijk MA, Bossuyt PMM, van der Post JAM, Grobman WA, et al. Why were the results of randomized trials on the clinical utility of Fetal

Fibronectin negative? A systematic review of their study Am J Perinatol 2011;28:145–150.

182. Hawe P, Shiell A, Riley T. Complex interventions: how "out of control" can a randomised controlled trial be? BMJ 2004;328:1561–1563.

183. Stolper E, Van de Wiel M, Van Royen P, Van Bokhoven M, Van der Weijden T, Dinant GJ. Gut feelings as a third track in general practitioners' diagnostic reasoning. J Gen Intern Med 2011;26:197–203.

184. Doubilet P, Herman PG. Interpretation of radiographs: effects of clinical history. AJR Am J Roentgenol 1981;137:1055–1058.

185. Clarke M. Standardising outcomes for clinical trials and systematic reviews. Trials. 2007;8:39–42.

186. Sinha I, Jones L, Smyth RL, Williamson PR. A systematic review of studies that aim to determine which outcomes to measure in clinical trials in children. PLoS Med 2008;5:e96.

187. Fineberg HV. Evaluation of Computed Tomography: Achievement and Challenge. AJR Am J Roentgenol 1978;131:1–4.

188. Einstein AJ. Effects of radiation exposure from cardiac imaging: how good are the data? J Am Coll Cardiol 2012;59:553–565.

189. Froehlich F, Gonvers JJ, Vader JP, Dubois RW, Burnand B. Appropriateness of gastrointestinal endoscopy: risk of complications. Endoscopy 1999;31:684–686.

190. Howard LM, Wessely S. Reappraising reassurance–the role of investigations. J Psychosom Res 1996;41:307–311.

191. Silberstein SD. Chronic daily headache. J Am Osteopath Assoc 2005;105(4 suppl 2):23S–29S.

192. Howard L, Wessely S, Leese M, Page L, McCrone P, Husain K, et al. Are investigations anxiolytic or anxiogenic? A randomised controlled trial of neuroimaging to provide reassurance in chronic daily headache. J Neurol Neurosurg Psychiatry 2005;76:1558–1564.

193. Bossuyt PMM, Lijmer JG. Traditional health outcomes in the evaluation of diagnostic tests. Acad Radiol. 1999;6(suppl 1):S77–S80.

194. Tsushima Y, Aoki J, Endo K. Contribution of the diagnostic test to the physician's diagnostic thinking: new method to evaluate the effect. Acad Radiol 2003;10:751–755.

195. Dickersin K, Manheimer E, Wieland S, Robinson KA, Lefebvre C, MacDonald S, Central Development Group. Development of the Cochrane Collaboration’s Central Register of Controlled Clinical Trials. Eval Health Prof 2002;25:38–64.

196. Lefebvre C, Eisinga A, McDonald S, Paul N. Enhancing access to reports of randomized trials published world-wide – the contribution of EMBASE records to the Cochrane Central Register of Controlled Trials (CENTRAL) in The Cochrane Library. Emerg Themes Epidemiol 2008;5:13–26.

197. Haynes RB, Wilczynski N, McKibbon KA, et al. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1:447–458.

198. Lefebvre C, Manheimer E, Glanville J. Chapter 6: Searching for studies. In: Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.0.1 [updated September 2008]. The Cochrane Collaboration, 2008. Available from URL: www.cochrane-handbook.org [Accessed 18th November 2008].

199. Saltvedt S, Almström H, Kublickas M, Valentin L, Bottinga R, Bui TH, et al. Screening for Down syndrome based on maternal age or fetal nuchal translucency: a randomized controlled trial in 39,572 pregnancies. Ultrasound Obstet Gynecol 2005;25:537–545.

200. Vretzakis G, Ferdi E, Argiriadou H, Papaziogas B, Mikroulis D, Lazarides M, et al. Influence of bispectral index monitoring on decision making during cardiac anesthesia. J Clin Anesth 2005;17:509–516.

201. East CE, Begg L, Colditz PB. Fetal pulse oximetry for fetal assessment in labour. Cochrane Database of Systematic Reviews 2007;(2):CD004075.

202. Dudeck O, Teichgraeber U, Podrabsky P, Lopez HE, Soerensen R, Ricke J. A randomized trial assessing the value of ultrasound-guided puncture of the femoral artery for interventional investigations. Int J Cardiovasc Imaging 2004;20:363–368.

203. Tozzi R, Malur S, Koehler C, Schneider A. Laparoscopy versus laparotomy in endometrial cancer: first analysis of survival of a randomized prospective study. J Minim Invasive Gynecol 2005;12:130–136.

204. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17:243e64.

205. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. Am J Epidemiol 1992;135:1060–1067.

206. Wittes J, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. J Chronic Dis 1968;21:287–301.

207. Spoor P, Airey M, Bennett C, Greensill J, Williams R. Use of the capture-recapture technique to evaluate the completeness of systematic literature searches. BMJ 1996;313:342–343.

208. Edwards P, Clarke M, DiGuiseppi C, Pratap S, Roberts I, Wentz R. Identification of randomized controlled trials in systematic reviews: accuracy and reliability of screening records. Stat Med 2002; 21:1635–1640.

209. Kastner M, Straus SE, McKibbon KA, Goldsmith CH. The capture-mark-recapture technique can be used as a stopping rule when searching in systematic reviews. J Clin Epidemiol 2009;62:149–157.

210. Bennett DA, Latham NK, Stretton C, Anderson CS. Capture-recapture is a potentially useful method for assessing publication bias. J Clin Epidemiol 2004;57:349–357.

211. Cormack RM. Interval Estimation for Mark-Recapture Studies of Closed Populations. Biometrika 1992;48:567–576.

212. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. Stat Med 1984;3:287–291.

213. Kirkwood BR, Sterne, JAC. Essentials of Medical Statistics. 2nd Edition. Oxford: Wiley–Blackwell.

214. Goodacre S, Dixon S. Is a chest pain observation unit likely to be cost effective at my hospital? Extrapolation of data from a randomised controlled trial. Emerg Med J 2005;22:418–422.

215. Brealey SD, DAMASK (Direct Access to Magnetic Resonance Imaging: Assessment for Suspect Knees) Trial Team. Influence of magnetic resonance of the knee on GPs' decisions: a randomised trial. Br J Gen Pract 2007;57:622–629.

216. Little P. Development and randomised controlled trial of dipsticks and diagnostic algorithms for the management of UTI. Current Controlled Trials ISRCTN Register [internet]. Available from URL: www.controlled-trials.com/isrctn/ [Accessed 23rd February 2010].

217. Breidthardt T, Laule K, Strohmeyer AH, Schindler C, Meier S, Fischer M, et al. Medical and economic long-term effects of B-type natriuretic peptide testing in patients with acute dyspnea. Clin Chem 2007;53:1415–1422.

218. Schietroma M, Cappelli S, Carlei F, Pescosolido A, Lygidakis NJ, Amicucci G. "Acute abdomen": early laparoscopy or active laparotomic-laparoscopic observation? Hepatogastroenterology 2007;54:1137–1141.

219. Peters NH, Borel Rinkes IH, Mali WP, van den Bosch MA, Storm RK, Plaisier PW, et al. Breast MRI in nonpalpable breast lesions: a randomized trial with diagnostic and therapeutic outcome - MONET - study. Trials 2007;8:40–47.

220. Nucifora G, Badano LP, Sarraf-Zadegan N, Karavidas A, Trocino G, Scaffidi G, et al. Comparison of early dobutamine stress echocardiography and exercise

electrocardiographic testing for management of patients presenting to the emergency department with chest pain. Am J Cardiol 2007;100:1068–1073.

221. Bernardi E, Camporese G, Buller HR, Siragusa S, Imberti D, Berchio A, et al. Ultrasonography for the diagnosis of clinically suspected deep vein thrombosis. a randomized study [abstract]. XXIst Congress of the International Society on Thrombosis and Haemostasis; 2007 Jul 6-12; Geneva.

222. Chauhan SP, Doherty DD, Magann EF, Cahanding F, Moreno F, Klausen JH. Amniotic fluid index vs single deepest pocket technique during modified biophysical profile: a randomized clinical trial. Am J Obstet Gynecol 2004;191:661–667.

223. Aaron SD, Vandemheen KL, Ferris W, Fergusson D, Tullis E, Haase D, et al. Combination antibiotic susceptibility testing to treat exacerbations of cystic fibrosis associated with multiresistant bacteria: a randomised, double-blind, controlled clinical trial. Lancet 2005;366:463–471.

224. Wang XH, Liu X, Sun YM, Gu JN, Shi HF, Zhou L, et al. Early identification and treatment of PV re-connections: role of observation time and impact on clinical results of atrial fibrillation ablation. Europace 2007;9:481–486.

225. Sameh WM. Value of intravenous urography before shockwave lithotripsy in the treatment of renal calculi: a randomized study. J Endourol 2007;21:574–577.

226. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174.

227. Luntz CA and Nebenzahl E. Behavior and interpretation of the K statistic: resolution of the two paradoxes J Clin Epidemiol 1996;49:431–434.

228. Whiting P, Rutjes AWS, Reitsma JB, Bossuyt PMM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003;3:25–38.

229. Department of Health [internet]. 2010. Available from URL: http://www.dh.gov.uk/en/Healthcare/index.htm [Accessed 10th January 2011].

230. Dixon-Woods M, Agarwal S, Jones D, Young B, Sutton A. Synthesising qualitative and quantitative evidence: a review of possible methods. J Health Serv Res Policy 2005;10:45–53.

231. Lusted LB. Introduction to medical decision making. Springfield IL: Charles C Thomas; 1968.

232. Doi K. Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. Phys Med Biol 2006;51:R5–R27.

233. Franco M, Cooper RS, Bilal U, Fuster V. Challenges and opportunities for cardiovascular disease prevention. Am J Med 2011;124:95–102.

234. Price CP, Christenson RH, editors. Evidence–based laboratory medicine from principles to practice. 2nd Edition. Washington DC: AACC Press; 2007.

235. Lundberg GD. The need for an outcomes research agenda for clinical laboratory testing. JAMA 1998;280:565–566.

236. Richards D, Toop L, Chambers S, Fletcher L. Response to antibiotics of women with symptoms of urinary tract infection but negative dipstick urine test results: double blind randomised controlled trial. BMJ 2005;331:143–148.

237. Egger M, Jüni P, Bartlett C, CONSORT Group. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996–1999.

238. Glenton C, Underland V, Kho M, Pennick V, Oxman AD. Summaries of findings, descriptions of interventions, and information about adverse effects would make reviews more informative. J Clin Epidemiol 2006;59:770–778.

239. Jacquier I, Boutron I, Moher D, Roy C, Ravaud P. The reporting of randomized clinical trials using a surgical intervention is in need of immediate improvement. A systematic review. Ann Surg 2006;244:677–683.

240. Boutron I, Tubach F, Giraudeau B, Ravaud P. Methodological differences in clinical trials evaluating nonpharmacological and pharmacological treatments of hip and knee osteoarthritis. JAMA 2003;290:1062–1070.

241. Jadad AR, Moore RA, Carroll D, Jenkinson C, Reynolds DJ, Gavaghan DJ, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials 1996;17:1–12.

242. Moher D, Cook DJ, Jadad AR, Tugwell P, Moher M, Jones A, et al. Assessing the quality of reports of randomised trials: implications for the conduct of meta-analyses. Health Technol Assess 1999;3:1–98.

243. Boutron I, Tubach F, Giraudeau B, Ravaud P. Blinding was judged more difficult to achieve and maintain in nonpharmacologic than pharmacologic trials. J Clin Epidemiol 2004;57:543–550.

244. Oxford Centre for Evidence–Based Medicine [internet]. Available at URL: http://www.cebm.net/?o=1025 [Accessed 25th October 2012].

245. Paterson C, Dieppe P. Characteristic and incidental (placebo) effects in complex interventions such as acupuncture. BMJ 2005 ;330:1202–1205.

246. Grimes DA, Schulz KF. Surrogate end points in clinical research: hazardous to your health. Obstet Gynecol. 2005;105:1114-1118.

247. Guillemin F. Primer: the fallacy of subgroup analysis. Nat Clin Pract Rheumatol 2007;3:401–413.

248. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ 1999;319:670-674.

249. Thabane L, Chu R, Cuddy K, Douketis J. What is the quality of reporting in weight loss intervention studies? A systematic review of randomized controlled trials. Int J Obes (Lond) 2007;31:1554–1559.

250. Gravel J, Opatrny L, Shapiro S. The intention-to-treat approach in randomized controlled trials: are authors saying what they do and doing what they say? Clin Trials 2007;4:350–356.

251. Kruse RL, Alper BS, Reust C, Stevermer JJ, Shannon S, Williams RH. Intention-to-treat analysis: who is in? Who is out? J Fam Pract 2002;51:969–971.

252. Choi SC, Lu IL. Effect of non-random missing data mechanisms in clinical trials. Stat Med 1995;14:2675–2684.

253. Altman DG. Missing outcomes in randomized trials: addressing the dilemma. Open Med 2009;3:e51–e53.

254. White IR, Kalaitzaki E, Thompson SG. Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an Internet-based alcohol trial. Stat Med 2011;30:3192–3207.

255. Deschartres A, Charles P, Hopewell S, Ravaud P, Altman DG. Reviews assessing the quality or reporting of randomized controlled trials are increasing over time but raised questions about how quality is assessed. J Clin Epidemiol 2011;64:136–144.

256. Biesheuvel CJ, Grobbee DE, Moons KGM. Distraction From Randomization in Diagnostic Research. Ann Epidemiol 2006;16:540–544.

257. Mushlin AI. Challenges and opportunities in economic evaluations of diagnostic tests and procedures. Acad Radiol 1999;6(suppl 1):S128-S131.

258. Gatsonis C. Design of evaluations of imaging technologies: development of a paradigm. Acad Radiol 2000;7:681–683.

259. Gatsonis C, Hillman BJ. Introduction. Acad Radiol 1999;6(suppl 1):S1–S2.

260. Bree, Dorfman GS, Greenfield S, Gatsonis C, Hillman B, Jarvik JG, et al. Intermediate Outcomes: Diagnostic and Therapeutic Impact – Discussion for Session 4. Acad Radiol 1999;6(suppl 1):S72–S76.

261. Znaniecki, F. The method of sociology. 1934. New York: Farrar & Rinehart.

262. Lindesmith, AR. Two comments on WS. Robinson's "The logical structure of analytic induction." Am Sociol Rev 1952;17:492–493.

263. Robinson, WS. The logical structure of analytic induction. Am Sociol Rev 1951;16:812–818.

264. Bossuyt PMM, McCaffery K. Multiple pathways and additional patient outcomes in evaluations of testing. Med Decis Making 2009;29:E30–E38.

265. Thornbury JR. Intermediate outcomes: Diagnostic and therapeutic impact. Workshop Session 4. Acad Radiol;1999:6(suppl 1):S58–S65.

266. Adriaensen MEAPM, Kock MCJM, Stijnen T, van Sambeek MRHM, van Urk H, Pattynama PMT, et al. Peripheral arterial disease: therapeutic confidence of CT versus Digital Subtraction Angiography and effects on additional imaging recommendations. Radiology 2004;233:385–391.

267. Bearcroft PWP, Guy S, Bradley M, Robinson F. MRI of the ankle: effect on diagnostic confidence and patient management. Am J Roentgenol 2006;187:1327–1331.

268. McCaffery KJ, Irwig L, Turner R, Chan SF, Macaskill P, Lewicka M et al. Psychosocial outcomes of three triage methods for the management of borderline abnormal cervical smears: an open randomised trial. BMJ 2010;340:b4491.

269. Haynes RB, Ackloo E, Sahota N, McDonald HP, Yao X. Interventions for enhancing medication adherence. Cochrane Database of Systematic Reviews 2008, Issue 2. Art. No.: CD000011. DOI: 10.1002/14651858.CD000011.pub3.

270. Claxton AJ, Cramer J, Pierce C. A systematic review of the associations between does regimens and medication compliance. Clin Ther 2001;23:1296-1310.

271. Pound P, Britten N, Morgan M, Yardley L, Pope C, Daker-White G, et al. Resisting medicines: a synthesis of qualitative studies of medicine taking. Soc Sci Med 2005;61:133-155.

272. DiMatteo MR, Giordani PJ, Lepper HS, Croghan TW. Patient adherence and medical treatment outcomes. Med Care 2002;40:794-811.

273. Coenen S, Michiels B, Van Royen P, Van der Auwera JC, Denekens J. Antibiotcs for coughing in general practice: a questionnaire study to quantify and condense the reasons for prescribing. BMC Fam Pract 2002;3:16–26.

274. Macfarlane J, Holmes W, Macfarlane R, Britten N. Influence of patients' expectations on antibiotic management of acute lower respiratory tract illness in general practice: questionnaire study. BMJ 1997;315:1211–1214.

275. Petursson P. GPs' reasons for "non-pharmacological" prescribing of antibiotics. A phenomenological study. Scand J Prim Health Care 2005;23:120–125.

276. Hauser MJ, Commons ML, Bursztajn HJ, Gutheil TG. Fear of Malpractice Liability and its Role in Clinical Decision Making. In: Gutheil TG, Bursztajn HJ, Brodsky A, Alexander V, editors. Decision making in psychiatry and the law. 1st Edition. Baltimore MD: Williams and Wilkins Co.; 1991. p.209–226.

277. McKinlay JB, Burns RB, Durante R, Feldman HA, Freund K, Harrow S, et al. Patient, physician and presentational influences on clinical decision making for breast cancer: results from a factorial experiment. J Eval Clin Pract 1997;3:23–57.

278. Little P, Dorward M, Warner G, Stephens K, Senior J, Moore M. Importance of patient pressure and perceived pressure and perceived medical need for investigations, referral, and prescribing in primary care: nested observational study. BMJ 2004;328:444–48.

279. Joffe M, Mindell J. Complex causal process diagrams for analyzing the health impacts of policy interventions. Am J Public Health 2006;96:473–479.

280. Hernán MA, Cole SR. Causal diagrams and measurement bias. Am J Epidemiol 2009; 170:959–962.

281. Cox E, Martin BC, Van Staa T, Garbe E, Siebert U, Johnson ML. Good research practices for comparative effectiveness research: approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources: the International Society for Pharmacoeconomics and Outcomes Research Good Research Practices for Retrospective Database Analysis Task Force Report--Part II. Value Health 2009;12:1053–1061.

282. Greenland S, Brumback B. An overview of relations among causal modelling methods. Int J Epidemiol 2002; 31:1030–1037.

283. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care 2007; 19:349–357.

284. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PloS Med 2009; 6:e1000097.

