Health Technology Assessment

12

Heidi Probst and Aarthi Ramlaul

12.1 Introduction

HTA involves the assessment of all new technologies used in radiography and radiotherapy. Effectiveness refers to the extent to which benefits are brought to patients in routine circumstances, and efficiency refers to the extent to which acceptable effectiveness is achieved with the best use of resources.

In this chapter we will discuss the following research methods used to assess health technologies in medical imaging and radiotherapy. In order to give each of these areas their deserved attention, each is covered within a separate sub-chapter, as follows.

- Researching diagnostic tests
- Researching therapies using randomized controlled trials (RCTs)
- Health economic assessment and RCTs
- Systematic reviews and meta-analyses of RCTs

H. Probst (✉)
Radiotherapy and Oncology, College of Health, Wellbeing and Life Sciences, Sheffield Hallam University, Sheffield, UK
e-mail: h.probst@shu.ac.uk

A. Ramlaul
Diagnostic Radiography and Imaging, School of Health and Social Work, University of Hertfordshire, Hatfield, Hertfordshire, UK
e-mail: a.ramlaul@herts.ac.uk

A. Ramlaul (ed.), *Medical Imaging and Radiotherapy Research: Skills and Strategies*, https://doi.org/10.1007/978-3-030-37944-5_12

12.2 Researching Diagnostic Tests

Diagnostic testing can be seen as the collection of information which will clarify a patient's clinical condition and help to determine prognosis. This information can include patient characteristics, signs and symptoms, clinical history, physical examination or clinical tests. Practitioners working in diagnostic imaging are particularly interested in providing high quality images which will permit an accurate medical diagnosis. Diagnostic imaging is a rapidly evolving specialty, and numerous imaging procedures such as Barium enemas, angiography and intravenous pyelography, to name a few, are being replaced by computed tomography (CT) and magnetic resonance imaging (MRI). New technologies, however, are complex and expensive and research is therefore required to evaluate them, in order to decide if and when they should be introduced into clinical practice.

The purpose of this section of the sub-chapter is to provide an overview of what is meant by evaluation of diagnostic technologies, focusing on research that measures the diagnostic performance (or accuracy) of an imaging modality and provides estimates of observer variability.

12.2.1 Evaluative Hierarchy of Diagnostic Technologies

Choices between alternative healthcare policies may be explored within healthcare evaluation, including the investigation of the efficacy and efficiency of available diagnostic technologies. It is not always apparent how the diagnostic technology itself brings about improvements in the prognosis or physical health of a patient. An imaging examination provides information from which a reporting radiographer or radiologist makes a report. This is then used by the clinician in combination with clinical findings and other tests to make or refine diagnosis and plan treatment which ultimately might affect patient outcomes. Therefore, to evaluate the effectiveness of imaging requires the measurement of a chain of events between the application of the technology and any potential influence on disease. With the development of CT in the 1970s, Fineberg and colleagues [1] suggested that hierarchy could be used to evaluate the effectiveness of diagnostic technologies. This has subsequently been extended to include whether the costs for a given examination are acceptable, providing an efficient use of resources [2]. Figure 12.1 presents the evaluative hierarchy as applied to the assessment of MRI [3].

Technical performance is the first level of the evaluative hierarchy and is concerned with whether, for example, MRI produces good quality images from which diagnostic and therapeutic decisions may be made [4].

The next level is diagnostic performance, which is concerned with whether imaging, such as MRI of the knee, correctly or incorrectly assesses the presence or absence of disease, such as meniscal or ligamentous injury, as corroborated by a 'gold standard' test (such as arthroscopy in this instance). Assessment of diagnostic performance is expressed using statistics such as sensitivity and specificity. Sensitivity is the percentage of correct abnormal diagnoses in patients with disease;

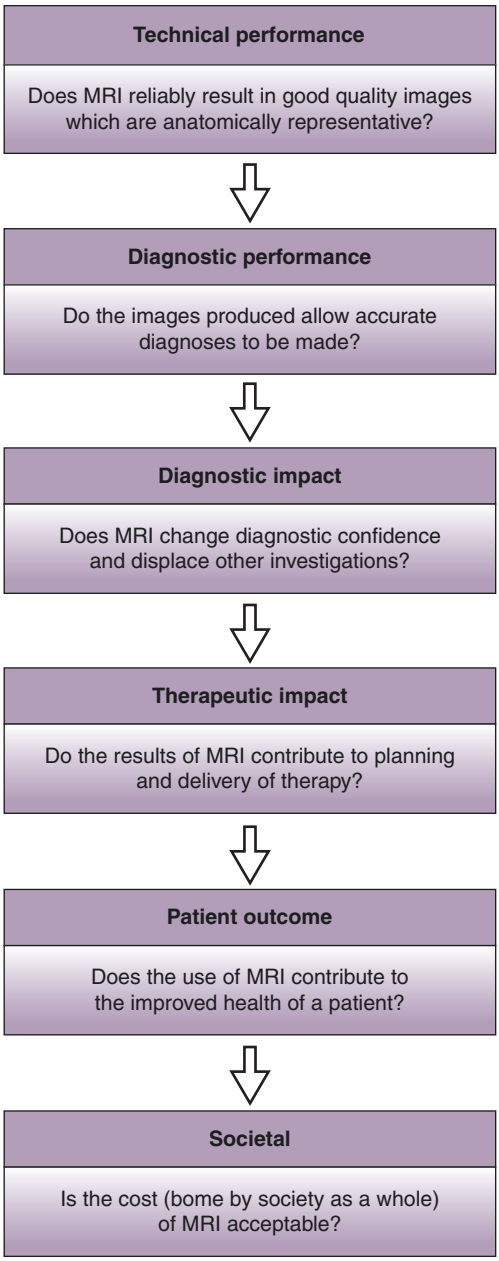

Fig. 12.1 The hierarchy used to evaluate MRI

and specificity is the percentage of correct normal diagnoses in patients without disease. Furthermore, observer variation in the interpretation of medical images is substantial and has been described as radiology's 'Achilles' heel' [5]. Thus, it is important to estimate observer variability, since the accuracy of the diagnostic test

can be a joint function of the images produced and the performance of the observers [2]. This level in the evaluation of a diagnostic technology is discussed further in the next section.

The following three levels of the evaluative hierarchy are concerned with:

- diagnostic impact, e.g., does MRI replace existing technologies?
- therapeutic impact, e.g., do MRI findings lead clinicians to make changes in treatments?
- patient outcome, e.g., does MRI improve patients' prognoses?

These levels of the hierarchy are often assessed using observational research designs. In these the technologies are simply observed and compared, without the experimental intervention that would take place in a randomized controlled trial. An example might be recording pre-imaging diagnosis and management plans and comparing this with post-imaging plans. Such studies assume that any change in diagnosis and management plan, or change in patient outcome, is attributable to MRI. The effectiveness of MRI, however, might be explained by the influence of other variables. One possibility is that there is a tendency for measured outcomes to 'average out' over time following the introduction of a new policy, due to random fluctuations in performance results, if enough results are taken. This is referred to statistically as 'regression towards the mean'. Another reason could be the Hawthorne or 'guinea pig' effect, which is the tendency for data to be biased because research subjects become aware they are being observed [6].

The best method for evaluating the effectiveness of technologies such as MRI is the randomized controlled trial (RCT), which will, in a controlled way, randomly allocate patients to receive either one diagnostic test or an alternative. Although there are logistical and financial implications to using RCTs, this method promotes study validity and provides a good basis for making statistical inferences [4]. The randomized controlled trial design is discussed later in the chapter.

The final level of the evaluative hierarchy moves beyond merely measuring the clinical effects of a technology to determining whether the cost of that technology is acceptable to society. For the policy maker entrusted with making resource allocations, it is necessary to assess the extent to which MRI is an efficient use of resources to provide benefits to society [2]. This could take the form of, for example, a cost-effectiveness study which involves computing a cost per unit of output for a medical technology such as cost per arthroscopy avoided by using MRI of the knee. The different methods of economic evaluation are discussed later in the chapter.

12.2.2 Studies of Diagnostic Test Accuracy

When diagnosing a patient, clinicians seldom have access to the gold standard or reference standard test for the disorders they suspect since these tests can be expensive, painful and/or invasive. There are many alternative tests that can be used for

patient diagnosis, such as taking a patient's history, physical examination, laboratory tests and diagnostic imaging. Diagnostic accuracy studies, which comprise the second level of the evaluative hierarchy, are vital to the assessment of imaging technologies, since they help to understand how they should be best used in clinical practice.

12.2.3 The Research Question

Sackett and Haynes [7, 8] identified four types of research questions that can be used to assess the real value of a diagnostic test such as an imaging modality.

- Do diagnostic test results in patients with the target disorder differ from those in normal people (a phase I question)?
- Are patients with certain diagnostic test results more likely to have the target disorder than patients with other test results (a phase II question)?
- Does the diagnostic test result distinguish patients with and without the target disorder among patients in whom it is clinically reasonable to suspect the disease is present (a phase III question)?
- Do patients who undergo this diagnostic test have better health outcomes than similar patients who are not tested?

Phase III questions are the most frequently asked in studies of diagnostic test performance and are concerned with the validity of the diagnostic test or rather whether it measures what it proposes to measure.

To evaluate whether a test can distinguish normal from abnormal patients during routine clinical practice requires the results of the test to be compared against the gold or reference standard that is acknowledged as being the best available test to accurately diagnose the patient's true disease status. To compare measurements, i.e., the diagnostic test and reference standard results, is to assess validity and this will be the main focus of this section. Studies of the diagnostic accuracy, or validity of a test, particularly for imaging modalities, should also consider whether the different observers responsible for interpreting medical images are doing this consistently; this provides an assessment of reliability. The design and analysis of reliability studies will also be briefly discussed.

12.2.4 Design of a Study of Validity

As described above, a diagnostic accuracy study involves the assessment of whether a diagnostic test can distinguish patients with and without the target disorder, as corroborated by gold or reference standard, among patients in whom it is clinically reason- able to suspect the presence of disease. If the study design is inadequate, there is experimental evidence that the performance of diagnostic tests might be exaggerated [9]. The STARD (Standards for the Reporting of Diagnostic accuracy

studies) statement, which is a checklist used to guide the reporting of studies of accuracy [9], and the QUADAS (Quality Assessment for Diagnostic Accuracy Studies), which is a generic tool used to appraise the quality of primary studies in systematic reviews of diagnostic accuracy [10], provide thorough descriptions of the relevant design issues when considering the validity of a diagnostic test. These design issues are also discussed in Chap. 10. In summary then, when designing a diagnostic accuracy study, it is important to consider the following areas as they pose an element of risk to the study validity [11].

- Patient selection—a consecutive series of patients suspected (but not known) to have the target disorder should be prospectively selected as a cohort of patients for inclusion in the study. There should be a clear description of the selection criteria and the setting, e.g., primary, secondary or tertiary care.
- Choice and application of the reference standard—the reference standard chosen should produce results close to the truth, or the performance of the diagnostic test will be poorly estimated.

  The reference standard should be applied within a clinically acceptable timeframe after the diagnostic test and preferably to the whole or at least a random sample of patients to avoid partial verification of patients. Nor should the index test form part of the reference standard.
- Measurement of results—a study should fully report indeterminate test results that occur due to factors such as technical faults or inferior image quality, and withdrawals that may occur due to patient death, move in residency or no longer wanting to cooperate. It is important to consider whether they are non-random exclusions and the effect on generalizability.
- Independence of interpretation—the reference standard should be interpreted blind, i.e., in total ignorance of the diagnostic test result and vice versa.

12.2.5 Analysis of a Study of Validity

Various measures can be used to assess how well a diagnostic test discriminates between patients with disease from those without disease. The diagnostic test will detect the presence of a disease, such as a lesion on a digital mammogram, and then be correctly classified as being present or absent by biopsy, as the reference standard. This 'binary' classification of results allows individuals to be classified either as true positives (TP) or true negatives (TN), which means that the test results are correct; or false positives (FP) and false negatives (FN), which means that the test results are incorrect (Fig. 12.2). Positive and negative refer to the presence or absence of the target disorder.

The number of individuals classified as TP, TN, FP and FN permits the calculation of sensitivity and specificity, predictive values and likelihood ratios to answer different questions as described below:

Sensitivity is the proportion of patients with disease who have a positive test result: i.e., how good is my diagnostic test in detecting patients with disease?

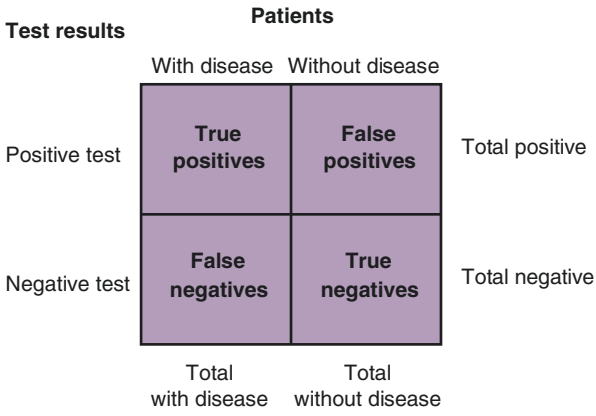

Fig. 12.2 Binary classification of results

$$\text{Sensitivity} \frac{1}{4} \text{TP} = \text{total with disease } \eth \text{TP} \text{ þ } \text{FN} \text{Þ}$$

Specificity is the proportion of patients without disease who have negative test results: i.e., how good is my diagnostic test in detecting patients without disease?

$$\text{Specificity} \frac{1}{4} \text{TN} = \text{total with disease } \eth \text{FP} \text{ þ } \text{TN} \text{Þ}$$

Positive predictive value is the proportion of patients with positive test results who have the disease: i.e., how well does a positive test result predict the presence of disease?

$$\text{Positive predictive value} \frac{1}{4} \text{TP} = \text{total positive } \eth \text{TP} \text{ þ } \text{FP} \text{Þ}$$

Negative predictive value is the proportion of patients with negative test results who do not have the disease: i.e., how well does a negative test result predict the absence of disease?

$$\text{Negative predictive value} \frac{1}{4} \text{TN} = \text{total negative } \eth \text{FN} \text{ þ } \text{TN} \text{Þ}$$

Positive likelihood ratio is the ratio of the true positive rate to the false positive rate: i.e., how much are the odds of the disease increased when a test is positive?

$$\text{LR} \text{ þ ve} \frac{1}{4} \text{sensitivity} = \eth 1 - \text{specificity} \text{Þ}$$

Negative likelihood ratio is the ratio of the false negative rate to the true negative rate: i.e., how much are the odds of the disease decreased when a test is negative?