285. Stroman DL, Bayouth CV, Kuhn JA, Westmoreland M, Jones RC, Fisher TL, et al. The role of computed tomography in the diagnosis of acute appendicitis. Am J Surg 1999;178:485–489.

286. Wise SW, Labuski MR, Kasales CJ, et al. Comparative assessment of CT and sonographic techniques for appendiceal imaging. Am J Roentgenol 2001;176:933–941.

287. Anderson BA, Salem L, Flum DR. A systematic review of whether oral contrast is necessary for the computed tomography diagnosis of appendicitis in adults. Am J Surg 2005;190:474–478.

288. Smith–Bindman R, Lipson J, Marcus R, Kim KP, Mahesh M, Gould R, et al. Radiation dose associated with common computed tomography examinations and the associated lifetime attributable risk of cancer. Arch Intern Med 2009;169:2078–2086.

289. Old JL, Dusing RW, Yap W, Dirks J. Imaging for suspected appendicitis. Am Fam Physician 2005;71:71–78.

290. Howell JM, Eddy OL, Lukens TW, Thiessen ME, Weingart SD, Decker WW, et al. Clinical policy: Critical issues in the evaluation and management of emergency department patients with suspected appendicitis. Ann Emerg Med 2010; 55:71–116.

291. Rao PM, Rhea JT, Novelline RA, Mostafavi AA, McCabe CJ. Effect of computed tomography of the appendix on treatment of patients and use of hospital resources. N Engl J Med 1998; 338:141–146.

292. Gwynn LK. The diagnosis of acute appendicitis: clinical assessment versus computed tomography evaluation. J Emerg Med 2001; 21:119–123.

293. Wijetunga R, Tan BS, Rouse JC, Bigg-Wither GW, Doust BD. Diagnostic accuracy of focused appendiceal CT in clinically equivocal cases of acute appendicitis. Radiology 2001;221:747–753.

294. Morris KT, Kavanagh M, Hansen P, Whiteford MH, Deveney K, Standage B. The rational use of computed tomography scans in the diagnosis of appendicitis. Am J Surg 2002; 183:547–550.

295. Ujiki MB, Murayama KM, Cribbins AJ, Angelos P, Dawes L, Prystowsky JB, et al. CT scan in the management of acute appendicitis. J Surg Res 2002;105:119–122.

296. Hong JJ, Cohn SM, Ekeh AP, Newman M, Salama M, Leblang SD. A Prospective randomized study of clinical assessment versus computed tomography for the diagnosis of acute appendicitis. Surg Infect 2003; 4:231–239.

297. Doria AS, Moineddin R, Kellenberger CJ, Epelman M, Beyene J, Schuh S, et al. US or CT for diagnosis of appendicitis in children and adults? A meta-analysis. Radiology 2006; 241:83–94.

298. Ng CS, Watson CJ, Palmer CR, See TC, Beharry NA, Housden BA, et al. Evaluation of early abdominopelvic computed tomography in patients with acute abdominal pain of unknown cause: prospective randomised study. BMJ 2002;325:1387–1891.

299. Birnbaum BA, Jeffrey RB Jr. CT and sonographic evaluation of acute right lower quadrant abdominal pain. AJR Am J Roentgenol 1998;170:361–371.

300. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med 2007;49:196–205.

301. Temple CL, Huchcroft SA, Temple WJ. The natural history of appendicitis in adults: a prospective study. Ann Surg 1995;221:278–281.

302. Andersson RE. The natural history and traditional management of appendicitis revisited: spontaneous resolution and predominance of prehospital perforations imply that a correct diagnosis is more important than an early diagnosis. World J Surg 2007;31:86–92.

303. Blomqvist PG, Andersson RE, Granath F, Lambe MP, Ekbom AR. Mortality after appendectomy in Sweden, 1987-1996. Ann Surg 2001;233:455–460.

304. Varadhan KK, Neal KR, Lobo DN. Safety and efficacy of antibiotics compared with appendicectomy for treatment of uncomplicated acute appendicitis: meta-analysis of randomised controlled trials. BMJ 2012;344:e2156.

305. Palazzo L, Girollet PP, Salmeron M, Silvain C, Roseau G, Canard JM, et al. Value of endoscopic ultrasonography in the diagnosis of common bile duct stones: comparison with surgical exploration and ERCP. Gastrointest Endosc 1995;42:225–231.

306. De Lisi S, Leandro G, Buscarini E. Endoscopic ultrasonography versus endoscopic retrograde cholangiopancreatography in acute biliary pancreatitis: a systematic review. Eur J Gastroenterol Hepatol 2011; 23:367–374.

307. Palazzo L, O'Toole D. EUS in common bile duct stones. Gastrointest Endosc 2002;56(4 Suppl):S49–S57.

308. Prat F, Amouyal G, Amouyal P, Pelletier G, Fritsch J, Choury AD, et al. Prospective controlled study of endoscopic ultrasonography and endoscopic retrograde cholangiography in patients with suspected common-bile duct lithiasis. Lancet 1996;347:75–79.

309. Norton SA, Alderson D. Prospective comparison of endoscopic ultrasonography and endoscopic retrograde cholangiopancreatography in the detection of bile duct stones. Br J Surg 1997;84:1366–1369.

310. Liu CL, Lo CM, Chan JKF, Poon RT, Lam CM, Fan ST, et al. Detection of choledocholithiasis by EUS in acute pancreatitis: a prospective evaluation in 100 consecutive patients. Gastrointest Endosc 2001;54:325–330.

311. Buscarini E, Tansini P, Vallisa D, Zambelli A, Buscarini L. EUS for suspected choledocholithiasis: do benefits outweigh costs? A prospective, controlled study. Gastrointest Endosc 2003;57:510–518.

312. Jafri IH, Saltzman JR, Colby JM, Krims PE. Evaluation of the clinical impact of endoscopic ultrasonography (EUS) in gastrointestinal disease [abstract]. Gastrointest Endosc 1994; 40:P65.

313. Nickl NJ, Bhutani MS, Catalano M, Hoffman B, Hawes R, Chak A, et al. Clinical implications of endoscopic ultrasound: the American Endosonography Club Study. Gastrointest Endosc 1996;44:371–377.

314. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. J Epidemiol Community Health 2002;56:119–127.

315. Oakley A, Strange V, Bonell C, Allen E, Stephenson J, RIPPLE study team. Process evaluation in randomised controlled trials of complex interventions. BMJ 2006;332:413-416.

316. Leeuw F, Vaessen J. Impact evaluations and development: Nonie guidance on impact evaluation. Available from URL: www.worldbank.org/ieg/nonie [Accessed 11th October 2012].

317. White H. Theory-based impact evaluation: principles and practice. International Initiative for Impact Evaluation; 2009. Working Paper 3. Available from URL: www.**3ieimpact.org**/media/filer/2012/05/07/Working_Paper_3.pdf [Accessed 3rd July 2012].

318. Bossuyt PMM. Effective medical testing [oral presentation]. EVIDENCE 2011. Available at URL: http://www.evidencelive.org/2011/programme [Accessed 3rd July 2012].

319. Lindsay B. Randomized controlled trials of socially complex nursing interventions: creating bias and unreliability? J Adv Nurs 2004;45:84–94.

320. Fergusson D, Aaron SD, Guyatt G, Hébert P. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ 2002;325:652–654.

321. Greiner T, Gold H, Cattell M, Travell J, Bakst H, Rinzler SH, et al. A method for the evaluation of the effects of drugs on cardiac pain in patients with angina of effort; a study of khellin (visammin). Am J Med 1950;9:143–155.

322. Schulz KF, Grimes DA. Blinding in randomised trials: hiding who got what. Lancet 2002;359:696–700.

323. Stame N. Theory-based evaluation and types of complexity. Evaluation 2004;10:58–76

324. Victora CG, Habicht JP, Bryce J. Evidence-based public health: moving beyond randomized trials. Am J Public Health 2004;94:400–405.

325. Pibouleau L, Boutron I, Reeves BC, Nizard R, Ravaud P. Applicability and generalisability of published results of randomised controlled trials and non-randomised studies evaluating four orthopaedic procedures: methodological systematic review. BMJ 2009:339:b4538.

326. Green J. The evolving randomised controlled trial in mental health: studying complexity and treatment process. Advances in Psychiatric Treatment 2006;12:268–279.

327. Perera R, Heneghan C, Yudkin P. Graphical method for depicting randomised trials of complex interventions. BMJ 2007;334:127–129.

Included test-treatment RCTs – Reference list

(T1) Babjuk M, Soukup V, Petrík R, Jirsa M, Dvorácek J. 5-aminolaevulinic acid-induced fluorescence cystoscopy during transurethral resection reduces the risk of recurrence in stage Ta/T1 bladder cancer. BJU Int 2005;96:798–802.

(T2) Farwell DJ, Sulke AN. A randomised prospective comparison of three protocols for head-up tilt testing and carotid sinus massage. Int J Cardiol 2005;105:241–249.

(T3) Kearon C, Ginsberg JS, Douketis J, Crowther MA, Turpie AG, Bates SM, et al. A randomized trial of diagnostic strategies after normal proximal vein ultrasonography for suspected deep venous thrombosis: D-dimer testing compared with repeated ultrasonography. Ann Intern Med 2005;142:490–496.

(T4) Modic MT, Obuchowski NA, Ross JS, Brant-Zawadzki MN, Grooff PN, Mazanec DJ, et al. Acute low back pain and radiculopathy: MR imaging findings and their prognostic role and effect on outcome. Radiology 2005;237:597–604.

(T5) Nikken JJ, Oei EH, Ginai AZ, Krestin GP, Verhaar JA, van Vugt AB, et al. Acute peripheral joint injury: cost and effectiveness of low-field-strength MR imaging--results of randomized controlled trial. Radiology 2005;236:958–967.

(T6) (No authors listed). After thrombolysis for myocardial infarction, early routine angiography reduces cardiac events and death compared with conservative treatment [Abstract]. Evidence based Healthcare and Public Health 2005;9:127–128.

(T7) Takemura Y, Ebisawa K, Kakoi H, Saitoh H, Kure H, Ishida H, et al. Antibiotic selection patterns in acutely febrile new outpatients with or without immediate testing for C reactive protein and leucocyte count. J Clin Pathol 2005;58:729–733.

(T8) Bartha JL, Romero-Carmona R, Martínez-Del-Fresno P, Comino-Delgado R. Bishop score and transvaginal ultrasound for preinduction cervical assessment: a randomized clinical trial. Ultrasound Obstet Gynecol 2005;25:155–159.

(T9) Liu CL, Fan ST, Lo CM, Tso WK, Wong Y, Poon RT, et al. Comparison of early endoscopic ultrasonography and endoscopic retrograde cholangiopancreatography in the management of acute biliary pancreatitis: a prospective randomized study.

Clin Gastroenterol Hepatol 2005;3:1238-1244.

(T10) Brooks S, Cicuttini FM, Lim S, Taylor D, Stuckey SL, Wluka AE. Cost effectiveness of adding magnetic resonance imaging to the usual management of suspected scaphoid fractures. Br J Sports Med 2005;39:75–79.

(T11) Faltin DL, Boulvain M, Floris LA, Irion O. Diagnosis of anal sphincter tears to prevent fecal incontinence: a randomized controlled trial. Obstet Gynecol 2005;106:6–13.

(T12) Kock MC, Adriaensen ME, Pattynama PM et al. DSA versus multi-detector row CT angiography in peripheral arterial disease: randomized controlled trial. Radiology 2005;237:727–737.

(T13) Cuddihy MT, Locke GR, Wahner-Roedler D, Dierkhising R, Zinsmeister AR, Long KH, et al. Dyspepsia management in primary care: a management trial. Int J Clin Pract 2005;59:194–201.

(T14) (No author listed). Early coronary angiography and revascularisation significantly reduces death and heart attacks in elderly people with acute coronary syndromes. Evidence based Healthcare and Public Health 2005;9:38–39.

(T15) Gray A, Elbourne D, Dezateux C, King A, Quinn A, Gardner F. Economic evaluation of ultrasonography in the diagnosis and management of developmental hip dysplasia in the United Kingdom and Ireland. J Bone Joint Surg Am 2005;87:2472–2479.

(T16) Ramakrishna G, Milavetz JJ, Zinsmeister AR, Farkouh ME, Evans RW, Allison TG, et al. Effect of exercise treadmill testing and stress imaging on the triage of patients with chest pain: CHEER substudy. Mayo Clin Proc 2005;80:322–329.

(T17) Larsen SS, Vilmann P, Krasnik M, Dirksen A, Clementsen P, Maltbaek N, et al. Endoscopic ultrasound guided biopsy performed routinely in lung cancer staging spares futile thoracotomies: preliminary results from a randomised clinical trial. Lung cancer 2005;49:377–385.

(T18) Rao AV, Hsieh F, Feussner JR, Cohen HJ. Geriatric evaluation and management units in the care of the frail elderly cancer patient. J Gerontol A Biol Sci Med Sci 2005;60:798–803.

(T19) Ouwendijk R, de VM, Pattynama PM, van Sambeek MR, de Haan MW, Stijnen T, et al. Imaging peripheral arterial disease: a randomized controlled trial comparing contrast-enhanced MR angiography and multi-detector row CT angiography. Radiology 2005;236:1094–1103.

(T20) Eisenberg MJ, Teng FF, Chaudhry MR, Ortiz J, Sobkowski W, Ebrahim I, et al. Impact of invasive management versus noninvasive management on functional status and quality of life following non-Q-wave myocardial infarction: a randomized clinical trial. Am Heart J 2005;149:813–819.

(T21) Oosterheert JJ, van Loon AM, Schuurman R, Hoepelman AI, Hak E, Thijsen S, et al. Impact of rapid detection of viral and atypical bacterial pathogens by real-time polymerase chain reaction for patients with lower respiratory tract infection. Clin Infect Dis 2005;41:1438–1444.

(T22) Rao HB, Saksena S, Mitruka R. Intra-cardiac echocardiography guided cardioversion to help interventional procedures (ICE-CHIP) study: study design and methods. J Interv Card Electrophysiol 2005;13:31–36.

(T23) Bruins M, Oord H, Bloembergen P, Wolfhagen M, Casparie A, Degener J, et al. Lack of effect of shorter turnaround time of microbiological procedures on clinical outcomes: a randomised controlled trial among hospitalised patients in the Netherlands. Eur J Clin Microbiol Infect Dis 2005;24:305–313.

(T24) Purushotham AD, Upponi S, Klevesath MB, Bobrow L, Millar K, Myles JP, et al. Morbidity after sentinel lymph node biopsy in primary breast cancer: results from a randomized controlled trial. J Clin Oncol 2005;23:4312–4321.

(T25) Wetzig NR, Gill PG, Ung O, Collins J, Kollias J, Gillett D, et al. Participation in the RACS sentinel node biopsy versus axillary clearance trial. ANZ J Surg 2005;75:98–100.

(T26) Berkwits M, Localio AR, Kimmel SE. The effect of cardiac troponin testing on clinical care in a veterans population: a randomized controlled trial. J Gen Intern Med 2005;20:584–592.

(T27) Tanahatoe SJ, Lambalk CB, Hompes PG. The role of laparoscopy in intrauterine insemination: a prospective randomized reallocation study. Human Reprod 2005;20:3225–3230.

(T28) Djais N, Kalim H. The role of lumbar spine radiography in the outcomes of patients with simple acute low back pain. APLAR Journal of Rheumatology 2005;8:45–50.

(T29) Conti A, Pieralli F, Sammicheli L, Camaiti A, Vanni S, Grifoni S, et al. Updated management of non-st-segment elevation acute coronary syndromes: selection of patients for low-cost care: an analysis of outcome and cost effectiveness. Med Sci Monit 2005;11:CR100–CR108.

(T30) Green BT, Rockey DC, Portwood G, Tarnasky PR, Guarisco S, Branch MS, et al. Urgent colonoscopy for evaluation and management of acute lower gastrointestinal hemorrhage: a randomized controlled trial. Am J Gastroenterol 2005;100:2395–2402.

(T31) Collinson PO, John C, Lynch S, Rao A, Canepa-Anson R, Carson E, et al. A prospective randomized controlled trial of point-of-care testing on the coronary care unit. Ann Clin Biochem 2004;41:397–404.

(T32) Kapetanopoulos A, Heller GV, Selker HP, Ruthazer R, Beshansky JR, Feldman JA. Acute resting myocardial perfusion imaging in patients with diabetes mellitus: Results from the Emergency Room Assessment of Sestamibi for Evaluation of Chest Pain (ERASE Chest Pain) trial. J Nucl Cardiol 2004;11:570–577.

(T33) Nascimento MA, Lira RP, Soares PH, Spessatto N, Kara-José N, Arieta CE. Are routine preoperative medical tests needed with cataract surgery? Study of visual acuity outcome. Current Eye Res 2004;28:285–290.

(T34) Little P. Development and randomised controlled trial of dipsticks and diagnostic algorithms for the management of UTI [ISRCTN03525333]. National Research Register, UK; 2004. Available at URL: [http://www nrr nhs uk/].

(T35) Takao M, Uchio Y, Naito K, Fukazawa I, Kakimaru T, Ochi M. Diagnosis and treatment of combined intra-articular disorders in acute distal fibular fractures. J Trauma 2004;57:1303–1307.

(T36) Gilbert FJ, Grant AM, Gillan MG, Vale L, Scott NW, Campbell MK, et al. Does early imaging influence management and improve outcome in patients with low back pain? A pragmatic randomised controlled trial. Health Technol Assess 2004;8:1–131.

(T37) Wallace HC, Newbegin CJ. Does ENT outpatient review at 1-week post ventilation tube insertion improve outcome at 1 month in paediatric patients? Clin Otolaryngol Allied Sci. 2004;29:595–597.

(T38) Klein AL, Murray RD, Becker ER, Culler SD, Weintraub WS, Jasper S,E et al. Economic analysis of a transesophageal echocardiography-guided approach to cardioversion of patients with atrial fibrillation: the ACUTE economic data at eight weeks. J Am Coll Cardiol 2004;43:1217–1224.

(T39) Pelletier-Fleury N, Meslier N, Gagnadoux F, Person C, Rakotonanahary D, Ouksel H, et al. Economic arguments for the immediate management of moderate-to-severe obstructive sleep apnoea syndrome. Eur Respir J 2004;23:53–60.

(T40) Laheij RJ, Hermsen JT, Jansen JB, Horrevorts AM, Rongen RJ, Van Rossum LG, et al. Empirical treatment followed by a test-and-treat strategy is more cost-effective in comparison with prompt endoscopy or radiography in patients with dyspeptic symptoms: a randomized trial in a primary care setting. Fam Pract 2004;21:238–243.