LR-ve ¼ ð1 − sensitivityÞ=specificity Likelihood ratios can be applied to clinical practice to estimate the chances of disease in a patient according to their test result using Bayes' theorem [12]. In order to calculate the post-test odds of disease, you need to specify the pre-test odds: i.e., the likelihood that the patient would have a specific disease prior to testing. The pre-test odds are usually related to the prevalence of the disease, though you might adjust it depending on characteristics of the individual patient. Once you have specified the pre-test odds, you multiply them by the likelihood ratio. This gives you the post-test odds. Suppose a woman had a negative mammogram when screening for breast cancer and the local prevalence of

cancer among women is 5% and the negative likelihood ratio for a mammogram is 0.20. Using Bayes' theorem we can estimate that the woman's probability of breast cancer prior to screening will be reduced after a negative mammogram from 5% to 1%.

- pre-test odds ¼ prevalence/(1 − prevalence) ¼ 0.05/0.95 ¼ 0.05
- post-test odds ¼ pre-test odds ∗ LR-ve ¼ 0.05 ∗ 0.20 ¼ 0.01
- post-test probability ¼ post-test odds/(1 þ post-test odds) ¼ 0.01/1.01 ¼ 0.01 (or 1%)

Sometimes, however, the test under evaluation might yield results as a continuous measurement or ordered categories. The images from MRI of the knee, for example, might be used to describe some anatomical feature such as degenerative changes in the menisci as definitely, probably or possibly present, and probably or definitely absent, and then confirmed as present or absent by arthroscopy. Sensitivity and specificity could still be calculated by combining categories above and below a threshold, such as combining definitely, probably or possibly present compared to combining probably or definitely absent.

Changing the threshold will alter the estimates of sensitivity and specificity. A more useful method, however, of measuring the performance of MRI across a range of thresholds, or 'cut-offs', is the receiver operating characteristic (ROC) curve (see also Chap. 8). The ROC curve, as shown in Fig. 12.3, shows

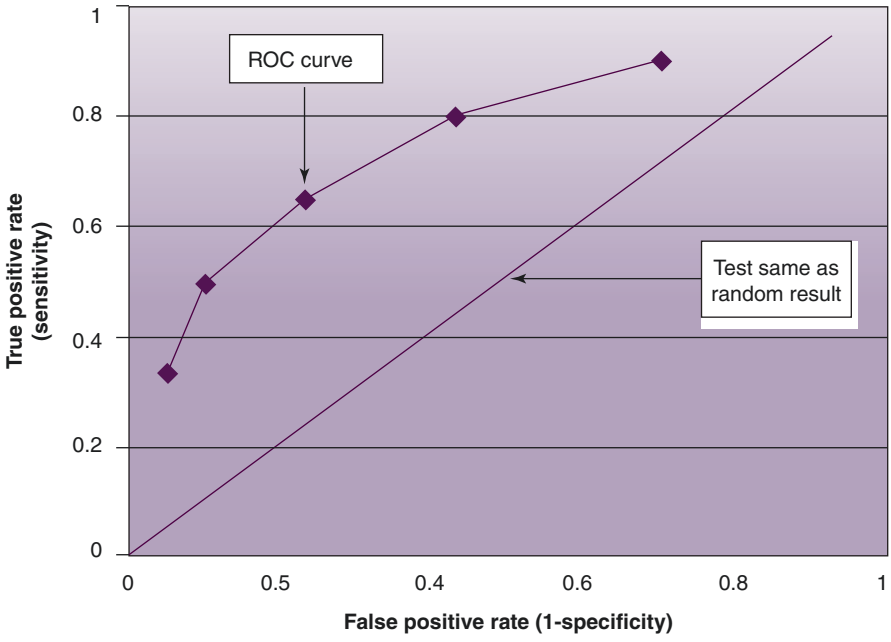

Fig. 12.3 Example ROC curve for an imaging procedure with ordinal categories

graphically the trade-offs at each cutoff for any diagnostic test that uses an ordinal or continuous variable. Ideally, the best cutoff value provides both the highest sensitivity and the highest specificity. This can be located on the ROC curve by finding the highest point on the vertical axis and the furthest to the left on the horizontal axis [13]. Alternatively, depending on the target disorder, it might be more important to exclude disease, so a higher sensitivity is chosen at the cost of lower specificity. Furthermore, it is possible to calculate the area under the ROC curve. When this is 0.5 (i.e., 50% sensitive and 50% specific) it represents a totally uninformative test, as shown in Fig. 12.3, by a straight diagonal line extending from the lower left corner to the upper right. A test that perfectly separates diseased from non-diseased patients would have an area under the curve of 1.0 (i.e., 100% sensitive and 100% specific). If the area under the curve of MRI of the menisci of the knee is 0.85, then the interpretation of the value is as follows. If two patients are drawn randomly from a sample of patients, in whom degeneration of menisci is present and absent, respectively, and, are both subjected to MRI to determine which patient had degeneration of the menisci, then MRI will be correct 85% of the time [14].

12.2.6 Assessment of Reliability

Diagnostic performance studies of imaging modalities require observers to interpret images and it is observer variability in this task that is considered to be the weakest aspect of clinical imaging [5]. It is important to estimate the variability of observers' performance, or the reproducibility with which an observer interprets an image, as this will influence the decisions made by clinicians and could ultimately affect patient outcome. The assessment of reliability involves different observers interpreting the same sample of images, known as an inter-observer test, or the same observers interpreting the same images on separate occasions, known as an intra-observer test [15]. We shall restrict our discussion of reliability to inter-observer variability as the principles of study design and analyses also apply to an assessment of intra-observer variability as well. In addition, inter-observer variability demonstrates observer consistencies within and between both sets of observers in the interpretation of images. As with studies of validity, similar principles apply to the design of a reliability study such as the need for a representative sample of patients and blinding in the interpretation of images. Selection bias is less likely when a consecutive or random sample of images is included, and blinding avoids the knowledge of one observer's interpretation influencing the interpretation of another observer. Availability of clinical data to observers should also be considered. It is important in a reliability study to carefully choose which observers are involved in the interpretation of images. For example, a study that includes highly specialist observers is likely to produce less generalizable results but in contrast could help to produce the best estimates of observer variability. Characteristics of observers that have been considered important in the assessment of reliability include the number of observers and their areas of training and expertise.