(T41) Eisenberg MJ, Blankenship JC, Huynh T, Azrin M, Pathan A, Sedlis S, et al. Evaluation of routine functional testing after percutaneous coronary intervention. Am J Cardiol 2004;93:744–747.

(T42) Piccioli A, Lensing AW, Prins MH, Falanga A, Scannapieco GL, Ieran M, et al. Extensive screening for occult malignant disease in idiopathic venous thromboembolism: a prospective randomized clinical trial. J Thromb Haemost 2004;2:884–889.

(T43) Horisberger T, Harbarth S, Nadal D, Baenziger O, Fischer JE. G-CSF and IL-8 for early diagnosis of sepsis in neonates and critically ill children - safety and cost effectiveness of a new laboratory prediction model: study protocol of a randomized controlled trial [ISRCTN91123847]. Crit Care 2004;8:R443–R450.

(T44) Lassen AT, Hallas J, Schaffalitzky de Muckadell OB. Helicobacter pylori test and eradicate versus prompt endoscopy for management of dyspeptic patients: 6.7 year follow up of a randomised trial. Gut 2004;53:1758–1763.

(T45) Cavallini GM, Saccarola P, D'Amico R, Gasparin A, Campi L. Impact of preoperative testing on ophthalmologic and systemic outcomes in cataract surgery. Eur J Ophthalmol 2004;14:369–374.

(T46) Laheij RJ, Van Rossum LG, Heinen N, Jansen JB. Long-term follow-up of empirical treatment or prompt endoscopy for patients with persistent dyspeptic symptoms? Eur J Gastroenterol Hepatol 2004;16:785–789.

(T47) Bryan S, Bungay HP, Weatherburn G, Field S. Magnetic resonance imaging for investigation of the knee joint: a clinical and economic evaluation. Int J Technol

Assess Health Care 2004;20:222–229.

(T48) Goodacre SW, Quinney D, Revill S, Morris F, Capewell S, Nicholl J. Patient and primary care physician satisfaction with chest pain unit and routine care. Acad Emerg Med 2004;11:827–833.

(T49) Kiss H, Petricevic L, Husslein P. Prospective randomised controlled trial of an infection screening programme to reduce the rate of preterm delivery. BMJ 2004;329:371–376.

(T50) Lowe MP, Zimmerman B, Hansen W. Prospective randomized controlled trial of fetal fibronectin on preterm labor management in a tertiary care center. Am J Obstet Gynecol 2004;190:358–362.

(T51) Marteau T, Senior V, Humphries SE, Bobrow M, Cranston T, Crook MA, et al. Psychological impact of genetic testing for familial hypercholesterolemia within a previously aware population: a randomized controlled trial. Am J Med Genet A 2004;128:285–293.

(T52) Viney RC, Boyer MJ, King MT, Kenny PM, Pollicino CA, McLean JM, et al. Randomized controlled trial of the role of positron emission tomography in the management of stage I and II non-small-cell lung cancer. J Clin Oncol 2004;22:2357–2362.

(T53) Shen WK, Decker WW, Smars PA, Goyal DG, Walker AE, Hodge DO, et al. Syncope Evaluation in the Emergency Department Study (SEEDS): a multidisciplinary approach to syncope management. Circulation 2004;110:3636–3645.

(T54) Hamilton E, Platt R, Gauthier R, McNamara H, Miner L, Rothenberg S, et al. The effect of computer-assisted evaluation of labor on cesarean rates. J Healthc Qual 2004;26:37–44.

(T55) Mueller C, Scholer A, Laule-Kilian K, Martina B, Schindler C, Buser P, et al. Use of B-type natriuretic peptide in the evaluation and management of acute dyspnea. N Engl J Med 2004;350:647–654.

(T56) Farwell DJ, Freemantle N, Sulke AN. Use of implantable loop recorders in the diagnosis and management of syncope. Eur Heart J 2004;25:1257–1263.

(T57) Wallace P, Barber J, Clayton W, Currell R, Fleming K, Garner P, et al. Virtual outreach: a randomised controlled trial and economic evaluation of joint teleconferenced medical consultations. Health Technol Assess 2004;8:1–106.

(T58) Blomgren L, Johansson G, Bergqvist D. Randomized clinical trial of routine preoperative duplex imaging before varicose vein surgery. Br J Surg 2005;92:688–694.

(T59) Bouza E, Torres MV, Radice C, Cercenado E, de Diego R, Sánchez-Carrillo C, et al. Direct E-test (AB Biodisk) of respiratory samples improves antimicrobial use in ventilator-associated pneumonia. Clin Infect dis 2007;44:382–387.

(T60) Lee CC, Golub R, Singer AJ, Cantu R, Levinson H. Routine versus selective abdominal computed tomography scan in the evaluation of right lower quadrant pain: a randomized controlled trial. Acad Emerg Med 2007;14:117–122.

(T61) East CE, Gascoigne MB, Doran CM, Brennecke SP, King JF, Colditz PB. A cost-effectiveness analysis of the intrapartum fetal pulse oximetry multicentre randomised controlled trial (the FOREMOST trial). BJOG 2006;113:1080–1087.

(T62) Canadian Critical Care Trials Group, Heyland D, Dodek P, Muscedere J, Day A. A randomized trial of diagnostic techniques for ventilator-associated pneumonia. N Engl J Med 2006;355:2619–2630.

(T63) Kearon C, Ginsberg JS, Douketis J, Turpie AG, Bates SM, Lee AY, et al. An evaluation of D-dimer in the diagnosis of pulmonary embolism: a randomized trial. Ann Intern Med 2006;144:812–821.

(T64) Huas D, Pouchain D, Gay B, Avouac B, Bouvenot G, French College Of Teachers In General Practice. Assessing chronic pain in general practice: are guidelines relevant? A cluster randomized controlled trial. Eur J Gen Pract 2006;12:52–57.

(T65) Hu WH, Lam SK, Lam CL, Wong WM, Lam KF, Lai KC, et al. Comparison between empirical prokinetics, Helicobacter test-and-treat and empirical endoscopy in primary-care patients presenting with dyspepsia: a one-year study. Worold J Gastroenterol 2006;12:5010–5016.

(T66) Bosmans J, de BM, van HH, van Marwijk H, Beekman A, Bouter L, et al. Cost-effectiveness of a disease management program for major depression in elderly primary care patients. J Gen Intern Med 2006;21:1020–1026.

(T67) Bloom SL, Spong CY, Thom E, Varner MW, Rouse DJ, Weininger S, et al. Fetal pulse oximetry and cesarean delivery. N Engl J Med 2006;355:2195–2202.

(T68) Norlund A, Marké LA, af Geijerstam JL, Oredsson S, Britton M, Octopus S. Immediate computed tomography or admission for observation after mild head injury: cost comparison in randomised controlled trial. BMJ 2006;333:469–473.

(T69) Martina B, Nordmann A, Dieterle T, Sigle JP, Bengel G, Kiefer G, et al. Impact of baseline echocardiography on treatment outcome in primary care patients with newly detected arterial hypertension: a randomized trial. Am J Hypertens 2006;19:1150–1155.

(T70) de VM, Ouwendijk R, Flobbe K, Nelemans PJ, Kessels AG, Schurink GH, et al. Peripheral arterial disease: clinical and cost comparisons between duplex US and contrast-enhanced MR angiography--a multicenter randomized trial. Radiology 2006;240:401–410.

(T71) Jarbol DE, Kragstrup J, Stovring H, Havelund T, Schaffalitzky de Muckadell OB. Proton pump inhibitor or testing for Helicobacter pylori as the first step for patients presenting with dyspepsia? A cluster-randomized trial. Am J Gastroenterol 2006;101:1200–1208.

(T72) Melniker LA, Leibner E, McKenney MG, Lopez P, Briggs WM, Mancuso CA. Randomized controlled clinical trial of point-of-care, limited ultrasonography for trauma in the emergency department: the first sonography outcomes assessment program trial. Ann Emerg Med 2006;48:227–235.

(T73) Perquin DA, DöRr PJ, de Craen AJ, Helmerhorst FM. Routine use of hysterosalpingography prior to laparoscopy in the fertility workup: a multicentre randomized controlled trial. Hum Reprod 2006;21:1227–1231.

(T74) Rodger MA, Bredeson CN, Jones G, Rasuli P, Raymond F, Clement AM, et al. The bedside investigation of pulmonary embolism diagnosis study: a double-blind randomized controlled trial comparing combinations of 3 bedside tests vs ventilation-perfusion scan for the initial investigation of suspected pulmonary embolism. Arch Intern Med 2006;166:181–187.

(T75) Wong HT, Poon WS, Jacobs P, Goh KY, Leung CH, Lau FL, et al. The comparative impact of video consultation on emergency neurosurgical referrals. Neurosurgery 2006;59:607–613.

(T76) Debby A, Malinger G, Harow E, Golan A, Glezerman M. Transvaginal ultrasound after first-trimester uterine evacuation reduces the incidence of retained products of conception. Ultrasound Obstet Gynecol 2006;27:61–64.

(T77) Williams J, Russell I, Durai D, Cheung WY, Farrin A, Bloor K, et al. What are the clinical outcome and cost-effectiveness of endoscopy undertaken by nurses when compared with doctors? A Multi-Institution Nurse Endoscopy Trial (MINuET). Health Technol Assess 2006;10:1–195.

(T78) Vayssière C, David E, Meyer N, Haberstich R, Sebahoun V, Roth E, et al. A French randomized controlled trial of ST-segment analysis in a population with abnormal cardiotocograms during labor. Am J Obstet Gynecol 2007;197:299.e1–299.e6.

(T79) Jamal A, Marsoosi V, Eslamian L, Noori K. A prospective trial of the fetal biophysical profile versus modified biophysical profile in the management of high risk pregnancies. Acta Med Iran 2007;45:204–208.

(T80) Westerhuis ME, Moons KG, van BE, Bijvoet SM, Drogtrop AP, van Geijn HP, et al. A randomised clinical trial on cardiotocography plus fetal blood sampling versus cardiotocography plus ST-analysis of the fetal electrocardiogram (STAN) for intrapartum monitoring. BMC Pregnancy Childbirth 2007;7:13–21.

(T81) Windrim R, Seaward PG, Hodnett E, Akoury H, Kingdom J, Salenieks ME, et al. A randomized controlled trial of a bedside partogram in the active management of primiparous labour. J Obstet Gynecol Can 2007;29:27–34.

(T82) Goldstein JA, Gallagher MJ, O'Neill WW, Ross MA, O'Neil BJ, Raff GL. A randomized controlled trial of multi-slice coronary computed tomography for evaluation of acute chest pain. J Am Coll Cardiol 2007;49:863–871.

(T83) Sabharwal NK, Stoykova B, Taneja AK, Lahiri A. A randomized trial of exercise treadmill ECG versus stress SPECT myocardial perfusion imaging as an initial diagnostic strategy in stable patients with chest pain and suspected CAD: cost analysis. J Nucl Cardiol 2007;14:174–186.

(T84) de BR, van der PL, Hoekstra OS, Kuik DJ, Uyl-de Groot CA, van Tinteren H, et al. A randomized trial of PET scanning to improve diagnostic yield of direct laryngoscopy in patients with suspicion of recurrent laryngeal carcinoma after radiotherapy. Contemp Clin Trials 2007;28:705–712.

(T85) Sala E, Watson CJ, Beadsmoore C, Groot-Wassink T, Fanshawe TR, Smith JC, et al. A randomized, controlled trial of routine early abdominal computed tomography in patients presenting with non-specific acute abdominal pain. Clin Radiol 2007;62:961-969.

(T86) Ross MA, Compton S, Medado P, Fitzgerald M, Kilanowski P, O'Neil BJ. An emergency department diagnostic protocol for patients with transient ischemic attack: a randomized controlled trial. Ann Emerg Med 2007;50:109–119.

(T87) Kokkali G, Traeger-Synodinos J, Vrettou C, Stavrou D, Jones GM, Cram DS, et al. Blastocyst biopsy versus cleavage stage biopsy and blastocyst transfer for preimplantation genetic diagnosis of beta-thalassaemia: a pilot study. Hum Reprod 2007;22:1443–1449.

(T88) de Leusse A, Vahedi K, Edery J, Tiah D, Fery-Lemonnier E, Cellier C, et al. Capsule endoscopy or push enteroscopy for first-line exploration of obscure gastrointestinal bleeding? Gastroenterology 2007;132:855–862.

(T89) Jeetley P, Burden L, Stoykova B, Senior R. Clinical and economic impact of stress echocardiography compared with exercise electrocardiography in patients with suspected acute coronary syndrome but negative troponin: a prospective randomized controlled study. Eur Heart J 2007;28:204–211.

(T90) Denzinger S, Burger M, Walter B, Knuechel R, Roessler W, Wieland WF, et al. Clinically relevant reduction in risk of recurrence of superficial bladder cancer using 5-aminolevulinic acid-induced fluorescence diagnosis: 8-year results of prospective randomized study. Urology 2007;69:675–679.

(T91) Anderson DR, Kahn SR, Rodger MA, Kovacs MJ, Morris T, Hirsch A, et al. Computed tomographic pulmonary angiography vs ventilation-perfusion lung scanning in patients with suspected pulmonary embolism: a randomized controlled trial. JAMA 2007;298:2743–2753.

(T92) Sharples L, Hughes V, Crean A et al. Cost-effectiveness of functional cardiac testing in the diagnosis and management of coronary artery disease: a randomised controlled trial. The CECaT trial. Health Technol Assess 2007;11:1–115.

(T93) Goossens V, De RM, De VA, Staessen C, Michiels A, Verpoest W, et al. Diagnostic efficiency, embryonic development and clinical outcome after the biopsy of one or two blastomeres for preimplantation genetic diagnosis [see comment]. Hum Reprod 2007;23:481–492.

(T94) Ness A, Visintine J, Ricci E, Berghella V. Does knowledge of cervical length and fetal fibronectin affect management of women with threatened preterm labor? A randomized trial. Am J Obstet Gynecol 2007;197:426.e1–426.e7.

(T95) Polkowski M, Regula J, Tilszer A, Butruk E. Endoscopic ultrasound versus endoscopic retrograde cholangiography for patients with intermediate probability of bile duct stones: a randomized trial comparing two management strategies. Endoscopy 2007;39:296–303.

(T96) Beanlands RS, Nichol G, Huszti E, Humen D, Racine N, Freeman M, et al. F-18-fluorodeoxyglucose positron emission tomography imaging-assisted management of patients with severe left ventricular dysfunction and suspected coronary disease: a randomized, controlled trial (PARR-2). J Am Coll Cardiol 2007;50:2002–2012.

(T97) Cals JW, Hopstaken RM, Butler CC, Hood K, Severens JL, Dinant GJ. Improving management of patients with acute cough by C-reactive protein point of care testing and communication training (IMPAC3T): study protocol of a cluster randomised controlled trial. BMC Fam Prac 2007;8:15–26.

(T98) Mastenbroek S, Twisk M, van Echten-Arends J, Sikkema-Raddatz B, Korevaar JC, Verhoeve HR, et al. In vitro fertilization with preimplantation genetic screening. N Engl J Med 2007;357:9–17.

(T99) Hirsch A, Windhausen F, Tijssen JG, Verheugt FW, Cornel JH, de Winter RJ, et al. Long-term outcome after an early invasive versus selective invasive treatment strategy in patients with non-ST-elevation acute coronary syndrome and elevated cardiac troponin T (the ICTUS trial): a follow-up study. Lancet 2007;369:827–835.

(T100) Moe GW, Howlett J, Januzzi JL, Zowall H, Canadian Multicenter Improved Management of Patients With Congestive Heart Failure (IMPROVE-CHF) Study Investigators. N-terminal pro-B-type natriuretic peptide testing improves the management of patients with suspected acute heart failure: primary results of the Canadian prospective randomized multicenter IMPROVE-CHF study. Circulation 2007;115:3103–3110.

(T101) Fearon WF, Tonino PA, De BB, Siebert U, Pijls NH, Fame S, I. Rationale and design of the Fractional Flow Reserve versus Angiography for Multivessel Evaluation (FAME) study. Am Heart J 2007;154:632–636.

(T102) Alfirevic Z, len-Coward H, Molina F, Vinuesa CP, Nicolaides K. Targeted therapy for threatened preterm labor based on sonographic measurement of the cervical length: a randomized controlled trial. Ultrasound Obstet Gynecol 2007;29:47–50.

(T103) Bridgman S, Richards PJ, Walley G, MacKenzie G, Clement D, McCall I, et al. The effect of magnetic resonance imaging scans on knee arthroscopy: randomized controlled trial. Arthroscopy 2007;23:1167–1173.

(T104) Rama Raju GA, Shashi KG, Krishna KM, Prakash GJ, Madan K. Assessment of uterine cavity by hysteroscopy in assisted reproduction programme and its influence on pregnancy outcome. Arch Gynecol Obstet 2006;274:160–164.

(T105) Marzio L, Coraggio D, Capodicasa S, Grossi L, Cappello G. Role of the preliminary susceptibility testing for initial and after failed therapy of Helicobacter pylori infection with levofloxacin, amoxicillin, and esomeprazole. Helicobacter 2006;11:237–242.

(T106) Estrada JN, Rolandi F, Bansilal S, Averbuj P, Natale E, Zafar MU, et al. Stress testing and troponin in unstable coronary syndromes: the status trial-clinical outcomes and resource use. Am Heart Hosp J 2006;4:252–258.

(T107) Brealey SD, Atwell C, Bryan S, Coulton S, Cox H, Cross B, et al. The DAMASK trial protocol: a pragmatic randomised trial to evaluate whether GPs should have direct access to MRI for patients with suspected internal derangement of the knee. BMC Health Serv Res 2006;6:133–142.

(T108) Rábago LR, Vicente C, Soler F, Delgado M, Moral I, Guerra I, et al. Two-stage treatment with preoperative endoscopic retrograde cholangiopancreatography (ERCP) compared with single-stage treatment with intraoperative ERCP for patients with symptomatic cholelithiasis with possible choledocholithiasis. Endoscopy 2006;38:779–786.

Appendices

Appendix A: Search strategy & study selection

This appendix provides additional details of the project search for published test-treatment RCTs and the study selection process.

A.1 Development of search strategy

Three initial strategies were tested in Ovid MEDLINE in May 2006. General diagnosis terms across all fields, for publication years 1966 – 2006. Truncations are denoted by $.