In studies of inter-observer variability, it is not assumed that one particular observer produces the correct report, but rather there is a genuine difference in interpretation of images between observers. The measure of performance used to analyse whether observers' reports agree is called the Kappa statistic [16]. It can be calculated when the classification of an image by an observer is binary, e.g., the presence or absence of a fracture on a plain radiograph, or ordinal, e.g., a normal mammogram, one which shows benign disease, the suspicion of cancer or the presence of cancer.

Kappa is defined as $K = (P_o - P_e)/(1 - P_e)$, where P_o is the observed proportion of agreement, and P_e is the proportion expected by chance. Kappa has a maximum of 1.0 when there is perfect agreement between observers and a value of zero indicates no better than chance. Kappa can be calculated for agreement between:

- a single observer interpreting the same image on two separate occasions
- two different observers on the same occasion
- comparisons of multiple observers [5].

When considering Kappa for ordinal categories, it might be preferable to use weighted Kappa which gives different weighting to disagreements in accordance with the extent of the discrepancy.

12.3 Researching Therapies Using Randomized Controlled Trials (RCTs)

12.3.1 Introduction

Diagnostic imaging and radiotherapy practitioners will be aware of the pace of technological change. However, the introduction of a new technology should be accompanied by a careful assessment of its value over existing methods. Meticulous assessment of any new technology should involve a controlled analysis of the new technology compared with the current approach [17]. The aim of this section is to provide an overview of randomized controlled trials (RCTs) and how they can be used within radiation therapy and imaging. By the end of this section practitioners should understand how to apply RCT designs for their own investigations as well as appraise RCTs published within the literature for applying evidence in practice. This section will start with a brief review of the benefits of RCTs and why they are considered a powerful research tool within HTA. Following this the specific characteristics and types of RCTs will be presented with examples of how the design characteristics could be used to investigate topics of relevance to clinical practitioners and those working in healthcare education.

The quality of a RCT, i.e., how stringent the design of the study is in limiting opportunities for bias, can influence potential outcomes by either overestimating or underestimating the benefit of the intervention. Such distortions have the potential to lead to ineffective treatments or interventions being employed and

effective treatments being discarded [18, 19]. Quality can be affected at many different stages of design and implementation and so throughout the following section attention will be paid to limitations of RCTs and the factors that may affect internal validity.

The final part of this section will focus on the use of economic evaluations alongside RCTs as part of HTA utilizing a case study from a radiotherapy trial as an example of how this can be of value.

12.3.1.1 Benefits of Randomized Controlled Trials

RCTs are a research design under the positivist research paradigm. For example, there is an emphasis on neutrality with an attempt to keep researcher and research participant's remote from each other to avoid any influence on the study results. Characteristically RCTs seek to explain the whole by a study of one aspect or parts. RCTs are based on a science model in which there is a belief in universal laws measuring and analysing relationships using numbers to quantify effects or behaviour. Objectivity is a primary aim and specific aspects of the approach are designed to provide neutrality and to avoid personal biases. Control over potential biases or confounding variables is integral to this approach. Owing to the strict controls and statistical strengths of RCTs, this design sits high within the hierarchy of evidence. Table 12.1 shows the Scottish Intercollegiate Guidelines Network (SIGN) hierarchy of evidence. It is clearly demonstrated in Table 12.1 that studies where there is a high risk of bias are given a lower ranking than similar designs where bias is deemed low [20].

Why Are RCTs So Useful?

Consider the following scenario.

Post-operative radiotherapy for breast cancer is the accepted treatment for the majority of women following surgery. Radiation treatment to the breast can lead to a mild skin reaction (erythema), reactions usually start in the second week of treatment and increase as the treatment course progresses. Traditionally skin care advice

Table 12.1 Levels of evidence from the Scottish Intercollegiate Guidelines Network

1++	High quality meta-analyses, systematic reviews of RCTs or RCTs with a very low risk of bias
1+	Well conducted meta-analyses, systematic reviews or RCTs with a low risk of bias
1-	Meta-analyses, systematic reviews or RCTs with a high risk of bias
2++	High quality systematic reviews of case control or cohort studies High quality case control or cohort studies with a very low risk of confounding or bias and a high probability that the relationship is causal
2+	Well conducted case control or cohort studies with a low risk of confounding or bias and a moderate probability that the relationship is causal
2-	Case control or cohort studies with a high risk of confounding or bias and a significant risk that the relationship is not causal
3	Non-analytic studies, e.g., case reports, case series
4	Expert opinion

has been to undertake a variety of practices with limited evidence base to support the advice given which may include the below three.

1. Washing with mild soap
2. Using mild creams in the treated area
3. Allowing air to get to the skin

Recently, barrier dressings have been developed to try to reduce the impact of friction from clothing exacerbating skin reactions and to reduce radiation induced erythema [21].

If investigators wanted to study the benefits of using a barrier film on radiotherapy patients to identify its usefulness in preventing erythema they might formulate a research question as follows.

> Does the use of a barrier film on the irradiated skin during breast cancer irradiation reduce the skin reactions experienced by patients?

Here it would be useful to take a few minutes to consider some of the different research approaches that can be used to answer this question. For the research question above, what would be the strengths and limitations of the research methods posed in Box 12.1

Box 12.1 Different Approaches That Can Be Used for Research on the Use of a Barrier Film on Irradiated Skin

Method 1

A prospective evaluation of skin reactions on all patients irradiated for breast cancer using a barrier film on the affected skin during treatment.

Method 2

A prospective evaluation of skin reactions on all patients irradiated for breast cancer using a barrier film on the affected skin during treatment compared with the results of a previous study to evaluate skin reactions in irradiated patients with breast cancer using conventional skin care instructions.

Method 3

A prospective evaluation of skin reactions on all patients irradiated for breast cancer using a barrier film on the affected skin during treatment, compared with a control group of irradiated patients who are given the conventional skin care instructions. Patients can opt for either the current skin care approach or the barrier film intervention.

Method 4

A prospective evaluation of skin reactions on all patients irradiated for breast cancer using a barrier film on one half of the affected skin during treatment, the other half of the breast or chest wall, patients use the conventional skin care instructions (no barrier film).

As method 1 has no comparison group it is not possible to place the results in any context, so we would still be unsure which skin care regimen was most effective.

In method 2 a comparison group is available to provide some way of assessing the performance of the new intervention. However, using a historical control group as the comparator has a number of problems and would mean any results obtained could be viewed as unsound. For example, if we assume the results identified a statistically significant reduction in erythema in patients using the barrier film, it is possible that this result may have occurred not because of the intervention but due to other extraneous factors including the following.

1. A difference in patient characteristics between the two study groups. If the historical control group contained a higher proportion of patients with larger breasts than in the barrier film group it is possible that this might account for the difference in skin reactions seen as it is known that breast size has an influence on subsequent adverse events [22].
2. Technological differences between the two periods of study. As time passes changes in technology may mean application of treatment is no longer the same. The introduction of a new planning technique or a change to the immobilization device between the two data collection periods could account for differences in skin reactions observed.

In method 3 the use of a comparison group treated in parallel with the intervention group eliminates potential confounding variables associated with a historical control group. However, patients choosing between treatments could mean that patient numbers might be unbalanced between the two skin care interventions and it is likely that patient characteristics would be unbalanced between the two study arms. Furthermore, where there is the option for choice it is possible that any patient reports of symptoms may be underplayed, especially where patients have read favourable information about a specific intervention, e.g., the benefits of using a barrier film, again limiting any confidence the researchers can have in the results obtained.

In method 4 the researchers would need to take care to make sure the radiation dose received to the skin underneath the barrier film was the same as that received by skin in the section of the breast not covered by the barrier film. They would also need to take care that the skin care used on the breast tissue not covered by the barrier film did not itself cause an increased irritation. For example, some topical creams can cause dryness or irritation that may exacerbate any radiation skin reaction.

Using the above scenario it is possible to see the need for strict control of possible confounding variables as well as the benefits of blinding participants to the intervention, and the use of methods to ensure a balance of patient characteristics between the intervention and control arms. RCTs allow rigorous evaluation of a single variable in a defined patient group. Within the RCT design it is possible to eradicate potential bias by comparing two or more groups with balance in patient characteristics. Where RCTs are used this also allows the opportunity for

meta-analysis comparing studies of the same investigation across different populations or geographical areas to provide a larger overall sample size and a potentially powerful analysis (see later in this section). In the next section the specific design characteristics of RCTs will be presented and some of the terminology associated with RCT design will be explained so practitioners can evaluate different RCTs presented in the literature.

12.3.1.2 Design Characteristics of RCTs

As the name indicates, RCTs involve random allocation of participants to treatment or control groups. Both groups are generally followed for a specific period and measurements taken at the same time points for both groups. The groups are analysed in terms of an outcome that is defined at the outset. For example, in the previous scenario a patient's skin reactions may be measured using a standard skin toxicity score such as the Radiation Induced Skin Reaction Assessment Scale (RISRAS) [23, 24] or the Radiation Therapy Oncology Group (RTOG) [25] scoring system at specific points throughout the treatment course. A pre-treatment (baseline) assessment of skin colouration should be undertaken to ensure patients do not have erythema, perhaps associated with sun exposure, prior to the start of radiotherapy that would alter any post-treatment results. This baseline measure would also be used to ensure parity between the two groups at the outset. Measurements may be taken weekly during the course of radiotherapy and also at 2 weeks post irradiation when skin reactions may be at their peak. The timing of outcome measurements is crucial to the accuracy of the study and thought needs to be given to this aspect of the study design.

Within the RCT design controlling bias is a main focus so researchers need to consider any potential confounding variables that may influence the outcome and control for these within the analysis. For example, using the skin study scenario we have already identified that patient size can influence the skin reactions experienced so it would be important to record patient size, either chest separation or breast volume, at the outset and test the two treatment arms for equality of this characteristic. Researchers would need to consider all possible confounding variables so other factors may include the level of homogeneity of the dose distribution [26] within the planning target volume (PTV). In the next few sections we will consider in a little more detail some of the specific design characteristics of RCTs.

Types of RCTs

RCTs are often defined by:

- the purpose of the study, i.e., explanatory, efficacy or pragmatic trials
- how participants are exposed to the intervention, i.e., parallel, cross-over or factorial designs
- number of participants
- how the intervention is assessed [27].

When assessing health technology, RCTs are usually pragmatic trials where the study is designed to reflect normal clinical activities. The aim of a pragmatic trial is to determine if the intervention works but also to describe any consequences of implementation of the technology. Pragmatic trials often have wider inclusion criteria to ensure the sample studied represents the normal group of patients that are likely to be seen in everyday practice. The comparison group in a pragmatic trial is often the current treatment or current imaging technique. Effectiveness trials aim to assess whether an intervention works in people who are offered the intervention. They tend to be pragmatic studies as the aim is to assess the effects under normal daily practice. They have simpler designs with less strict inclusion criteria than efficacy studies allowing participants to accept or reject the intervention offered. An example of an effectiveness study would be the early evaluation of breast cancer screening where RCTs were used to identify the impact of a screening intervention. Patients would be called for screening but may opt not to attend. Follow-up of this arm would include all patients offered screening irrespective of whether they attended the screen or not and compared with patients in a control group (who were not offered any intervention) [28, 29]. An efficacy study is where the aim is to identify if an intervention works in those that receive it. Figure 12.4. shows a pictorial presentation of two basic RCT designs [27].

In its simplest form an RCT has two arms, an intervention arm, that may be a new process or technology being tested, compared with either a control arm, that receives no intervention, or a second intervention arm, which in HTA is usually the current treatment or current imaging modality. Cross-over designs can be a powerful way to study the impact of a new technology (see design (**b**) in Fig. 12.4). Here patients or subjects are used as their own control and this avoids the need for matching

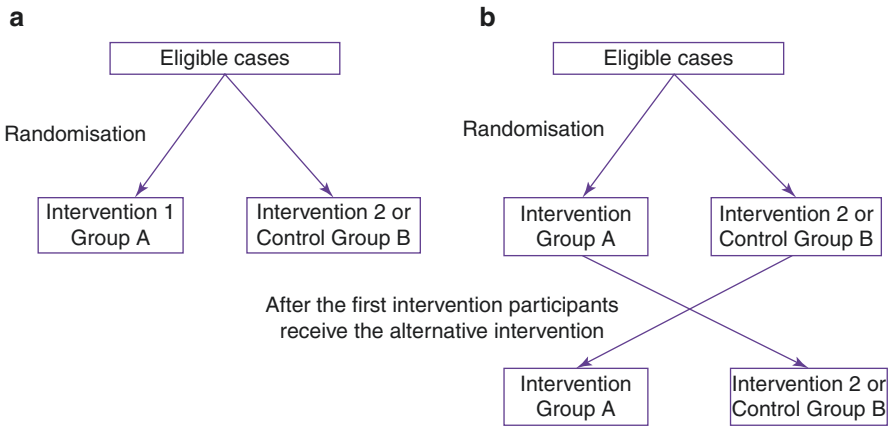

Fig. 12.4 Diagrammatic presentation of simple randomization and cross-over randomized controlled trial designs. (**a**) Simple randomization, (**b**) Cross-over or within-subjects design where participants receive both interventions in a different order (or the intervention and the control depending on the design)

characteristics across groups with different subjects that occur with the simple parallel design. However, cross-over or repeated measures designs can only be used in HTA where the first intervention has no lasting effect on the primary outcome measure. So, in the scenario used above it would not be appropriate to use a barrier film for the first two weeks of treatment and then apply traditional skin care for the remainder of the treatment course as the effect of the first skin care regimen would impact on subsequent skin reactions measured during application of the second regimen. In educational studies this design could not be used where subjects would learn through the first phase of the study. For example, if you wanted to test the effectiveness of two formats of patient information on a patient's ability to perform a breath hold technique during radiotherapy planning, you could use either of the below.