A.1.1 Haynes sensitive RCT methods filter

Methods filters are predefined search strategies designed to maximise search precision by filtering content in bibliographic databases according to methodologic characteristics. Haynes and colleagues developed a series of methods filters by handsearching key journals, and identifying the optimal search strings using index and text terms from all relevant studies. These were tested and validated specifically for use in the MEDLINE database. The Haynes RCT filters limit search results to the RCT study design, and was chosen to maximise the project search's precision as it has demonstrated the highest sensitivity and amongst the highest specificities of 19 similar filters, when carried out in MEDLINE[A1].

The first strategy uses the sensitive Haynes RCT filter, which uses a broad range of terms to identify articles associated with an RCT study. These could hence also include secondary evidence in addition to the primary RCTs sought by the project.

Diagnosis:	sensitive$.mp.; diagnosis$.mp.; di.fs ; "sensitivity and specificity"/ (combined OR)
RCT:	sensitive RCT methods filter (Haynes et al)
Treatment:	intervention.mp.; experimental.mp.; study group$.mp.; treatment.mp.; treatment outcome/ (combined OR)

Yield = 187,895

A.1.2 Haynes specific RCT methods filter and control$.mp search term

The second strategy uses the specific RCT methods filter, which uses a more closely defined set of terms to limit findings to primary RCT reports (e.g. rather than systematic reviews).

Diagnosis:	sensitive$.mp.; diagnosis$.mp.; di.fs ; “sensitivity and specificity”/ (combined OR)
RCT:	specific RCT methods filter (Haynes et al)
Treatment:	intervention.mp.; experimental.mp.; study group$.mp.; treatment.mp.; treatment outcome/; control$.mp. (combined OR)

Yield = 51,699

A.1.3 RCT.pt and control$.mp search terms

The third strategy refers solely to the RCT as a publication type, which targets only articles indexed as primary RCT reports.

Diagnosis:	sensitive$.mp.; diagnosis$.mp.; di.fs ; “sensitivity and specificity”/ (combined OR)
RCT:	RCT.pt.
Treatment:	intervention.mp.; experimental.mp.; study group$.mp.; treatment.mp.; treatment outcome/; control$.mp. (combined OR)

Yield = 31,896

A.2 Abstract/Full Paper Inclusion Proforma

Title:	
Authors:	
Publication details:	**Ref ID:**

Screen:		
Study Type	Is this a randomised controlled trial?	Y / N
	Do participants have a clinical complaint? (i.e. not healthy or asymptomatic)	Y / N
	Are patients randomised to different diagnostic strategies?	Y / N
	Is the test used to classify suspected or existing disease for the purposes of treatment planning?	Y / N
	Is the test used repeatedly to monitor disease progression or titrate treatment?	Y / N
	Is treatment given as a result of dianostic information? (either **explicitly** mentioned or **implicit** from type of outcomes evaluated)	Y / N
	Are patient outcomes assessed after treatment?	Y / N
	INCLUDE / EXCLUDE	

A.3 Common examples of excluded trials

The table below lists some common examples of exclusions:

Title	Reason for exclusion	Source
Acute barrier disruption by adhesive tapes is influenced by pressure, time and anatomical location: integrity and cohesion assessed by sequential tape stripping. A randomized, controlled study	Test RCT, development of use of healthy volunteers	[A2]
The Ekman 60 Faces Test as a diagnostic instrument in frontotemporal dementia	Non-randomised evaluation of a test	[A3]
Measurement of the intraocular pressure with the 'transpalpebral tonometer' TGDc-01 in comparison with applanation tonometry	Non RCT, Test accuracy study	[A4]
Brain Imaging and Mental Disorders of Aging Intervention	Treatment RCT	[A5]
A prospective randomized comparative study on the safety and tolerability of transnasal esophagogastroduodenoscopy	Test RCT with no treatment phase	[A6]
Autofluorescence bronchoscopy with white light bronchoscopy compared with white light bronchoscopy alone for the detection of precancerous lesions: a European randomised controlled multicentre trial	Test RCT, Accuracy	[A7]
Application of contrast-enhanced ultrasound to increase the diagnostic rate of liver tumor by biopsy	Test-treatment RCT with no downstream patient outcomes	[A8]
A comparison of an evidence based regime with the standard protocol for monitoring postoperative observation: a randomised controlled trial	Test RCT evaluating monitoring test	[A9]
Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models	Test RCT evaluating asymptomatic screening	[A10]

A.4 References

[A1] Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR for the Hedges Team. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ 2005;330(7501):1179–85.

[A2] Breternitz M, Flach M, Prässler J, Elsner P, Fluhr JW. Acute barrier disruption by adhesive tapes is influenced by pressure, time and anatomical location: integrity and cohesion assessed by sequential tape stripping. A randomized, controlled study. Br J Dermatol 2007; 156: 231–240.

[A3] Diehl-Schmid J, Pohl C, Ruprecht C, Wagenpfeil S, Foerstl H, Kurz A. The Ekman 60 Faces Test as a diagnostic instrument in frontotemporal dementia. Arch Clin Neuropsychol 2007; 22:459–464.

[A4] Sandner D, Böhm A, Kostov S, Pillunat L. Measurement of the intraocular pressure with the "transpalpebral tonometer" TGDc-01 in comparison with applanation tonometry. Graefes Arch Clin Exp Ophthalmol. 2005; 243:563–569.

[A5] Small GW. Brain Imaging and Mental Disorders of Aging Intervention. ClinicalTrials.gov NCT00267163 [internet]. Available from: http://www.clinicaltrials.gov/ct2/show/NCT00267163?term=NCT00267163&rank=1

[A6] Yagi J, Adachi K, Arima N, Tanaka S, Ose T, Azumi T, et al. A prospective randomized comparative study on the safety and tolerability of transnasal esophagogastroduodenoscopy. Endoscopy 2005; 37:1226–1231.

[A7] Häussinger K, Becker H, Stanzel F, Kreuzer A, Schmidt B, Strausz J, et al. Autofluorescence bronchoscopy with white light bronchoscopy compared with white light bronchoscopy alone for the detection of precancerous lesions: a European randomised controlled multicentre trial. Thorax 2005; 60:496–503.

[A8] Wu W, Chen MH, Yan K, Yin SS, Dai Y, Fan ZH, et al. Application of contrast-enhanced ultrasound to increase the diagnostic rate of liver tumor by biopsy. Zhonghua Yi Xue Za Zhi 2006; 86:116–120.

[A9] Fernandez R, Griffiths R. A comparison of an evidence based regime with the standard protocol for monitoring postoperative observation: a randomised controlled trial. Aust J Adv Nurs 2005; 23:15–21.

[A10] Davis A, Smith P, Ferguson M, Stephens D, Gianopoulos I. Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models. Health Technol Assess 2007; 11:1-294.

Appendix B:
Diagnostic characteristics

This appendix lists all 108 included test-treatment RCTs, and provides basic details of trial design, patient population, the diagnostic interventions and the purpose of testing.

The following abbreviations are used throughout:

CCU	Coronary care unit
CK	Creatine kinase
CKMB	Creatine kinase MB
CPU	Chest pain unit
CT	Computed tomography
CTG	Cardiotocography
ECG	Electrocardiography
EUS	Endoscopic ultrasound
ERCP	Endoscopic retrograde cholangiopancreatography
FDG-PET	Fludeoxyglucose (18F) enhanced Positron emission tomography
FISH	Fluorescence in situ hybridization
FFN	Fetal fibronectin
FNA	Fine needle aspiration
FPO	Fetal pulse oximetry
MRI	Magnetic resonance imaging
NT–proBNP	N-terminal prohormone of brain natriuretic peptide
PCR	Polymerase chain reaction
PSA	Prostate antigen biomarker
PSG	Polysomnography
SPECT	Single-photon emission computed tomography
SPECT MPI	SPECT Myocardial perfusion imaging
US	Ultrasound
V/Q	Ventilation–perfusion scan

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T1	2–arm Standard Parallel	Suspected primary or recurrent superficial urinary bladder tumours	Control	White light cystoscopy during transurethral resection	Additional	To locate bladder tumours during transurethral resection
			Experimental	Fluorescence cystoscopy plus white light cystoscopy		
T2	3–arm Standard Parallel	Secondary care patients with recurrent unexplained syncope	Control	Drug-free tilt table testing	Replacement	To determine the cause of vasovagal syncope and determine the need for further investigation and/or treatment
			Experimental	Tilt table testing with glyceryl trinitrate spray (GTN)		
			Experimental	Tilt table testing with Adenosine		
T3	2–arm Standard Parallel	Patients referred to secondary care for suspected first episode of DVT	Control	Repeat compression ultrasonography	Replacement	To rule out thrombus and determine the need for treatment
			Experimental	D-dimer (± Venography)		
T4	2–arm Randomised non-disclosure	Patients referred to secondary care with acute onset (<3 weeks) of lower back pain or radiculopathy	Control	Clinical assessment	Additional	To rule out serious pathology and inform the treatment approach
			Experimental	Early MRI results (within 48hrs) plus clinical assessment		
T5	6–arm Standard Parallel (stratified according to sit of injury)	Secondary care patients with recent injury of wrist, ankle or knee requiring radiography	Control	Radiography	Additional	To detect fracture and inform the treatment approach
			Experimental	Extremity MRI AND Radiography		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T6	2–arm Standard Parallel	Patients referred for investigation of ST-segment elevated myocardial infarction	Control	Conservative pharmacological Rx, followed by angiography only if clinically indicated (spontaneous recurrent ischaemia with ECG changes or an abnormal non-invasive stress test)	Replacement	To determine the extent of coronary arterial disease and inform the need for revascularisation
			Experimental	Routine early angiography (within 24hrs)		
T7	2–arm Standard Parallel	Patients presenting to primary care with acute fever and suspected infection	Control	Clinical examination (No C-Reactive Protein and White Blood cell Count test results at initial consultation)	Additional	To differentiate between bacterial and non-bacterial infection, in order to select appropriate treatment
			Experimental	C-Reactive Protein and White Blood Cell count results available at Clinical examination		
T8	2–arm Random disclosure parallel	Women admitted for induction of labour	Control	Bishop Score + Blinded Transvaginal US	Replacement	To evaluate the readiness of the cervix for labour
			Experimental	Transvaginal Ultrasonography + Blinded Bishop score		
T9	2–arm Standard Parallel	Patients with a first episode of acute pancreatitis with suspected biliary cause, referred for surgery	Control	Diagnostic ERCP	Replacement	To detect cholelithiasis and remove stones from the bile duct
			Experimental	EUS		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T10	2–arm Standard Parallel	Emergency patients with possible occult scaphoid fracture and normal/inconclusive wrist radiograph	Control	Standard investigation - "usual care"; control group diagnosis based on "the modality or modalities that their treating doctor used"	Additional	To confirm fracture of scaphoid and determine suitability for further treatment
			Experimental	MRI within 2-5days and Standard investigation		
T11	2–arm Standard Parallel	Women with a second-degree perineal tear due to vaginal delivery	Control	Clinical examination	Additional	To determine the severity of tears to the anal sphincter and inform the need for treatment
			Experimental	Clinical exam AND Ultrasound examination of anal sphincter		
T12	2–arm Standard Parallel	Patients with symptomatic peripheral arterial disease referred for evaluation of feasibility of revascularisation	Control	Digital Subtraction Angiography	Replacement	To determine the location and severity of stenoses in order to inform the treatment approach
			Experimental	Multi-detector row CT angiography		
T13	4–arm Standard Parallel	Patients presenting to primary care with dyspepsia of >4week duration	Control	Empirical Rx	Replacement	To find the cause of dyspeptic symptoms (peptic ulcer disease vs. non-ulcer dyspepsia) and direct the need for treatment
			Exp #1	Hp Serology		
			Exp #2	Hp Carbon-13 Urea Breath Test		
			Exp #3	Endoscopy		
T14	2–arm Standard Parallel	Candidates for coronary revascularisation with a recent episode of angina (<24 hrs) and myocardial infarction without ST–segment elevation	Control	Medical therapy in all, coronary angiography only if pt meets clinical criteria for failure of medical treatment	Replacement	To determine the presence of recurrent ischaemia and select between multiple predefined treatment options
			Experimental	Early mandatory coronary angiography (4-48hrs)		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T15	2–arm Standard Parallel	Neonates with clinically suspected hip instability	Control	Standard: Clinical examination only	Additional	To diagnose congenital hip dislocation and determine the need for treatment
			Experimental	Ultrasonography plus standard investigation		
T16	2–arm Standard Parallel	Emergency patients with unstable angina and intermediate risk of short-term cardiovascular events	Control	Usual care: monitored bed in cardiology service of hospital (same management options available but not standardised)	Triage	To identify patients who can be discharged home safely
			Experimental	Chest pain unit: CKMB ± functional testing (treadmill exercise test, nuclear stress test or echo stress test)		
T17	2–arm Standard Parallel	Patients with suspected or newly diagnosed non–small cell lung cancer who are candidates for invasive staging prior to curative surgery	Control	Selective endoscopic ultrasound guided FNA if indicated by positive CT scan (enlarged lymphs or spread) Note: both groups also underwent conventional work-up including mediastinoscopy prior to randomisation	Replacement	To stage disease and detect mediastinal spread, and determine whether surgery can still be curative
			Experimental	Routine endoscopic ultrasound guided FNA. Note: both groups also underwent conventional work-up including mediastinoscopy prior to randomisation		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T18	4–arm Factorial 2x2	Hospitalised elderly patients in a frail condition	Control	No targeted detection of geriatric syndromes unless symptoms develop (inpatient and outpatient care)	Replacement	To assess patients' physiologic, functional and cognitive attributes and diagnose possible comorbidity in order to direct all aspects of care
			Exp #1	Targeted inpatient care plan (Prompt and detailed patient assessments [physical and history, functional status, nutritional status, caregiver conditions]); Standard untargeted outpatient care.		
			Exp #2	Standard untargeted inpatient care; Targeted outpatient care plan		
			Exp #3	Targeted inpatient and outpatient care plan		
T19	2–arm Standard Parallel	Patients with symptomatic peripheral arterial disease referred for evaluation of feasibility of revascularisation	Control	CT Angiography	Replacement	To determine the location and severity of stenoses in order to inform the treatment approach
			Experimental	MR Angiography		
T20	2–arm Standard Parallel	Hospitalised patients with chest pain indicative of cardiac ischaemia, with documented non-Q-wave myocardial infarction	Control	Stress testing (stress echocardiography OR stress nuclear perfusion imaging) ± Angiography	Replacement	To detect narrowing or obstruction of the heart vessels and inform the need for treatment
			Experimental	Angiography		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T21	2–arm Random disclosure	Hospitalised patients requiring immediate antimicrobial treatment for lower respiratory tract infection	Control	Standard microbiological samples only (PCR masked)	Additional	To determine whether the cause of infection is viral or bacterial, and inform the most appropriate treatment
			Experimental	Real-time PCR + standard microbiological samples		
T22	2–arm Standard Parallel	Patients with atrial fibrillation undergoing clinically indicated cardiac catheterization procedures who require electrical cardioversion	Control	No Test (treatment in all)	Replacement	To detect systemic thromboembolism or left atrial pathology during interventional cardiology procedures to indicate safety of immediate cardioversion
			Experimental	Intracardiac echocardiography		
T23	6–arm 3 x parallel randomisations	Hospitalised patients with a bacterial infection confirmed by culture	Control	Study Period 1/2/3: Standard (overnight) identification and susceptibility testing	Replacement	To identify the causative microorganism(s) and determine the susceptibility to antibiotics
			Exp #1	Study period 1: Vitek 2 - rapid identification and susceptibility testing with overnight results		
			Exp #2	Study period 2: Vitek 2 - rapid identification and susceptibility testing with some same day results)		
			Exp #3	Study period 3: Vitek 2 - rapid identification and susceptibility testing with all results returned same day		
T24	2–arm Standard Parallel	Histologically confirmed early invasive breast cancer	Control	Axillary lymph node dissection (level 2)	Triage	To determine the presence of lymph node metastases indicating a change in treatment
			Experimental	Sentinel Lymph Node Biopsy ± axillary lymph node dissection		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T25	2–arm Standard Parallel	Women with histologically or cytologically confirmed operable invasive breast cancer	Control	Sentinel Lymph Node Biopsy plus immediate Axilliary Clearance	Replacement	To determine the presence of lymph node metastases indicating a change in treatment
			Experimental	Sentinel Lymph Node Biopsy +/- Axillary clearance if positive		
T26	2–arm Random disclosure	Emergency patients or outpatients presenting with symptoms prompting cardiac enzyme testing	Control	CK/CK-MB	Additional	To diagnose the presence of acute coronary syndromes and determine the need for further investigation and/or treatment
			Experimental	CK/CH-MB + Troponin I		
T27	2–arm Standard Parallel	Women referred to secondary care for investigation and treatment of infertility	Control	Diagnostic laparoscopy	Triage	To detect and treat possible biological cause of infertility (e.g. tubal pathology or endometriosis)
			Experimental	Rx only ± diagnostic laparoscopy		
T28	2–arm Standard Parallel	Patients referred to secondary care with acute onset (<3 months) lower back pain	Control	No test presumed; states "Usual care"	Replacement	To detect possible cause of LBP
			Experimental	Lumbar spine radiography		
T29	2–arm Standard Parallel	Hospitalised patients with chest pain and non–ST–segment elevated acute coronary syndromes, with suspected underlying ischemic heart disease, and at intermediate risk of cardiac events	Control	CCU care (dedicated cath lab cardiologists)	Replacement	To narrow the differential diagnosis and inform the need for treatment
			Experimental	CPU care (emergency department physicians)		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T30	2–arm Standard Parallel	Patients admitted to hospital with lower gastrointestinal bleed.	Control	Selective colonoscopy or technetium RBC scan, according to bleeding status	Replacement	To detect the origin of the bleed and target treatment
			Experimental	Urgent purge preparation colonoscopy		
T31	2–arm Standard Parallel	Patients admitted to hospital with chest pain who are at high risk of acute coronary syndromes	Control	Central lab testing cardiac troponin plus standard investigation (included continuous ECG monitoring, daily 12 lead ECG, CK)	Replacement	To diagnose the cause of chest pain; To identify pts without acute myocardial infarction who would benefit from further early investigations (rapid rule out group)
			Experimental	POCT cardiac troponin + standard investigations		
T32	2–arm Standard Parallel	Emergency patients with suspected acute cardiac ischemia	Control	Standard clinical evaluation strategy (not specified, varies by centre)	Additional	To determine the presence of ischaemia and cause of chest pain in order to establish the need for further investigation and/or treatment
			Experimental	SPECT Tc 99m sestamibi AND standard clinical evaluation strategy		
T33	2–arm Standard Parallel	Patients scheduled to undergo cataract surgery	Control	Routine preoperative testing (12-l ECG, complete blood, serum glucose)	Triage	To identify conditions or characteristics which risk adverse events during surgery
			Experimental	Selective preoperative testing (No pre–op tests unless pt develops with new or worsening condition that would warrant testing even if surgery were not planned)		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T34	5–arm Standard parallel	Non-pregnant women presenting with a suspected uncomplicated urinary tract infection	Control	Patient report/signs and symptoms	Replacement	To confirm infection and inform treatment selection
			Exp #1	Patient report/signs and symptoms		
			Exp #2	Symptom score (two or more of urine cloudy, urine offensive smell, moderately severe dysuria or nocturia)		
			Exp #3	Dipstick algorithm (nitrites or leucocytes and a trace of blood)		
			Exp #4	Mid-stream urine results (symptomatic treatment until MSU results available)		
T35	2–arm Standard Parallel	Fracture (type B Weber) to distal fibula, referred for operative treatment	Control	Visual inspection/palpation only	Additional	To detect intra-articular lesions during operative reduction of fibular fracture and treat them.
			Experimental	Arthroscopy + inspection/palpation		
T36	2–arm Standard parallel	Patients with lower back pain referred by GP to consultant, on whom indication for imaging is clinically uncertain	Control	MRI/CT - Early, routine	Triage	To rule out serious pathology
			Experimental	MRI/CT - Delayed, selective (based on change in condition)		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T37	2–arm Standard Parallel	Children undergoing ventilation tube insertion	Control	No follow up visit	Replacement	To identify postoperative complications and determine the need for treatment
			Experimental	Clinical assessment of otological complications following grommet insertion		
T38	2–arm Standard Parallel	Candidates for electrical cardioversion, with atrial fibrillation of >2 days duration	Control	No test, treatment only	Replacement	To detect systemic thromboembolism or left atrial pathology during interventional cardiology
			Experimental	Transesophageal echocardiography		
T39	2–arm Standard Parallel	Patients with suspected obstructive sleep apnoea attending a sleep clinic	Control	Delayed PSG (Observation)	Replacement	To confirm the diagnosis of obstructive sleep apnoea and inform the need for treatment
			Experimental	Immediate PSG		
T40	3–arm Standard Parallel	Patients with persistent dyspepsia referred by GP for endoscopy	Control	Upper gastrointestinal endoscopy	Triage [EXP1]	To discriminate between acid and non-acid causes in order to identify those in need of further investigations and/or treatment
			Exp #1	Empirical Rx ± endoscopy OR H.pylori serology		
			Exp #2	Upper gastrointestinal radiography	Replacement [EXP2]	

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T41	2–arm Standard Parallel	Patients recovering from recent complete coronary revascularisation	Control	Selective testing: Exercise treadmill test on clinical indication	Replacement	To detect restenosis or postoperative complication and treat as indicated
			Experimental	Routine testing: Exercise treadmill test (with nuclear perfusion imaging) at 6 weeks (and 6 months)		
T42	2–arm Standard Parallel	Patients with a first episode of symptomatic deep-vein thrombosis of the lower extremity, or pulmonary embolism	Control	No Test	Replacement	To identify patients with occult malignancy in need of further investigation and/or treatment
			Experimental	Extensive screening including: US, CT, gastroscopy, colonoscopy, hemoccult, sputum cytology, tumour markers, mammogram, Pap smear, prostate US, PSA		
T43	2–arm Cluster; Random Disclosure Clusters (by day)	Critically ill newborns and children admitted to neonatal or paediatric intensive care unit with suspected bacterial infection	Control	Routine management	Additional	To confirm presence of bacterial infection and inform the treatment approach
			Experimental	Routine management + blood or tracheal aspirate specimens		
T44	2–arm Standard Parallel	Patients with dyspepsia of at least 2 week's duration referred to secondary care for further investigation	Control	Endoscopy	Triage	To detect the presence of H.pylori bacterium and treat as indicated
			Experimental	Carbon-13 Urea Breath Test (H.pylori Test) ± endoscopy		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T45	2–arm Standard Parallel	Patients undergoing elective cataract surgery	Control	Routine medical tests and electrocardiogram	Replacement	To identify conditions or characteristics which risk adverse events during surgery
			Experimental	No preoperative testing (tests performed but results not released), use of a pulsimeter during operation		
T46	2–arm Standard Parallel	Patients with persistent dyspepsia referred by GP for endoscopy	Control	Upper gastrointestinal endoscopy	Triage	To determine the cause of dyspepsia and inform the subsequent treatment approach
			Experimental	Empirical Rx ± H.pylori serology ± Endoscopy		
T47	2–arm Standard parallel	Patients referred to secondary care for management of knee problems, in whom arthroscopy is being considered	Control	Arthroscopy	Triage	To diagnose the cause of knee pain and inform the need for treatment
			Experimental	MRI +/- arthroscopy		
T48	2–arm Cluster Parallel	Emergency patients with acute undifferentiated chest pain who are undiagnosed after initial ED assessment	Control	Routine assessment (CK-MB, Troponin and presumably ECG available but not standardised, patients admitted if in need of further investigation)	Replacement	To determine a cardiac cause for chest pain, and distinguish between those needing further investigation/treatment and those who do not.
			Experimental	Chest Pain Unit: including ECG, CK-MB, Troponin T, continuous ST segment monitoring, exercise treadmill.		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T49	2–arm Randomised disclosure	Pregnant women without subjective complaints presenting for routine prenatal visit	Control	No test (results masked)	Replacement	To diagnose the presence of candida, trichomonas vaginalis or bacterial vaginosis infection and treat appropriately
			Experimental	Smear and Gram Stain		
T50	2–arm Standard Parallel	Women with signs and symptoms of preterm labour	Control	Cervical examination	Additional	To determine the risk of preterm labour and inform the need for treatment
			Experimental	Fetal fibronectin		
T51	2–arm Cluster Parallel	Individuals with diagnosed or suspected heterozygous Familial Hypercholesterolemia	Control	Standard investigation: routine clinical assessment for familial hypercholesterolemia (analysis of fasting lipid profile, personal and family history of premature CHD and hyperlipidemia and clinical history)	Additional	To diagnose the presence of familial hypercholesterolemia and inform the need for treatment
			Experimental	Molecular analysis of DNA plus standard investigation		
T52	2–arm Standard Parallel	Patients with histologically or cytologically confirmed non–small cell lung cancer referred for surgical staging	Control	No Test	Replacement	To stage the cancer and identify those with inoperable disease
			Experimental	FDG PET		
T53	2–arm Standard Parallel	Emergency patients with syncope of undetermined cause, at intermediate risk of adverse cardiovascular events	Control	Routine investigation at emergency physician's discretion (usually admission for further evaluation)	Replacement	To diagnose the cause of syncope and identify those in need of further investigation or treatment
			Experimental	Syncope Unit: Continuous cardiac monitoring, hourly vital signs + orthostatic blood, ECG, tilt-table testing, echocardiogram)		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T54	2–arm Standard Parallel	First-time mothers near term with baby presenting in normal head first position	Control	Standard evaluation of labour - cervical dilatation plotted against time only (no reference range).	Additional	To identify whether progress of cervical dilation is 'abnormal', determine women undergoing difficult labour and determine the need for assisted delivery
			Experimental	Computer assisted evaluation of labour - cervical dilatation plotted against time with superimposed computerised reference range		
T55	2–arm Standard Parallel	Emergency patients with acute dyspnoea with no obvious traumatic cause	Control	Standard investigation (including: Clinical history, physical exam, ECG, pulse oximetry, blood tests, chest x-ray, echocardiography and pulmonary function tests recommended)	Additional	To establish or rule out the diagnosis of heart failure, and determine the need for further investigation and/or treatment
			Experimental	B-type natriuretic peptide AND standard investigation		
T56	2–arm Standard Parallel	Secondary care patients with recurrent syncope of undetermined cause	Control	Standard investigation (not described but includes ECG diagnosis in some cases)	Replacement	To identify the cause of syncope and treat as indicated
			Experimental	Implantable loop recorder		
T57	2–arm Standard Parallel	Any patients referred from primary to outpatient care	Control	Standard outpatient specialist appointment	Replacement	To diagnose and treat according to the indication
			Experimental	Video conferencing		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T58	2–arm Standard Parallel	Patients with primary varicose veins referred for surgery	Control	No preoperative test	Replacement	To determine the precise location of obstruction in order to guide subsequent surgical intervention
			Experimental	Preoperative Duplex US		
T59	2–arm Standard Parallel	Hospitalised patients with suspected lower respiratory tract infection acquired during mechanical ventilation	Control	Conventional microbiology [Gram Stain (tracheal aspirate)]	Additional	To determine the cause of infection, and identify if organisms are susceptible to antimicrobial treatment, in order to guide subsequent treatment
			Experimental	Conventional microbiology + Antimicrobial susceptibility E-test		
T60	2–arm Standard Parallel	Emergency patients presenting with acute lower right quadrant abdominal pain, suspected appendicitis	Control	Mandatory CT	Triage	To confirm appendicitis and inform the treatment approach
			Experimental	Signs and Symptoms and lab tests ± CT		
T61	2–arm Standard Parallel	Women in labour showing a nonreassuring cardiotocograph	Control	CTG only	Additional	To confirm the presence of fetal distress and inform the need for assisted delivery
			Experimental	CTG and FPO		
T62	2–arm Factorial 2x2	Hospitalised patients with confirmed pneumonia due to mechanical ventilation	Control	Endotracheal aspiration with nonquantitative culture	Replacement	To identify the microorganisms causing infection and guide subsequent treatment
			Experimental	Bronchoalveolar lavage with quantitative culture		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T63	2–arm 2x parallel randomisations	Patients with suspected pulmonary embolism referred to a thrombosis service	Control #1	V/Q Scan (Low)	Replacement	To rule out pulmonary embolism and identify those in need of treatment
			Control #2	Serial US (high)		
			Exp #1	No Test	Replacement	
			Exp #2	No Test		
T64	2–arm Cluster Parallel	Patients managed in secondary care with chronic daily musculoskeletal pain necessitating regular analgesics	Control	No Test	Replacement	To identify the severity of pain more precisely and thus enable prescription of more appropriate analgesics
			Experimental	Visual Analogue Scale for pain intensity and Hospital Anxiety and Depression scale		
T65	3–arm Partial Random disclosure	Patients presenting to the GP with a first presentation of dyspepsia	Control	Prompt endoscopy	Triage	To determine the cause of dyspepsia and identify those in need of further investigations and/or treatment
			Exp #1	No test (Empirical treatment)		
			Exp #2	13-Carbon urea breath test	Triage	
T66	2–arm Cluster Parallel	Primary care patients with major depression	Control	Untrained clinical examination	Replacement	To determine the degree of depression and establish the need for treatment
			Experimental	Trained clinical examination		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T67	2–arm Standard Parallel	Women in labour at low risk of complications	Control	CTG only (FPO masked)	Additional	To identify fetuses in distress that require assistance in delivery
			Experimental	CTG and FPO		
T68	2–arm Standard Parallel	Emergency patients with mild head injury	Control	Observation in hospital	Replacement	To rule out intracerebral injury and determine the need for treatment
			Experimental	Immediate CT		
T69	2–arm Random Disclosure	General medical outpatients with hypertension	Control	Routine clinical work-up (prior tests) [Echo results withheld (information on left ventricular hypertrophy through ECG only)]	Additional	To determine the presence/extent of target organ damage in order to inform the need for treatment
			Experimental	Echocardiogram + Routine clinical work-up (prior tests)		
T70	2–arm Standard Parallel	Patients with peripheral arterial disease scheduled for surgery, referred for pre–operative assessment	Control	Duplex US	Replacement	To determine the location and extent of stenoses in order to guide subsequent surgery
			Experimental	Contrast-enhanced MR angiography		
T71	3–arm Cluster Parallel	Primary care patients presenting with dyspepsia (at least 2 week's duration) and decision taken to investigate/treat	Control	Hp test (13C Urea breath test) and eradicate	Replacement (EXP1)	To identify the cause of dyspepsia and clarify the need for further investigation and/or treatment
			Exp #1	No test (Empirical therapy 1wk)		
			Exp #2	No test (Empirical therapy 1wk) ± Hp test and eradicate with continuing symptoms.	Triage (EXP2)	

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T72	2–arm Standard Parallel	Emergency patients with suspected blunt or penetrating trauma to torso	Control	Standard testing (several tests, including CT)	Additional	To diagnose the presence, location and nature of trauma to the torso in order to inform the need for further investigation and /or treatment
			Experimental	Focussed assessment with sonography + standard testing		
T73	2–arm Standard Parallel	Women referred for investigation of subfertility	Control	Laparoscopy	Triage	To detect obstruction of fallopian tubes and treat as indicated
			Experimental	Hysterosalpingography +/- Laparoscopy		
T74	2–arm Double blind Parallel	Hospitalised patients with suspected pulmonary embolism with referral for V/Q scan	Control	V/Q Scan (BIOPED masked)	Triage	To confirm the presence of embolus and inform the need for treatment
			Experimental	BIOPED (3 tests: D-dimer, 7-variable clinical model [Well's], AVDSf [Alveolar dead space fraction]) ± V/Q Scan		
T75	3–arm Standard Parallel	Patients requiring emergency neurosurgical consultation in secondary or tertiary care	Control #1	Telephone consultation	Replacement	To investigate the cause of neurological symptoms and identify those in need of further investigations and/or treatment in a neurosurgical centre
			Control #2	Teleradiology	Replacement	
			Experimental	Video-conferencing		
T76	2–arm Standard parallel	Women undergoing termination of pregnancy	Control	No Test	Replacement	To detect retained products of conception and treat accordingly
			Experimental	Transvaginal ultrasound immediately following surgery		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T77	2–arm Standard parallel	Secondary care patients with gastrointestinal symptoms referred for either upper GI endoscopy or flexible sigmoidoscopy	Control	Nurse-led upper gastrointestinal endoscopy or Flexible Sigmoidoscopy	Replacement	To determine the cause of symptoms and determine the need for treatment
			Experimental	Doctor-led upper gastrointestinal endoscopy or Flexible Sigmoidoscopy		
T78	2–arm Standard Parallel	Women (at low-risk of complications) in labour with suspected fetal distress	Control	CTG	Additional	To confirm fetal distress and inform the need for assisted delivery
			Experimental	CTG + ST-segment analysis		
T79	2–arm Standard Parallel	Women (at high-risk of complications) in labour	Control	Biophysical profile	Replacement	To detect fetal distress and inform the need for assisted delivery
			Experimental	Modified biophysical profile		
T80	2–arm Standard Parallel	Women in labour with a medical indication for electronic fetal monitoring	Control	CTG + Fetal Blood Sampling	Replacement	To confirm fetal distress and inform the need for assisted delivery
			Experimental	CTG + ST–segment analysis		
T81	2–arm Standard Parallel	Women (at low-risk of complications) in labour	Control	Notes	Additional	To detect fetal distress and inform the need for assisted delivery
			Experimental	Partogram + Notes		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T82	2–arm Standard Parallel	Emergency patients with acute chest pain indicative of cardiac ischaemia, but at low-risk of cardiac events	Control	Standard "rule-out myocardial infarction" diagnostic protocol	Triage	To rule out myocardial infarction, and identify those in need of further investigation and/or treatment
			Experimental	Multislice CT ± Standard "rule-out myocardial infarction" diagnostic protocol		
T83	2–arm Standard Parallel	Secondary care patients with symptoms suggesting of coronary artery disease	Control	Exercise ECG	Replacement	To confirm the presence of coronary artery disease and determine the need for treatment
			Experimental	Gated SPECT MPI		
T84	2–arm Standard Parallel (Blinding optional per centre)	Patients with clinically suspected recurrence of laryngeal carcinoma after radiotherapy, with indication for direct laryngoscopy with biopsy	Control	Layngoscopy (direct laryngoscopy under general anaesthesia with taking of biopsies)	Triage	To confirm recurrence and indicate the need for further investigation
			Experimental	FDG–PET based strategy (only direct laryngoscopy under general anaesthesia with taking of biopsies if FDG–PET is positive or equivocal)		
T85	2–arm Standard Parallel	emergency and hospitalised patients with non-specific acute abdominal pain	Control	Plain X-ray - abdominal and erect chest	Replacement	To identify the cause of symptoms and determine the need for treatment
			Experimental	CT - Early abdominopelvic		
T86	2–arm Standard Parallel	Emergency patients with a diagnosis of transient ischemic attack	Control	Assessment in hospital	Replacement	To detect progression to full stroke and indicate the need for further investigation and/or treatment
			Experimental	Accelerated diagnostic protocol in emergency department		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T87	2–arm Standard Parallel	Women undergoing fertility treatment at risk of passing on B-thalassaemia	Control	Blastomere (cleavage stage) biopsy (Day 3)	Replacement	To rule out homozygous B-thalassaemia, and determine the suitability of further treatment
			Experimental	Trophectoderm (blastocyte stage) biopsy (Day 5)		
T88	2–arm Cross-over if negative	Hospitalised patients referred for investigation of an obscure gastrointestinal bleed	Control	Push Enteroscopy (+ Capsule Endoscopy if negative)	Replacement	To confirm the bleed and identify its location in order to guide treatment
			Experimental	Capsule Endoscopy (+ Push Enteroscopy if negative)		
T89	2–arm Standard parallel	Emergency patients with acute chest pain, normal or indeterminate ECG and multiple risk factors for CAD	Control	Exercise ECG	Replacement	To identify patients with significant underlying coronary artery disease and inform the need for treatment
			Experimental	Stress echocardiography within 24hrs admission		
T90	2–arm Standard Parallel	Patients with endoscopically suspected bladder carcinoma	Control	White-light cystoscopy with resection	Additional	To confirm the presence and location of carcinoma to guide the treatment approach
			Experimental	White-light and Fluorescence cystoscopy with resection		
T91	2–arm Standard Parallel	Patients with clinically suspected acute pulmonary embolism referred to a thrombosis clinic	Control	V/Q Scan	Replacement	To confirm the presence of pulmonary embolism and direct the need for further investigation and/or treatment
			Experimental	CT Pulmonary Angiography		
T92	4–arm Standard Parallel	Patients with known or suspected coronary artery disease referred for non-urgent coronary angiography to a tertiary centre	Control	Coronary angiography	Triage	To determine the functional significance of coronary stenosis, and determine the suitability for revascularisation
			Exp #1	SPECT		
			Exp #2	Stress cardiac MRI	Triage	
			Exp #3	Stress Echo	Triage	