- A patient education video versus
- A traditional written information leaflet

The participant's ability to hold their breath following exposure to the video would influence subsequent performance so it would be difficult to distinguish if the video or the pamphlet had the impact on overall ability to perform the technique.

However, this cross-over design can be used successfully when used with consideration for potential learning effects. For example, it has been used to assess the impact of work speed on the accuracy of setting up a patient for a complex technique using a phantom [30]. In this study each pair of staff was asked to set up the phantom as they would for a normal treatment, twice, under two different conditions. In condition 1, participants were given a scenario whereby they could take as much time as they needed and a radiographic image was taken of the final setup position to assess positional accuracy. In condition 2, participants were given the same technique to apply but the scenario was that they were treating a child that was distressed and so it was important to work fast but accurately, in order to assess the impact that a time pressure might have on treatment accuracy. It was important that groups alternated in the order in which they undertook the test, i.e., condition 1, then condition 2, for one group and condition 2, then condition 1 for another group, to ensure if there were any learning effects these would not affect the overall results.

Factorial trials offer the opportunity to test individual interventions as well as studying the impact of two or more interventions applied together. In imaging the factorial design has been used in experimental conditions to test the factors that influence image quality and radiation dose [31].

In addition, trials can be described as being single-blind, double-blind or triple-blind. Blinding refers to either the participants being blind to the intervention, i.e., they are unaware of which intervention they have been allocated to, or the investigators, such as the statistician. The purpose of blinding is to minimize opportunities for bias as a direct result of knowledge of the intervention received either by the participants or the investigators applying the treatment or collecting the data.

For example, where the participants have knowledge of the intervention they are to receive, there is a possibility any patient self-reports may be influenced by this knowledge. Where possible patients should be blind to the intervention; this is not always possible as it may often be obvious which of the interventions participants

have received. For example, in the skin care example above patients that are randomized to receive the barrier film will know that is the group they have been allocated, it is impossible to blind the patients in this scenario. Further opportunities for bias can occur during the assessment of the study outcomes. Researcher knowledge of the intervention arm can influence interpretation of key outcome measures, especially where the researcher has a hypothesis to test. For example, in a study to evaluate the effectiveness of using tattoos to improve radiotherapy treatment accuracy during breast irradiation compared with gentian violet pen for marking the skin, it was necessary to blind the researcher undertaking the analysis to the intervention each subject was allocated to during the measurement of the treatment images that were used to establish treatment accuracy [32]. Knowledge of the intervention could have resulted in the favourable measurement of some images in order to prove the hypothesis being tested. To completely reduce the opportunity for any bias researchers, where possible, should aim to blind the patient, the researchers undertaking measurement of the outcomes, and the researchers undertaking the statistical analysis (i.e., single-, double- or triple-blinding) [27]. Treatment effects may be overestimated by approximately 17% where double-blinding is not employed as compared with studies where double-blinding is used [18]. The impact of not blinding patients has been shown to have a significant impact on patient reported outcomes, with non-blinded patients giving more optimistic reports of the intervention (exaggerated in the region of 0.56 SDs) [19]. Where studies are reporting true intervention effect sizes this could mean exaggeration of effect by over 100% [19].

Randomization

The rationale for using randomization is to prevent bias occurring as a result of inequalities between the treatment options or intervention arms. For example, when looking at the effectiveness of breast cancer screening, it would be important for researchers to ensure equity of characteristics between the screening group and the control arm, such as age at time of entry into the study, as incidence of breast cancer is known to increase with age [33].

There are a number of methods available to researchers for achieving random allocation. The simplest way is by tossing a coin, throwing a dice or use of a table of random numbers. For example, it can be agreed at the start of the study that heads on a coin will indicate treatment arm A and tails treatment arm B or the control group. However, simple randomization methods such as this may still result in unequal numbers or unbalanced characteristics between the groups [27], especially in small trials [34]. To overcome this one method is the use of block randomization. Generally, blocks of four are used for a simple RCT design with two intervention arms, A and B, as follows: AABB, ABBA, BBAA, BABA, BAAB, and ABAB. One of the six possible combinations is selected and participants allocated to an intervention arm based on the sequence of four; the process is repeated as required depending on the sample size.

Even with block randomization some inequalities may still arise simply by chance, hence researchers need to be aware of this possibility and test baseline characteristics between the groups for equality. Where differences occur, it may be necessary to control for imbalances in subsequent analyses of the outcome data.

Alternatively, stratifying randomization by an important characteristic may reduce the potential for inequality. When considering the study to look at the impact of a barrier dressing to reduce skin reactions during breast cancer radiotherapy discussed above, it may help to stratify on patient size, i.e., large or small patients, as this is a contributing factor for skin reactions during breast irradiation. In this case a separate list of block sequences would be produced for each stratum, although as you increase the number of strata the risk of errors in application also increases [34]. A further alternative is the use of a technique called minimization. This method is successful at obtaining equality between groups for a set of relevant characteristics even in trials with small samples [34]. Here for the characteristics that require balance, e.g., age, patient size, menopausal status, etc., a running total of how many participants have been allocated with each characteristic to each intervention arm is kept. Following random allocation of the first participant, subsequent participant randomizations are weighted to the intervention arm that would maximize balance, i.e., minimize inequalities, with totals for each arm updated after each participant is entered into the study.

A further option for researchers is the use of cluster randomization. In contrast to most randomized trials where the individual is randomized, with cluster randomization groups of participants are randomized [27, 35]; clusters can be either general practitioner practices or imaging/oncology departments. The benefit of cluster randomization is a possible reduction in contamination of the control arm. For example, if you wanted to investigate the impact of a new electronic information service for patients, it is possible that those in the experimental arm might pass on to patients in the control arm, simply by chatting while in the waiting room, useful information they have gleaned as a result of the intervention. Cluster randomization may not be necessary for the majority of trial designs and therefore individual randomization should be used where possible to avoid some of the limitations of cluster randomization (see Box 12.2 for details [35]).

Box 12.2 Limitations of Cluster Randomization

1. Selection bias—different types of participants maybe recruited into different arms of the study due to the geographical locations of the clusters which may result in differences, for example, in socio-economic status between arms.
2. Selection bias—in cluster trials participants are not asked to consent to the study but to consent to being included in the study analysis; if a substantial proportion of the cluster participants refuse, then an imbalance will occur between the trial arms.

Cluster trials need larger sample sizes than trials that use individual randomization to ensure sufficient statistical power. If there is not full uptake of the intervention within the cluster, then a dilution effect may further influence the power of the study.

Concealment of Randomization

Randomization is generally accepted as the best way of removing opportunities for selection bias by removing any predictability in the assignment process. Yet the process of randomization itself can be fraught with opportunities for bias that may invalidate or reduce the quality of the subsequent results. A common approach adopted by novice researchers to the issue of randomization is to alternate participants to interventions as they are referred to the clinic or department, as they consider referral to be in itself, a random process (Table 12.2).