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T93	2–arm Standard Parallel	Women undergoing fertility treatment at risk of passing on a genetic abnormality	Control	1-cell biopsy, PCR and/or FISH	Replacement	To rule out genetic abnormality and inform the suitability of further treatment
			Experimental	2-cell biopsy, PCR and/or FISH		
T94	2–arm Random disclosure	Women admitted for suspected preterm labour	Control	Clinical examination (speculum) and observation	Additional	To identify women at high risk of preterm labour and inform the suitability of subsequent treatment
			Experimental	Clinical examination + Transvaginal US (for cervical length) + FFN		
T95	2–arm Standard Parallel	Secondary care patients with clinically suspected stones of the bile duct	Control	ERCP	Replacement	To confirm the presence of stones and remove according to indication
			Experimental	EUS ± ERCP		
T96	2–arm Standard Parallel	Patients with severe left ventricle dysfunction due to coronary artery disease	Control	No Test	Replacement	To identify viable recoverable myocardium in order to guide subsequent revascularisation
			Experimental	FDG-PET		
T97	4–arm Factorial 2x2, cluster	Patients presenting to primary care with an acute cough, suspected lower respiratory tract infection	Control	Usual Care	Replacement	To diagnose the presence of pneumonia and direct the prescription of antibiotics
			Experimental	Point–of–care C-reactive protein test		
T98	2–arm Double Blind Parallel	Women undergoing fertility treatment at risk of passing on a genetic abnormality due to advanced age	Control	Embryo morphologic score	Additional	To determine health of embryo, and suitability for implantation
			Experimental	Preimplantation genetic screen + Morphological score		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T99	2–arm Standard Parallel	Hospitalised patients with chest pain indicative of cardiac ischaemia	Control	No Test +/- Angiography	Replacement	To confirm the presence of coronary artery disease and inform further treatment
			Experimental	Angiography (within 48hrs)		
T100	2–arm Standard Parallel	Emergency patients with dyspnoea of suspected cardiac origin	Control	Standard battery of tests only	Additional	To confirm the presence of heart failure and inform the need for further investigation and /or treatment
			Experimental	NT-proBNP		
T101	2–arm Standard Parallel	Patients with epicardial stenosis in whom multivessel percutaneous coronary intervention is planned	Control	Coronary angiography	Additional	To determine which stenoses are causing ischaemia in order to inform the subsequent treatment approach
			Experimental	Coronary angiography AND fractional flow reserve		
T102	2–arm Standard Parallel	Women suspected of preterm labour	Control	No Test	Replacement	To confirm the likelihood of preterm labour and inform the need for treatment
			Experimental	Transvaginal US		
T103	2–arm Random disclosure	Patients with suspected knee derangement referred for arthroscopy	Control	None (Surgeon given blank MRI report card)	Additional	To rule out a cause that requires therapeutic arthroscopy
			Experimental	MRI (Surgeon given MRI report)		
T104	2–arm Standard Parallel	Women with primary infertility and history of previous failed in–vitro fertilisation	Control	No Test	Replacement	To identify uterine abnormalities causing failure and treat as indicated
			Experimental	Hysteroscopy		
T105	2–arm Standard Parallel	Outpatients attending a gastroenterology unit with dyspeptic symptoms and positive 13C-urea breath test	Control	No Test	Replacement	To identify the microorganism strain causing infection in order to guide subsequent treatment
			Experimental	Upper gastrointestinal Endoscopy with biopsy for culture and susceptibility test		

Trial Ref	Trial Design	Patient presentation	Arm	Diagnostic intervention	Comparison	Purpose of test
T106	2–arm Standard Parallel	Emergency patients with chest pain indicative of cardiac ischaemia at intermediate risk of cardiac events	Control	Stress Test (Treadmill exercise using modified Bruce protocol; dobutamine stress echocardiography; nuclear stress test with sestamibi imaging) [+ masked Troponins]	Replacement	Investigation to diagnose the possibility of myocardial ischaemia and determine the need for admission with further investigation and/or treatment
			Experimental	Troponin-I		
T107	2–arm Standard Parallel	Patients presenting to primary care with suspected internal derangement of the knee and referral to orthopaedic	Control	Orthopaedic consultation	Replacement	To confirm the presence of injury or disease requiring therapeutic arthroscopy
			Experimental	MRI +/- orthopaedic consultation		
T108	2–arm Standard Parallel	Secondary care patients with clinically suspected gall bladder stones and bile duct stones	Control	ERCP	Triage	To detect bile duct stones and treat as indicated
			Experimental	Intraoperative cholangiography (during laparoscopic cholecystectomy) +/- intraoperative ERCP		

Appendix C:
Additional publications

This appendix lists the additional publications, associated with included test-treatment trials, that were used to supplement the analysis of reporting quality (chapter 5) and methodological quality (chapter 6).

T4

Ash LM, Modic MT, Obuchowski NA, Ross JS, Brant-Zawadzki MN, Grooff PN. Effects of diagnostic information, per se, on patient outcomes in acute radiculopathy and low back pain. AJNR Am J Neuroradiol 2008;29:1098–1103.

T6

Fernandez-Avilés F, Alonso JJ, Castro-Beiras A, Vázquez N, Blanco J, Alonso-Briales J, et al. Routine invasive strategy within 24 hours of thrombolysis versus ischaemia-guided conservative approach for acute myocardial infarction with ST-segment elevation (GRACIA-1): a randomised controlled trial. Lancet 2004;364:1045–1053.

T12

Adriaensen ME, Kock MC, Stijnen T, van Sambeek MR, van Urk H, Pattynama PM, et al. Peripheral arterial disease: therapeutic confidence of CT versus digital subtraction angiography and effects on additional imaging recommendations. Radiology 2004;233:385–391.

T14

Cannon CP, Weintraub WS, Demopoulos LA, Robertson DH, Gormley GJ, Braunwald E. Invasive versus conservative strategies in unstable angina and non-Q-wave myocardial infarction following treatment with tirofiban: rationale and study design of the international TACTICS-TIMI 18 Trial. Treat Angina with Aggrastat and determine Cost of Therapy with an Invasive or Conservative Strategy. Thrombolysis In Myocardial Infarction. Am J Cardiol 1998;82:731–736.

Weintraub WS, Culler SD, Kosinski A, Becker ER, Mahoney E, Burnette J, et al. Economics, health-related quality of life, and cost-effectiveness methods for the TACTICS (Treat Angina With Aggrastat [tirofiban]] and Determine Cost of Therapy with Invasive or Conservative Strategy)-TIMI 18 trial. Am J Cardiol 1999;83:317–322.

Cannon CP, Weintraub WS, Demopoulos LA, Vicari R, Frey MJ, Lakkis N, et al. Comparison of early invasive and conservative strategies in patients with unstable coronary syndromes treated with the glycoprotein IIb/IIIa inhibitor tirofiban. N Engl J Med 2001;344:1879–1887.

Bach RG, Cannon CP, Weintraub WS, DiBattiste PM, Demopoulos LA, Anderson HV, et al. The effect of routine, early invasive management on outcome for elderly patients with non-ST-segment elevation acute coronary syndromes. Ann Intern Med 2004;141:186–195.

T15

Elbourne D, Dezateux C, Arthur R, Clarke NM, Gray A, King A, et al. Ultrasonography in the diagnosis and management of developmental hip dysplasia (UK Hip Trial): clinical and economic results of a multicentre randomised controlled trial. Lancet 2002;360:2009–2017.

T16

Farkouh ME, Smars PA, Reeder GS, Zinsmeister AR, Evans RW, Meloy TD, et al. A clinical trial of a chest-pain observation unit for patients with unstable angina. Chest Pain Evaluation in the Emergency Room (CHEER) Investigators. N Engl J Med 1998;339:1882–1888.

T18

Cohen HJ, Feussner JR, Weinberger M, Carnes M, Hamdy RC, Hsieh F, et al. A controlled trial of inpatient and outpatient geriatric evaluation and management. N Engl J Med 2002;346:905–912.

T32

Udelson JE, Beshansky JR, Ballin DS, Feldman JA, Griffith JL, Handler J, et al. Myocardial perfusion imaging for evaluation and triage of patients with suspected acute cardiac ischemia: a randomized controlled trial. JAMA 2002;288:2693–2700.

T34

Little P, Turner S, Rumsby K, Warner G, Moore M, Lowes JA, et al. Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort and qualitative study. Health Technol Assess 2009;13:1–73.

T36

Gilbert FJ, Grant AM, Gillan MG, Vale LD, Campbell MK, Scott NW, et al. Low Back Pain: Influence of Early MR Imaging or CT on Treatment and Outcome—Multicenter Randomized Trial. Radiology 2004;231:343–351.

T38

The ACUTE Study Investigators. Design of a clinical trial for the assessment of cardioversion using transesophageal echocardiography (The ACUTE Multicenter Study). Steering and Publications Committees of the ACUTE Study. Am J Cardiol 1998;81:877–883.

Klein AL, Grimm RA, Murray RD, Apperson-Hansen C, Asinger RW, Black IW, et al. Use of transesophageal echocardiography to guide cardioversion in patients with atrial fibrillation. N Engl J Med 2001;344:1411–1420.

Klein AL, Grimm RA, Jasper SE, Murray RD, Apperson-Hansen C, Lieber EA, et al. Efficacy of transesophageal echocardiography-guided cardioversion of patients with atrial fibrillation at 6 months: a randomized controlled trial. Am Heart J 2006;151:380–389.

T41

Saririan M, Cugno S, Blankenship J, Huynh T, Sedlis S, Starling M, et al. Routine versus selective functional testing after percutaneous coronary intervention in patients with diabetes mellitus. J Invasive Cardiol 2005;17:25–29.

T44

Lassen AT, Pedersen FM, Bytzer P, Schaffalitzky de Muckadell OB. Helicobacter pylori test-and-eradicate versus prompt endoscopy for management of dyspeptic patients: a randomised trial. Lancet 2000;356:455–460.

T46

Laheij RJ, Severens JL, Jansen JB, van de Lisdonk EH, Verbeek AL. Management in general practice of patients with persistent dyspepsia. A decision analysis. J Clin Gastroenterol 1997;25:563–567.

Laheij RJF, Stevens JL, Van De Lisdonk EH, Verbeek ALM, Jansen JBMJ. Randomized controlled trial of omeprazole or endoscopy in patients with persistent dyspepsia: a cost–effectiveness analysis. Aliment Pharmacol Ther 1998;12:1249–1256.

T47

Bryan S, Weatherburn G, Bungay H, Hatrick C, Salas C, Parry D, et al. The cost-effectiveness of magnetic resonance imaging for investigation of the knee joint. Health Technol Assess 2001;5:1–95

T48

Goodacre SW, Morris FM, Campbell S, Arnold J, Angelini K. A prospective, observational study of a chest pain observation unit in a British hospital. Emerg Med J 2002;19:117–121.

Goodacre S, Nicholl J. A randomised controlled trial to measure the effect of chest pain unit care upon anxiety, depression, and health-related quality of life [ISRCTN85078221]. Health Qual Life Outcomes 2004;2:39–47.

Goodacre S, Nicholl J, Dixon S, Cross E, Angelini K, Arnold J, et al. Randomised controlled trial and economic evaluation of a chest pain observation unit compared with routine care. BMJ 2004;328:254–230.

Goodacre S, Dixon S. Is a chest pain observation unit likely to be cost effective at my hospital? Extrapolation of data from a randomised controlled trial. Emerg Med J 2005;22:418–422.

T49

Kiss H, Pichler E, Petricevic L, Husslein P. Cost effectiveness of a screen-and-treat program for asymptomatic vaginal infections in pregnancy: towards a significant reduction in the costs of prematurity. Eur J Obstet Gynecol Reprod Biol 2006;127:198–203.

T55

Mueller C, Laule-Kilian K, Scholer A, Frana B, Rodriguez D, Schindler C, et al. Use of B-type natriuretic peptide for the management of women with dyspnea. Am J Cardiol 2004;94:1510–1514.

Mueller C, Laule-Kilian K, Frana B, Rodriguez D, Rudez J, Scholer A, et al. The use of B-type natriuretic peptide in the management of elderly patients with acute dyspnoea. J Intern Med 2005;258:77–85.

Mueller C, Laule-Kilian K, Scholer A, Nusbaumer C, Zeller T, Staub D, et al. B-type natriuretic peptide for acute dyspnea in patients with kidney disease: insights from a randomized comparison. Kidney Int 2005;67:278–284.

Mueller C, Laule-Kilian K, Schindler C, Klima T, Frana B, Rodriguez D, et al. Cost-effectiveness of B-type natriuretic peptide testing in patients with acute dyspnea. Arch Intern Med 2006;166:1081–1087.

T56

Farwell DJ, Freemantle N, Sulke N. The clinical impact of implantable loop recorders in patients with syncope. Eur Heart J 2006;27:351–356.

T57

Wallace P, Haines A, Harrison R, Barber J, Thompson S, Jacklin P, et al. Joint teleconsultations (virtual outreach) versus standard outpatient appointments for patients referred by their general practitioner for a specialist opinion: a randomised trial. Lancet 2002;359:1961–1968.

Wallace P, Haines A, Harrison R, Barber JA, Thompson S, Roberts J, et al. Design and performance of a multi-centre randomised controlled trial and economic evaluation of joint tele-consultations [ISRCTN54264250]. BMC Fam Pract 2002;3:1–9.

Jacklin PB, Roberts JA, Wallace P, Haines A, Harrison R, Barber JA, et al. Virtual outreach: economic evaluation of joint teleconsultations for patients referred by their general practitioner for a specialist opinion. BMJ 2003;327:84–92.

T58

Blomgren L, Johansson G, Bergqvist D. Quality of life after surgery for varicose veins and the impact of preoperative duplex: results based on a randomized trial. Ann Vasc Surg 2006;20:30–34.

T61

East CE, Brennecke SP, King JF, Chan FY, Colditz PB, FOREMOST Study Group. The effect of intrapartum fetal pulse oximetry, in the presence of a nonreassuring fetal heart rate pattern, on operative delivery rates: a multicenter, randomized, controlled trial (the FOREMOST trial). Am J Obstet Gynecol 2006;194:606.e1-16.

T66

Bijl D, van Marwijk HWJ, Beekman ATF, de Haan M, van Tilburg W. A randomized controlled trial to improve the recognition, diagnosis and treatment of major depression in elderly people in general practice: design, first results and feasibility of the West Friesland Study. Primary Care Psych 2003;8:135–140.

T68

af Geijerstam JL, Oredsson S, Britton M; OCTOPUS Study Investigators. Medical outcome after immediate computed tomography or admission for observation in patients with mild head injury: randomised controlled trial. BMJ 2006;333:465–471.

T71

Jarbol DE, Bech M, Kragstrup J, Havelund T, Schaffalitzky de Muckadell OB. Economic evaluation of empirical antisecretory therapy versus Helicobacter pylori test for management of dyspepsia: a randomized trial in primary care. Int J Technol Assess Health Care 2006;22:362–371.

T73

Perquin DA, Beersma MF, de Craen AJ, Helmerhorst FM. The value of Chlamydia trachomatis-specific IgG antibody testing and hysterosalpingography for predicting tubal pathology and occurrence of pregnancy. Fertil Steril 2007;88:224–226.

T89

Jeetley P, Burden L, Senior R. Stress echocardiography is superior to exercise ECG in the risk stratification of patients presenting with acute chest pain with negative Troponin. Eur J Echocardiogr 2006;7:155–164.

T90

Burger M, Zaak D, Stief CG, Filbeck T, Wieland WF, Roessler W, et al. Photodynamic diagnostics and noninvasive bladder cancer: is it cost-effective in long-term application? A Germany-based cost analysis. Eur Urol 2007;52:142–147.

Filbeck T, Pichlmeier U, Knuechel R, Wieland WF, Roessler W. Clinically relevant improvement of recurrence-free survival with 5-aminolevulinic acid induced fluorescence diagnosis in patients with superficial bladder tumors. J Urol 2002;168:67–71.

T96

Beanlands R, Nichol G, Ruddy TD, deKemp RA, Hendry P, Humen D, et al. Evaluation of outcome and cost-effectiveness using an FDG PET-guided approach to management of patients with coronary disease and severe left ventricular dysfunction (PARR-2): rationale, design, and methods. Control Clin Trials 2003;24:776–794.

T97

Cals JW, Butler CC, Hopstaken RM, Hood K, Dinant GJ. Effect of point of care testing for C reactive protein and training in communication skills on antibiotic use in lower respiratory tract infections: cluster randomised trial. BMJ 2009;338:b1374.

Cals JW, Chappin FH, Hopstaken RM, van Leeuwen ME, Hood K, Butler CC, et al. C-reactive protein point-of-care testing for lower respiratory tract infections: a qualitative evaluation of experiences by GPs. Fam Pract 2010;27:212–218.

T99

de Winter RJ, Windhausen F, Cornel JH, Dunselman PH, Janus CL, Bendermacher PE, et al. Early invasive versus selectively invasive management for acute coronary syndromes. N Engl J Med 2005;353:1095–1104.

T101

Tonino PA, De Bruyne B, Pijls NH, Siebert U, Ikeno F, van' t Veer M, et al. Fractional flow reserve versus angiography for guiding percutaneous coronary intervention. N Engl J Med 2009;360:213–224.

T107

Brealey SD, DAMASK (Direct Access to Magnetic Resonance Imaging: Assessment for Suspect Knees) Trial Team. Influence of magnetic resonance of the knee on GPs' decisions: a randomised trial. Br J Gen Pract 2007;57:622–629.

DAMASK (Direct Access to Magnetic Resonance Imaging: Assessment for Suspect Knees) Trial Team. Cost-effectiveness of magnetic resonance imaging of the knee for patients presenting in primary care. Br J Gen Pract 2008;58:e10–e16.

DAMASK (Direct Access to Magnetic Resonance Imaging: Assessment for Suspect Knees) Trial Team. Effectiveness of GP access to magnetic resonance imaging of the knee: a randomised trial. Br J Gen Pract 2008;58:e1–e8.

Appendix D:
Reporting quality summary data

Below are listed summaries of the data abstraction for appraising the reporting quality of included trials, which was presented in chapter 5.

The following abbreviations are used throughout:

CCU	Coronary care unit
ECG	Electrocardiogram
ED	Emergency department
FISH	Fluorescence in situ hybridization
PCR	Polymerase chain reaction
QoL	Quality of Life

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T1	Control	Yes	No	Yes	No	Partial	1	Yes	Recurrence free interval (time from transurethral resection to initial recurrence)	Yes
	Experimental	Yes	No	Yes	No					
T2	Control	Yes	No	No	No	Partial	n/a	No	Not defined	No
	Exp #1	Yes	No	No	No					
	Exp #2	Yes	No	No	No					
T3	Control	Tech	No	No	No	Fully	1	Yes	Symptomatic venous thrombosis rate	Yes
	Experimental	Yes	Yes	No	No					
T4	Control	No	No	No	Name only	Partial	7	Yes	Patient satisfaction	Partial
	Experimental	Yes	No	No	Name only			Yes	QoL	Partial
								Yes	Visual pain analog scale	No
								Yes	Absenteeism	No
								Yes	Self-efficacy perception score	No
								Yes	Fear avoidance questionnaire score	No
								Power calculation	Patients with 50% or more improvement in function	Yes

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T5	Control	Yes	No	No	No	Fully	1	Power calculation	Mean total cost	Yes
	Experimental	Yes	No	No	No					
T6	Control	No	Partial	No	No	Partial	1	Yes	Combined rate of death, non-fatal reinfarction, or ischaemia-based post-discharge revascularisation	Yes
	Experimental	Tech	Partial	Yes	Yes					
T7	Control	No	No	No	Name only	Partial	3	Study aim	% prescribed amoxicillin (non-pneumonic acute Upper/Lower respiratory tract infection)	Yes
	Experimental	Yes	Yes	No	Name only			Study aim	% patients prescribed antibiotics (non-pneumonic Upper/Lower respiratory tract infection)	Yes
								Study aim	% prescribed new extended spectrum antibiotics (non-pneumonic Upper/Lower respiratory tract infection)	Yes

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T8	Control	Tech	Yes	Yes	Yes	Partial	1	Yes	% women with unripe cervix (given prostaglandin as preinduction agent)	Yes
	Experimental	Tech	Yes	Yes	Yes					
T9	Control	Yes	No	Yes	No	Partial	1	Power calculation	Morbidity of acute pancreatitis	No
	Experimental	Yes	Yes	Yes	No					
T10	Control	No	No	No	No	Partial	1	Power calculation	Days unnecessarily immobilised	No
	Experimental	Yes	Partial	No	No					
T11	Control	No	No	No	Yes	Fully	2	Yes	Fecal incontinence (Any)	Yes
	Experimental	Tech	No	Yes	Yes			Yes	Fecal incontinence (Severe)	Yes
T12	Control	Yes	Yes	No	No	Fully	1	Power calculation	Mean strategy cost per patient (mean total diagnostic cost)	Yes
	Experimental	Yes	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T13	Control	N/a	n/a	n/a	No	Fully	1	Yes	Symptom-free status	Partial
	Exp #1	No	No	No	No					
	Exp #2	No	No	No	No					
	Exp #3	No	No	No	No					
T14	Control	Tech	Yes	No	No	Partial	1	Yes	Combined incidence of death, nonfatal myocardial infarction, and rehospitalisation for an acute coronary syndrome	Partial
	Experimental	Tech	No	No	No					
T15	Control	No	No	Yes	Name only	Partial	1	Yes	Radiological appearance abnormal	Yes
	Experimental	Yes	No	Yes	Name only					
T16	Control	No	No	No	No	Fully	2	Yes	Occurrence of serious adverse event (nonfatal MI, death, acute congestive heart failure, stroke, out-of-hospital cardiac arrest)	Partial
	Experimental	Yes	Yes	Yes	No			Yes	Event-free survival	No

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T17	Control	Yes	Yes	Yes	No	Partial	1	Yes	Number of futile thoracotomies	Yes
	Experimental	Yes	Yes	Yes	No					
T18	Control	N/a	n/a	No	No	Partial	9	Yes	Survival (all-cause mortality)	Yes
	Exp #1	Part name only	No	No	No			Yes	Health-related QoL: Physical functioning	Yes
	Exp #2	N/A	N/A	No	No			Yes	Health-related QoL: Physical limitations	Yes
	Exp #3	Name only	No	No	No			Yes	Health-related QoL: Emotional limitations	Yes
								Yes	Health-related QoL: bodily pain	Yes
								Yes	Health-related QoL: energy	Yes
								Yes	Health-related QoL: mental health	Yes
								Yes	Health-related QoL: social activity	Yes
								Yes	Health-related QoL: general health	Yes

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T19	Control	Yes	Yes	No	No	Fully	1	Yes	Mean total diagnostic cost per patient	Yes
	Experimental	Yes	Yes	No	No					
T20	Control	Tech	No	Yes	No	Partial	1	Yes	Maximal endurance by exercise testing	Yes
	Experimental	Tech	single	Yes	No					
T21	Control	Yes	No	No	No	Partial	1	Yes	Change in antibiotic treatment based on results of PCR	Partial
	Experimental	Yes	No	No	No					
T23	Control	Yes	No	No	No	Partial	1	Power calculation	Mortality (All-cause)	Partial
	Exp #1	Yes	No	No	No					
	Exp #2	Yes	No	No	No					
	Exp #3	Yes	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T24	Control	Tech	No	No	Name only	Fully	15	Yes	Physical morbidity: Swelling - arm volume changes	Partial
	Experimental	Yes	No	Yes	No			Yes	physical morbidity: shoulder mobility - Flexion	Partial
								Yes	Psychological morbidity: Depression	Partial
								Yes	Psychological morbidity: Anxiety	Partial
								Yes	Psychological morbidity: psychiatric disturbance	Partial
								Yes	Psychological morbidity: overall distress	Partial
								Yes	Psychological morbidity: coping responses	Partial
								Yes	Physical morbidity: Swelling - lymphedema	Partial

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T24 cont.								Yes	physical morbidity: shoulder mobility - Extension	Partial
								Yes	physical morbidity: shoulder mobility - Abduction	Partial
								Yes	physical morbidity: shoulder mobility - Internal rotation	Partial
								Yes	physical morbidity: shoulder mobility - External rotation	Partial
								Yes	physical morbidity: any paresthesia	No
								Yes	physical morbidity: seroma formation present	No
								Yes	physical morbidity: seroma formation aspirated	No
T26	Control	Yes	No	No	No	Fully	2	Yes	inhospital cardiac catheterization (for patients who were admitted)	No
	Experimental	Yes	Yes	No	No			Yes	ED discharge	Yes

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T27	Control	Tech	Yes	Yes	Yes	Fully	1	Yes	% abnormal laparoscopies leading to a change in treatment	No
	Experimental	Yes	Yes	Yes	Yes					
T28	Control	N/a	n/a	No	No	Partial	3	Study aim	Health-related QoL	Yes
	Experimental	Tech	No	No	No			Study aim	Functional disability	Yes
								Study aim	Pain	Yes
T29	Control	No	No	No	No	Partial	n/a	No	Not defined	No
	Experimental	Tech	No	No	No					
T30	Control	Tech	Yes	Yes	No	Fully	2	Yes	Rebleeding - early	Partial
	Experimental	Yes	Yes	Yes	Yes			Yes	Rebleeding - late	Partial
T31	Control	Tech	Yes	Yes	No	Fully	n/a	No	Not defined	No
	Experimental	Tech	Yes	Yes	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T32	Control	No	No	No	No	Fully	2	Yes	Appropriateness of initial ED triage decision (Inppropriate discharge home of patients with acute cardiac ischemia)	Partial
	Experimental	Yes	Yes	No	No			Yes	Appropriateness of initial ED trage decision (inappropriate hospitalisation of patients without acute cardiac ischemia)	Partial
T33	Control	Tech	No	No	No	Partial	1	Study aim	Rate of ocular surgical complications	Yes
	Experimental	Tech	No	No	No					
T34	Control	No	No	No	Yes	Partial	1	Yes	Daily symptoms of urinary tract infection	Partial
	Exp #1	No	No	No	No					
	Exp #2	No	Yes	Yes	Yes					
	Exp #3	Yes	Yes	Yes	Yes					
	Exp #4	Yes	Yes	Yes	Yes					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T35	Control	No	No	No	Yes	Fully	1	No	Ankle-Hindfoot function and pain (mean score)	Partial
	Experimental	Yes	Yes	Yes	Yes					
T36	Control	No	No	No	No	Fully	1	Power calculation	Lower back pain	Partial
	Experimental	No	No	No	No					
T37	Control	N/a	n/a	No	No	Fully	1	Yes	Incidence postoperative otorrhoea	Yes
	Experimental	Yes	No	No	No					
T38	Control	N/a	n/a	N/A	No	Partial	1	Yes	Composite incidence of Stroke, transient ischaemic attack or systemic embolism	Yes
	Experimental	No	Yes	Yes	No					
T39	Control	N/a	n/a	N/A	N/A	Partial	6	Yes	QoL: Physical mobility	Partial
	Experimental	Tech	No	Yes	No			Yes	QoL: social isolation	Partial
								Yes	QoL: Pain	Partial
								Yes	QoL: emotional reactions	Partial
								Yes	QoL: energy	Partial
								Yes	QoL: sleep	Partial

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T40	Control	No	No	No	No	Partial	1	Power calculation	Complete symptom relief	Partial
	Exp #1	Tech	No	Yes	Yes					
	Exp #2	No	No	No	No					
T41	Control	Tech	No	No	No	Partial	1	Yes	Functional status (Metabolic Equivalents)	Yes
	Experimental	Tech	No	No	No					
T42	Control	N/a	n/a	No	No	Fully	1	Yes	Cancer-related mortality	Yes
	Experimental	Tech	No	No	No					
T44	Control	No	No	Yes	Yes	Fully	1	Yes	Median % symptomless days	Yes
	Experimental	Tech	No	Yes	Yes					
T45	Control	No	No	No	No	Partial	1	Yes	Rate of ocular adverse events (intra- and post-operative)	Yes
	Experimental	N/A	N/A	No	No					
T46	Control	No	No	No	No	Partial	1	Power calculation	Number symptom-free days	Yes
	Experimental	Tech	No	Yes	Yes					
T47	Control	No	No	No	No	Fully	1	Power calculation	Rate of arthroscopy	No
	Experimental	No	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T48	Control	No	No	No	No	Fully	1	Yes	% admissions	Partial
	Experimental	Yes	Yes	Yes	No					
T49	Control	N/a	n/a	No	No	Partial	1	Yes	Rate of spontaneous preterm delivery	Partial
	Experimental	Tech	No	Yes	Yes					
T50	Control	No	No	No	No	Partial	1	Power calculation	Length of stay in antepartum ward	No
	Experimental	Yes	Yes	No	No					
T51	Control	Tech	No	No	No	Partial	4	Yes	Perceived fatalism	Partial
	Experimental	Yes	No	No	No			Yes	Perceived control over familial hypercholesterolaemia	Partial
								Yes	Perceived control over cholesterol	Partial
								Yes	Perceived control over heart disease	Partial
T52	Control	N/a	n/a	N/A	No	Fully	1	Yes	% patients able to avoid thoracotomy	Partial
	Experimental	Yes	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T53	Control	No	No	No	No	Partial	2	Yes	Hospital admission rate	Partial
	Experimental	Yes	No	No	No			Yes	Diagnostic yield (% patients with a diagnosis)	Partial
T54	Control	Tech	No	No	Name only	Partial	1	Yes	Total Cesarean section rate	Partial
	Experimental	Tech	No	No	Name only					
T55	Control	Tech	Yes	Yes	Yes	Partial	2	Yes	Total cost of treatment	Partial
	Experimental	Yes	Yes	Yes	No			Yes	Time to discharge	Yes
T56	Control	No	No	No	No	Partial	1	Yes	Time to ECG diagnosis	No
	Experimental	Yes	No	No	No					
T57	Control	No	No	No	No	Fully	1	Yes	% patients offered follow–up appointments after consultation	Yes
	Experimental	No	No	No	No					
T58	Control	N/a	n/a	No	Yes	Fully	1	Power calculation	Recurrence rate (varicose veins)	Yes
	Experimental	Yes	Yes	Yes	Yes					
T59	Control	Yes	No	No	No	Partial	n/a	No	Not defined	No
	Experimental	Yes	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T60	Control	Yes	No	No	No	Partial	1	Power calculation	Negative appendectomy rate	Partial
	Experimental	Yes	No	No	No					
T61	Control	Tech	No	Yes	No	Fully	1	Yes	Operative birth (cesarean, forceps, vacuum) for nonreassuring fetal status	Partial
	Experimental	Yes	No	Yes	No					
T62	Control	Yes	Yes	Yes	No	Partial	1	Yes	Mortality rate	Partial
	Experimental	Yes	Yes	Yes	No					
T63	Control #1	No	Yes	Yes	No	Fully	1	Yes	Frequency of venous thromboembolism	Yes
	Control #2	No	No	No	No					
	Exp #1	N/A	N/A	N/A	N/A					
	Exp #2	N/A	N/A	N/A	N/A					
T64	Control	N/a	n/a	No	No	None	1	Yes	Pain relief	Yes
	Experimental	Partly	No	No	No					
T65	Control	Name only	Yes	Yes	Yes	Fully	1	Power calculation	Mean symptom score	Partial
	Exp #1	No	No	No	Yes					
	Exp #2	Yes	No	Yes	Yes					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T66	Control	Tech	No	No	No	Fully	3	Yes	Severity of symptoms	Partial
	Experimental	Tech	No	No	Yes			Yes	Clinical Global Impression	No
								Yes	Recovery from depression	Yes
T67	Control	Tech	No	No	No	Fully	1	Yes	Rate overall cesarean delivery	Partial
	Experimental	Tech	No	No	No					
T68	Control	No	No	No	No	Fully	1	Yes	Full recovery	Yes
	Experimental	No	Partial	Yes	No					
T69	Control	No	Yes	Yes	No	Partial	1	Yes	Mean left ventricular mass index	Yes
	Experimental	Tech	Yes	Yes	No					
T70	Control	No	No	No	No	Fully	1	Yes	Mean total cost	Yes
	Experimental	No	No	No	No					
T71	Control	Tech	No	Yes	Yes	Fully	1	Yes	% days without dyspeptic symptoms	Yes
	Exp #1	n/a	n/a	n/a	Yes					
	Exp #2	Tech	No	Yes	Yes					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T72	Control	No	No	No	No	Fully	1	Yes	Mean time from ED arrival to direct transfer to operative care	Yes
	Experimental	Yes	No	No	No					
T73	Control	Yes	Yes	No	No	Partial	1	Yes	Pregnancy rate	Yes
	Experimental	Yes	Yes	Yes	No					
T74	Control	No	No	No	No	Fully	1	Yes	Recurrence of enous thromboembolism (in patients not taking anticoagulants)	Yes
	Experimental	Yes	Yes	No	No					
T75	Control #1	Tech	No	No	No	Partial	1	Yes	Favourable outcome (Glasgow outcome score 4-5)	Partial
	Control #2	Tech	No	No	No					
	Experimental	Tech	No	No	No					
T76	Control	N/A	N/A	N/A	N/A	Partial	1	Study aim	Incidence of retained products of conception	Partial
	Experimental	Yes	Yes	Yes	Yes					
T77	Control	No	No	No	No	Fully	1	Yes	Gastrointestinal Symptoms	Yes
	Experimental	No	No	No	No					
T78	Control	Tech	No	No	No	Partial	1	Yes	& operative delivery for non–reassuring fetal heart rate	Partial
	Experimental	Tech	Yes	Yes	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T79	Control	Tech	Partial	Yes	No	Partial	n/a	No	Not defined	No
	Experimental	No	No	No	No					
T81	Control	No	Yes	Yes	No	Partial	1	Yes	% Cesarean section	Yes
	Experimental	Yes	Yes	Yes	Yes					
T82	Control	Yes	Yes	Yes	No	Partial	n/a	No	Not Defined	No
	Experimental	Yes	Yes	Yes	No					
T83	Control	Yes	Yes	Yes	No	Partial	1	Yes	Mean diagnostic cost per patient	Yes
	Experimental	Yes	Unclear	Yes	No					
T85	Control	No	No	No	No	Partial	1	Yes	Length hospital stay	Yes
	Experimental	Yes	No	No	No					
T86	Control	No	No	No	No	Partial	1	Yes	Length of stay	Yes
	Experimental	No	No	Yes	Name only					
T87	Control	Yes	No	No	Yes	Partial	1	Study aim	Implantation rate	Partial
	Experimental	Yes	No	No	Yes					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T88	Control	Yes	Yes	No	No	Partial	1	Yes	Overall diagnostic yield (certain or highly probable sources of bleeding only)	Yes
	Experimental	Yes	Yes	No	No					
T89	Control	Yes	Yes	Yes	No	Partial	1	Yes	Cardiac death, non-fatal AMI or coronary revascularisation	Partial
	Experimental	Yes	Yes	Yes	No					
T90	Control	Yes	Yes	Yes	No	Fully	1	Yes	Time to first recurrence	Yes
	Experimental	Yes	Yes	Yes	No					
T91	Control	Yes	Yes	Yes	No	Fully	1	Yes	Frequency of venous thromboembolism (composite of proximal deep vein thrombosis or pulmonary embolism) in patients in whom pulmonary embolism was ruled out	Yes
	Experimental	Yes	Yes	Yes	No					
T92	Control	Yes	Yes	No	No	Fully	1	Yes	Exercise treadmill time	Yes
	Exp #1	Yes	No	No	No					
	Exp #2	Yes	No	No	No					
	Exp #3	Yes	Yes	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T93	Control	Yes	No	No	No	Partial	3	Yes	Live birth rate	Yes
	Experimental	Yes	No	No	No			Yes	Diagnostic Efficiency-PCR (% embryos with diagnosis)	Yes
								Yes	Diagnostic Efficiency-FISH (% embryos with diagnosis)	Yes
T94	Control	Tech	No	No	No	Fully	1	Yes	Time to discharge	Yes
	Experimental	Yes	Yes	Yes	No					
T95	Control	Tech	No	Yes	Name only	Fully	1	Yes	% patients with a negative Outcome	Yes
	Experimental	Yes	No	Yes	Name only					
T96	Control	N/a	n/a	No	Name only	Partial	1	Yes	Composite adverse event rate	Yes
	Experimental	Yes	Yes	Yes	Name only					
T97	Control	No	No	No	Name only	Partial	1	Yes	Antibiotic prescription	Yes
	Experimental	Yes	Yes	No	Name only					
T98	Control	Yes	Yes	Yes	Yes	Fully	1	Yes	Pregnancy (ongoing)	Yes
	Experimental	Yes	Yes	Yes	Yes					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T99	Control	Tech	No	Yes	Name only	Partial	1	Yes	Adverse cardiac events	Yes
	Experimental	Tech	No	Yes	Name only					
T100	Control	No	No	No	No	Partial	2	Yes	Length of ED stay	No
	Experimental	Tech	Yes	No	No			Yes	Total direct medical cost	Yes
T101	Control	Tech	No	No	No	Fully	1	Yes	% patients with major cardiac events	Yes
	Experimental	Yes	Yes	Yes	No					
T102	Control	n/a	n/a	n/a	Name only	Fully	1	Yes	% patients in whom pregnancy ongoing after corticosteroids	Yes
	Experimental	Tech	Yes	Yes	Name only					
T103	Control	N/a	n/a	No	Partial	Fully	1	Yes	% Patients not proceeding arthroscopy	No
	Experimental	Yes	No	No	Partial					
T104	Control	n/a	n/a	n/a	n/a	Partial	1	Yes	clinical pregnancy rate	No
	Experimental	Yes	No	No	No					

Trial Ref	Arm	Intervention description				Participant flow reported	Number primary outcomes	Outcome defined	Outcome name	Outcome Measurement method reported
		Test Method	Diagnostic Decision	Treatment Decision	Treatment Method					
T105	Control	N/a	n/a	n/a	Yes	Partial	1	Study aim	Eradication rate	Partial
	Experimental	Yes	Yes	Yes	Yes					
T106	Control	No	Yes	No	No	Partial	1	Yes	% Requiring admission to CCU	Yes
	Experimental	Yes	Yes	Yes	No					
T107	Control	No	No	No	No	Partial	2	Yes	Physical function (subscale SF-36)	Yes
	Experimental	No	No	Yes	Name only			Yes	Knee-specific QoL	Yes
T108	Control	Tech	No	No	No	Fully	1	Study aim	Morbidity %	Partial
	Experimental	Yes	No	No	Yes					

Appendix E:
Methodological quality summary data

This appendix lists summary data for all methodological quality items appraised during the review presented in chapter 6. It is arranged in two tables; the first presents quality judgements for reported methods of sequence generation and allocation concealment, as well as reports of blinding conduct and judgements of blinding feasibility. The second is a catalogue of sample sizes and methods of analysis.