Looking at the process in Table 12.2, can you foresee any problems with this approach? Primarily there is an identifiable pattern that may introduce bias. For example, where the pattern is known, there is the opportunity for researchers to selectively change the detail of the information given to potential participants. This is done to discourage entry into the trial where that patient has co-morbid disease or any potential characteristic that the researcher considers may influence or skew the results in an unfavourable direction. Inadequate concealment of this nature can result in overestimation of the potential effect of the intervention in the order of 40% when compared with trials with adequate concealment of randomization [18].

One method used to reduce the opportunity for bias during randomization is to use sealed opaque envelopes containing random allocations. However, this system may be prone to interference. Clinicians can open envelopes in advance, or view allocations by holding the envelope up to a bright light. Block randomization of four is used, if three of the previous participant allocations are known, the fourth can be predicted allowing the clinician to reserve entering patients into a trial until specific participants present with desired characteristics. Subversion of allocation concealment has also been shown in one study to have a significant impact on the age of patients enrolled in the experimental arm compared with the control arm [36]; the median age in the experimental arm was 59 years compared with a median age of 63 years in the control arm when a sealed envelope system was used. For lone researchers undertaking a simple RCT as part of perhaps an undergraduate or postgraduate course of study the use of sealed opaque envelopes may be the only practical solution on offer; in these circumstances researchers should be aware of the potential for interference and subsequent effects on the study quality. In most cases attempts should be made to use a system that removes the randomization process from the researchers, such as a central randomization service available through local trials units [37].

Table 12.2 One method sometimes used by students or novice researchers to randomize participants

Participant number	Allocation
1	Intervention A
2	Intervention B
3	Intervention A
4	Intervention B
5	Intervention A

Sample Size Requirements

As well as randomization of patients into the control or intervention arms, RCTs rely on statistical analysis of the primary outcome to demonstrate effectiveness of the intervention. In order to demonstrate a statistically significant difference in treatments between the study groups it is important that an adequate sample is studied to demonstrate an effect. In HTA, improvements in outcomes may be small and therefore where studies have small sample sizes it may not be possible to demonstrate a difference even where a difference exists [38]. For this reason, researchers undertaking RCTs must consider at the outset what improvement in the primary outcome would be appropriate for a clinically significant improvement or benefit and then calculate the sample size required to establish this statistically. This calculation is referred to as a power calculation. For example, in a study to establish the effectiveness of a radiotherapy protocol to reduce lung morbidity for patients undergoing breast or chest wall irradiation following surgery for breast cancer, it was calculated that a sample of 200 patients in each group would be required to detect a difference of 0.3 (in the primary outcome measure) with 5% significance and 80% power [39] (see Chap. 10 for more information on power calculations).

Recruitment of Subjects

Recruitment of patients into clinical trials is often problematic. In the study of the effectiveness of a radiotherapy protocol to reduce patient reports of lung morbidity mentioned above [39], recruitment of subjects to the study was slow despite a feasibility study indicating sufficient eligible patients were available in the host centre. Recruitment was hampered by:

- clinicians forgetting to mention the study to eligible patients,
- patients refusing to participate partly due to poor information about the possible side effects of treatment at the early referral stage. Within the patient information sheet for the study details of lung morbidity were highlighted, and patients unknowing of this aspect of their treatment feared that inclusion in the study would cause unwanted respiratory side effects, even though this was a possible corollary of treatment regardless of inclusion in the study,
- limited patient awareness of clinical trials during the early stages of the study,
- a strong preference for one of the intervention arms with patients not wishing to take a chance of receiving the alternative option through randomization.

Of 452 patients assessed as eligible for inclusion in the study, 92 (20%) refused to participate [39], which is similar to reports from other cancer trials [40]. As well as an effect on the overall sample size, this loss of potential participants can have an effect on the generalizability of the results as the sample recruited may not fully represent the population of patients as intended. Where studies include a placebo arm it is possible that a reduction in acceptance to randomization may also occur [41]. Generally factors reported as influential in a patient's decision to join a study include the belief that they may help future patients, or that they may

benefit from inclusion [41]; hence researchers should ensure potential participants are aware of the benefits of the study during the recruitment stage. In many cancer trials a lack of participants can be reflective of strict inclusion criteria excluding a substantial proportion of patients, perhaps in the region of 30% [40, 42]. Hence more pragmatic trials with less strict inclusion criteria may enhance the proportion of patients eligible for study and thus increase the potential for recruitment and generalizability [40, 43].

A comparison study of two community-based RCTs undertaking similar palliative care interventions identified a number of positive recruitment strategies. The more successful of the two trials, studied in terms of reaching an adequate sample size, employed the following strategies to maximize recruitment [43]:

- use of an inflated sample size to account for expected high attrition from early withdrawal or death
- maximal inclusion criteria and minimal exclusion criteria
- dedicated recruitment nurse
- triage process to screen for eligible patients
- recruitment interview included key messages
- patients approached for consent before GP consent was requested
- extensive marketing to raise the profile of the study topic
- effort was placed on ensuring clinician input to the study to encourage feelings of inclusion and reduce concerns
- realistic timeframe to recruit sufficient sample size
- adequate funding to support an extensive recruitment strategy.

Other strategies that have been shown to have a beneficial effect is telephone reminders to non-responders [44]. Recruitment may be hampered where potential participants or referring clinicians have strong preferences for one of the intervention arms that leads to a refusal to be randomized. Again, where these patients refuse consent to randomization, a reduction in generalizability of the results may be a consequence. Furthermore, where patients with a strong preference accept being randomized, subsequent results may be biased by strong beliefs about the treatment received where blinding of the patient is not possible [45]. A solution to this dilemma is the use of patient preference trials and there are a number of different designs currently being used (Fig. 12.5) [27, 45, 46].

While patient preference designs may allow a greater proportion of patients to be included in a study, the disadvantage of such designs is the resultant unknown or uncontrolled confounding variables in the preference arms [45]. It is suggested that the analysis for these studies includes comparison of the two randomized arms alone and perhaps an analysis using randomization status as a co-variate [45]. A concern of using the Zelen design, where participants are randomized before giving consent, and where those randomized to the standard treatment only consent to treatment and not to participation in a study, is a possible ethical implication in therapeutic scenarios [46]. However, it has been suggested that this design is specifically helpful for population-based screening studies [46].