The following abbreviations are used throughout:

NR	Not reported
n/a	Not applicable
Prot devs	Protocol deviations

E.1 Sequence generation, allocation concealment & blinding

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T1	Unclear	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T2	Unclear	Unclear	NR	NR	No primary outcome	Impossible	Impossible	No primary outcome	n/a
T3	Adequate	Adequate	NR	NR	NR	Impossible	Difficult	Feasible	Subjective
T4	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T5	Adequate	Unclear	Absent	Absent	NR	Feasible	Impossible	impossible	Objective
T6	Unclear	Adequate	Absent	Absent	Present	Impossible	Impossible	Feasible	Objective
T7	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T8	Adequate	Adequate	NR	NR	NR	Feasible	Impossible	impossible	Subjective
T9	Unclear	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T10	Adequate	Unclear	Absent	Absent	Absent	Feasible	Impossible	Unclear	Subjective
T11	Adequate	Adequate	Absent	Absent	Absent	Impossible	Impossible	impossible	Subjective
T12	Adequate	Adequate	Absent	Absent	NR	Impossible	Impossible	impossible	Objective
T13	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T14	Unclear	Unclear	NR	NR	Present	Difficult	Difficult	Feasible	Objective
T15	Adequate	Adequate	NR	NR	Present	Feasible	Impossible	Feasible	Subjective
T16	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	Feasible	Objective

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T17	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T18	Adequate	Unclear	Absent	Absent	Present	Impossible	Impossible	Feasible; Impossible	Subjective; Objective
T19	Adequate	Adequate	Absent	Absent	NR	Impossible	Impossible	impossible	Objective
T20	Unclear	Adequate	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T21	Adequate	Unclear	Absent	Absent	Absent	Feasible	Impossible	Feasible	Objective
T23	Inadequate	Inadequate	NR	Absent	NR	Impossible	Impossible	Feasible	Objective
T24	Adequate	Unclear	Absent	Absent	NR	Impossible	Impossible	Difficult; Impossible	Subjective; Objective
T26	Adequate	Unclear	NR	NR	Absent	Feasible	Feasible	Feasible	Objective
T27	Adequate	Adequate	Absent	Absent	NR	Impossible	Impossible	impossible	Subjective
T28	Inadequate	Unclear	Absent	Absent	Absent	Feasible	Impossible	impossible	Subjective
T29	Unclear	Unclear	NR	NR	No primary outcome	Impossible	Impossible	No primary outcome	n/a
T30	Adequate	Unclear	NR	Absent	NR	Impossible	Impossible	Feasible; Impossible	Objective
T31	Adequate	Unclear	NR	NR	No primary outcome	Feasible	Impossible	No primary outcome	n/a

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T32	Unclear	Adequate	NR	NR	Present	Impossible	Impossible	Feasible	Subjective
T33	Unclear	Unclear	NR	NR	Present	Feasible	Impossible	Feasible	Subjective
T34	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T35	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T36	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T37	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T38	Unclear	Unclear	NR	NR	Present	Impossible	Impossible	Feasible	Objective
T39	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T40	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T41	Unclear	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T42	Unclear	Unclear	NR	NR	Present	Impossible	Impossible	Feasible	Objective
T44	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T45	Unclear	Unclear	NR	NR	NR	Feasible	Impossible	impossible	Objective
T46	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T47	Unclear	Adequate	Absent	Absent	NR	Impossible	Impossible	Feasible	Objective
T48	Unclear	Inadequate	Absent	NR	NR	Impossible	Impossible	Feasible	Objective
T49	Adequate	Adequate	NR	NR	NR	Feasible	Impossible	Feasible	Objective

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T50	Adequate	Adequate	NR	Absent	NR	Difficult	Impossible	Feasible	Objective
T51	Unclear	Adequate	Absent	NR	Absent	Impossible	Impossible	impossible	Subjective
T52	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T53	Unclear	Unclear	Absent	Absent	Absent	Impossible	Impossible	Feasible	Subjective; Objective
T54	Unclear	Unclear	NR	Absent	NR	Feasible	Impossible	Feasible	Objective
T55	Adequate	Unclear	Absent	Absent	Present	Feasible	Impossible	Feasible; Impossible	Objective
T56	Adequate	Unclear	Absent	NR	NR	Impossible	Impossible	impossible	Subjective
T57	Adequate	Unclear	NR	Absent	Present	Impossible	Impossible	Feasible	Objective
T58	Unclear	Unclear	NR	Absent	NR	Feasible	Impossible	Feasible	Subjective
T59	Adequate	Unclear	NR	NR	No primary outcome	Feasible	Impossible	No primary outcome	n/a
T60	Adequate	Unclear	Absent	Absent	Present	Feasible	Impossible	Feasible	Objective
T61	Unclear	Adequate	Absent	Absent	NR	Feasible	Impossible	Feasible	Objective
T62	Unclear	Adequate	NR	Absent	Absent	Impossible	Difficult	Feasible	Objective
T63	Adequate	Adequate	NR	Absent	Present	Feasible	Feasible	Feasible	Objective
T64	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	impossible	Subjective

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T65	Adequate	Unclear	NR	NR	NR	Difficult	Impossible	Difficult	Subjective
T66	Unclear	Unclear	Absent	Absent	Present	Difficult	Impossible	Feasible	Subjective
T67	Unclear	Unclear	NR	NR	Present	Difficult	Impossible	Feasible	Objective
T68	Adequate	Adequate	Absent	NR	Absent	Difficult	Impossible	Difficult	Subjective
T69	Unclear	Unclear	Present	NR	Present	Feasible	Impossible	Feasible	Objective
T70	Adequate	Adequate	Absent	Absent	NR	Difficult	Impossible	impossible	Objective
T71	Adequate	Adequate	Absent	Absent	Absent	Feasible	Impossible	Feasible	Subjective
T72	Adequate	Unclear	Absent	Absent	Present	Feasible	Impossible	Feasible	Objective
T73	Unclear	Adequate	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T74	Unclear	Adequate	Present	Present	Present	Feasible	Feasible	Feasible	Subjective
T75	Unclear	Unclear	NR	NR	NR	Impossible	Impossible	feasible	Subjective
T76	Adequate	Unclear	NR	NR	NR	Feasible	Impossible	Difficult	Objective
T77	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T78	Unclear	Adequate	NR	NR	NR	Feasible	Impossible	Feasible	Objective
T79	Unclear	Unclear	NR	NR	No primary outcome	Impossible	Impossible	No primary outcome	n/a
T81	Unclear	Adequate	NR	NR	NR	Difficult	Impossible	Feasible	Objective

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T82	Adequate	Unclear	NR	NR	No primary outcome	Impossible	Impossible	No primary outcome	n/a
T83	Adequate	Unclear	Absent	Absent	Absent	Impossible	Impossible	impossible	Objective
T85	Adequate	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T86	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T87	Adequate	Unclear	NR	NR	NR	Feasible	Impossible	Feasible	Objective
T88	Unclear	Adequate	NR	NR	NR	Impossible	Impossible	impossible	Subjective
T89	Adequate	Unclear	NR	NR	NR	Difficult	Feasible	Feasible	Objective
T90	Unclear	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T91	Adequate	Adequate	Absent	Present	Present	Difficult	Difficult	Feasible	Subjective
T92	Adequate	Adequate	Absent	Absent	Present	Impossible	Impossible	Feasible	Objective
T93	Adequate	Unclear	NR	NR	NR	Feasible	Feasible	Feasible	Objective
T94	Adequate	Adequate	NR	NR	NR	Feasible	Impossible	Feasible	Objective
T95	Adequate	Unclear	NR	NR	Absent	Impossible	Impossible	Feasible	Objective
T96	Adequate	Unclear	NR	NR	Present	Difficult	Impossible	Feasible	Objective
T97	Unclear	Inadequate	Absent	Absent	Absent	Feasible	Impossible	Feasible	Objective

Trial Ref	Sequence Generation	Allocation Concealment	Blinding Conduct			Blinding Feasibility			Objectivity of outcome
			Patients	Clinicians	Assessors	Patients	Clinicians	Assessors	
T98	Adequate	Adequate	Present	Present	NR	Feasible	Feasible	Feasible	Objective
T99	Unclear	Adequate	NR	NR	Present	Difficult	Impossible	Feasible	Objective
T100	Unclear	Unclear	Present	Present	NR	Feasible	Impossible	Feasible	Objective
T101	Adequate	Unclear	NR	NR	Present	Feasible	Impossible	Feasible	Objective
T102	Adequate	Adequate	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T103	Adequate	Unclear	Present	NR	NR	Feasible	Impossible	impossible	Subjective
T104	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	Feasible	Objective
T105	Adequate	Unclear	NR	NR	NR	Difficult	Impossible	Feasible	Objective
T106	Unclear	Unclear	Absent	Absent	NR	Difficult	Feasible	Feasible	Objective
T107	Unclear	Adequate	Absent	Absent	Absent	Impossible	Impossible	impossible	Subjective
T108	Adequate	Unclear	NR	NR	NR	Impossible	Impossible	Unclear	Unclear

E.2 Sample sizes, attrition & intention−to−treat

The second table lists target sample sizes, actual sample size, presence of missing data and exclusions, analysis of patients according to randomised allocation and trial reports of intention–to–treat analysis

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T1	NR	128	Yes	Yes	Yes	No	Excluded	NR
T2	NR	214	Yes	NR	No	Apparent	none missing	NR
T3	800	810	Yes	Yes	No	No	Excluded	NR
T4	NR	255	Unclear	Yes	Yes	No	Mixed-Imputation partial records (method NR)	NR
T5	240	500	Yes	Yes	Yes	No	Excluded	Present
T6	500	500	Unclear	Yes	Yes	No	Included-Outcome assumed absent (method NR)	Present
T7	NR	305	Unclear	Yes	No	No	Unclear whether imputed	NR
T8	76	80	Yes	NR	No	Apparent	none missing	NR
T9	140	140	Yes	NR	No	Apparent	none missing	NR
T10	10	28	Yes	Yes	No	Yes	Included-method NR	NR
T11	744	752	Yes	Yes	Yes	No	Excluded	Present
T12	108	145	Yes	Yes	No	No	Excluded	Present
T13	NR	43	Yes	No	No	Yes	none missing	NR
T14	2220	2220	Yes	Yes	No	Yes	Included-outcome assumed absent (method NR)	NR
T15	700	629	Yes	Yes	No	Yes	Included-outcome assumed absent (method NR)	Present
T16	424	424	Yes	No	No	Yes	none missing	NR

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T17	160	104	Yes	NR	No	No	Excluded	Present
T18	1400	1388	Yes	Yes	No	Yes	Included-outcome assumed missing (method NR)	NR
T19	156	157	Yes	Yes	No	No	Excluded	Present
T20	NR	88	Yes	Yes	No	No	Excluded	Present
T21	100	107	No	NR	No	Apparent	none missing	Present
T23	1776	1883	Yes	Yes	No	No	Excluded	NR
T24	300	298	Yes	Yes	No	No	Excluded	Present
T26	NR	392	Yes	No	No	Yes	Excluded	Present
T27	154	154	Yes	Yes	Yes	No	Excluded	Present
T28	NR	101	Unclear	Yes	No	No	Excluded	Present
T29	NR	210	Unclear	Yes	No	No	Excluded	NR
T30	116	105	Yes	Yes	Yes	No	Excluded	Present
T31	NR	263	Yes	NR	No	Apparent	none missing (prot devs inc)	Present
T32	NR	2475	Yes	Yes	No	No	Excluded	Present
T33	1000	1025	No	Yes	Yes	No	Excluded	NR
T34	260	309	Unclear	Yes	NR	No	Excluded	NR
T35	NR	72	Yes	No	No	Yes	none missing	NR

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T36	800	782	Yes	Yes	No	No	Excluded	Present
T37	NR	66	No	Yes	Yes	No	Excluded	NR
T38	3000	1222	Yes	Yes	No	Yes	Included-outcome assumed absent (method NR)	Present
T39	250	263	No	Yes	Yes	No	Excluded	Absent
T40	300	199	Yes	Yes	No	No	Excluded	NR
T41	NR	348	Unclear	Yes	No	No	Unclear whether imputed	Present
T42	1000	201	Yes	No	No	Yes	none missing (prot devs included)	Present
T44	500	500	Yes	Yes	Yes	No	Excluded	Absent
T45	1276	1276	Yes	NR	No	Apparent	none missing	NR
T46	46	84	No	Yes	Yes	No	Excluded	NR
T47	100	118	Yes	No	No	Yes	none missing (prot devs included)	Present
T48	988	972	Yes	No	No	Yes	none missing	NR
T49	4000	4429	Unclear	Unclear	Yes	Unclear	Unclear whether imputed	NR
T50	100	110	Unclear	Unclear	Yes	Unclear	Unclear whether any missing	NR
T51	192	341	No	Yes	Yes	No	Excluded	NR
T52	174	184	Yes	Yes	Yes	No	Excluded	Present

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T53	200	103	Yes	NR	No	Apparent	none missing	NR
T54	NR	4993	Yes	No	No	Yes	none missing	NR
T55	444	452	Unclear	Yes	No	No	Excluded	Present
T56	200	201	Unclear	Yes	No	No	Excluded	Present
T57	1950	2094	Yes	Yes	No	No	Mixed: assumed poor outcome for lost to follow–up; Missing letters excluded	Present
T58	300	308	No	Yes	Yes	No	Excluded	Present
T59	NR	250	Yes	NR	No	Apparent	none missing	NR
T60	140	152	Yes	Yes	No	No	Excluded	Present
T61	600	601	Yes	Yes	No	No	Excluded	Present
T62	740	740	Unclear	Yes	No	No	Excluded	Present
T63	2000	456	Yes	Yes	No	No	Excluded	NR
T64	712	772	Yes	Yes	NR	No	Unclear whether any missing	NR
T65	210	234	Yes	Yes	NR	No	Excluded	Present
T66	136	145	Yes	Yes	No	No	Excluded	Present
T67	10000	5341	Yes	No	No	Yes	none missing (prot devs included)	Present

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T68	2500	2602	Yes	Yes	Yes	No	Excluded (ITT denominators not used)	Present
T69	182	177	Yes	Yes	No	Yes	Included-Last obs carried forward	Present
T70	357	357	No	Yes	Yes	No	Excluded	Present
T71	750	722	Unclear	Yes	No	No	Excluded	Present
T72	246	262	Yes	Yes	Yes	No	Excluded	Present
T73	750	344	Yes	No	No	Yes	none missing (prot devs included)	Present
T74	570	399	Yes	Yes	Yes	No	Excluded	Present
T75	561	710	Yes	No	No	Yes	none missing	Present
T76	NR	809	Yes	Yes	No	Yes	Included-method NR	NR
T77	1500	4128	Yes	Yes	Yes	No	Excluded	Present
T78	800	799	Yes	NR	No	Apparent	none missing	Present
T79	NR	200	Yes	NR	No	Apparent	none missing	NR
T81	1732	1932	Yes	No	No	Yes	none missing	NR
T82	NR	203	Unclear	Yes	No	No	Unclear whether imputed	Present
T83	NR	457	Yes	NR	No	Apparent	none missing (prot devs included)	Present

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T85	190	205	No	Yes	Yes	No	Excluded	Present
T86	150	151	Unclear	Unclear	Yes	Unclear	Unclear whether any missing	Present
T87	NR	20	Yes	NR	No	Apparent	none missing	NR
T88	90	89	Yes	Yes	Yes	No	Excluded	Present
T89	432	433	Yes	Yes	No	No	Excluded	NR
T90	NR	301	Yes	Yes	Yes	No	Excluded	NR
T91	1380	1417	Yes	Yes	No	No	Excluded	Present
T92	896	898	Yes	Yes	Yes	No	Excluded	Present
T93	566	3377	Yes	No	No	Yes	none missing (prot devs included)	NR
T94	70	100	Yes	Yes	No	No	Excluded	Present
T95	100	100	Yes	Yes	Yes	No	Excluded	Absent
T96	412	430	Yes	Yes	No	No	Excluded	Present
T97	400	431	Yes	No	No	Yes	none missing	Present
T98	372	408	Yes	No	No	Yes	none missing (prot devs included)	Present
T99	1200	1200	Yes	Yes	No	Yes	Included-censored at the time of the last contact.	Present

Trial Ref	Target sample size	Number Randomised	Analysed as Randomised	Missing data	Exclusions	Analysis Complete	How were missing data dealt with?	ITT Reported
T100	478	500	Yes	No	No	Yes	none missing	Present
T101	852	1005	Yes	Yes	No	Yes	Included-assumed no events (method NR)	Present
T102	40	41	Yes	No	No	Yes	none missing	Present
T103	385	252	Yes	No	No	Yes	none missing (prot devs inc)	Present
T104	NR	520	Yes	Yes	Yes	No	Excluded	NR
T105	NR	163	Yes	NR	No	Apparent	none missing	Present
T106	240	241	Yes	NR	No	Apparent	none missing	NR
T107	500	553	Unclear	Yes	No	No	Excluded	Present
T108	242	129	No	Yes	Yes	No	Excluded	Present