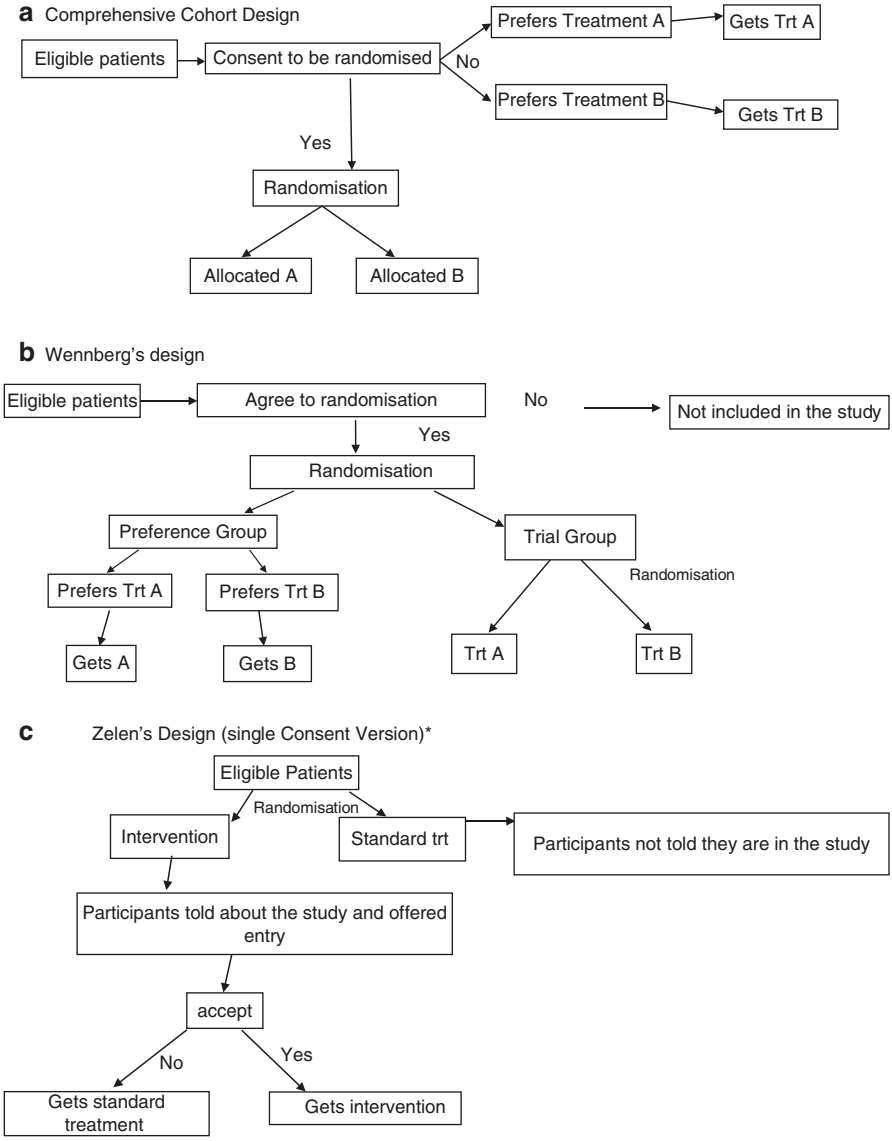

Fig. 12.5 Patient preference study designs [37, 46]. (**a**) Comprehensive cohort design. (**b**) Wennberg's design. (**c**) Zelen's design (single Consent Version)∗. ∗ In the double consent version participants are told which intervention they have been randomized to and offered the opportunity to switch to the alternative treatment [46]

Attrition

Even when researchers manage to recruit sufficient numbers to their trials problems with attrition can lead to a reduction in the strength of the reported findings. It is common for participants to fail to completely finish the allocated treatment or

intervention for a number of reasons: the patient may move to a different geographical area, the intervention may cause adverse side effects and the participant opts to withdraw leaving an incomplete data set. In addition, there may be missing data as a result of incomplete collection, perhaps due to staff absence at the time of collection, participants not adhering to the protocol or staff failing to record the information on the correct forms.

Attrition through a loss of patients to follow-up or incomplete data sets can bias results when the characteristics of those with missing data differ between the randomized groups [47]. It is therefore advised that missing data be presented by researchers in publications by providing normal baseline characteristics data for the whole study sample, and also separate data on those lost to follow-up from those remaining in the analysis so that readers can judge any imbalances between the intervention arms as a result of the missing data [47]. Using strategies to minimize attrition is beneficial and these may include minimizing patient burden by attention to the data collection methods [43]. For example, reducing the need for patients to attend clinics by visiting them at home may increase cooperation and reduce missing data; although more costly than other approaches this method was successful in the comparison of two community-based palliative care trials [43]. Ensuring all staff involved in data collection are fully informed and included in the trial process may ensure adequate data recording and protocol compliance. In addition, regular assessment of data accrual may highlight the need for a change in strategy where rising missing data becomes apparent.

12.3.1.3 Protocol Deviations

In circumstances where there have been protocol deviations it is appropriate to use an 'intention to treat' (ITT) analysis where participants are analysed as part of the group they were assigned to irrespective of whether they competed their allocated treatment/intervention or not [27, 48]. The ITT analysis should be applied to a full data set [49] but frequently protocol deviations are accompanied by missing data. Failure to include participants with missing data can result in an overestimation of the benefit of the intervention [27]. Consider the barrier film example used previously, if patients stopped using the barrier film because of exacerbation of the skin reaction, or left the study due to adverse reactions, this would result in missing data for some patients. If these data are excluded from the analysis, the assumed benefit of this product may be exaggerated [27].

When considering missing data it may be appropriate to use a 'sensitivity analysis' [27] or imputation [50]. Here it is proposed that either a worst case scenario is used, or a value is chosen that is credible given the rest of the patient's data set [50]. Sometimes it may be appropriate to use the last recorded response or to assume that responses remained constant [49]. For example, in the study of the effectiveness of a radiotherapy protocol to reduce patient reports of lung morbidity mentioned previously, 'no symptoms' were used for missing data in both groups as this was a plausible outcome given the rest of the data set [39]. However, imputations of this nature are provided to give some estimation of treatment effect and should be considered carefully; producing a range of potential outcomes for readers using different imputation methods may be the most beneficial policy [48].

It is suggested that the use of the ITT analysis is the most cautious approach to take when handling protocol deviations [51]. However, it is proposed that using an ITT approach can lead to type II errors and there may be justified circumstances when patients with specific criteria could be excluded from the analysis [51]. These would include participants that were randomized for inclusion in a study but who were in fact ineligible, i.e., they did not meet the eligibility criteria for the study [51]. Even in these cases it is prudent to consider individual exclusions with care. In addition, the ITT approach is appropriate for effectiveness or pragmatic studies where the aim of the study is to evaluate the impact of an intervention under normal clinical circumstances where it is likely that some deviations from protocol would also occur [49]. Hollis and Campbell suggest a strategy for the full implementation of the ITT approach that researchers designing RCTs may find helpful [49].

When evaluating RCTs presented in the literature it is important to consider how protocol deviations were handled by the researchers. In a study of RCTs published in a number of high impact journals, Hollis and Campbell discovered only 50% of RCTs published in 1 year stated explicitly that results were analysed on an ITT basis [49]. Of those stating they used an ITT analysis 13% did not actually analyse patients as randomized (which is the criteria for the ITT approach). Furthermore, the handling of missing data was variable across the studies, emphasizing the need for practitioners to undertake rigorous appraisals of published RCT study results before considering applying the evidence to practice.

Drug Trials

When a new drug is developed the development process can be time-consuming. Initially, safety and efficacy of the drug will be tested through animal studies. The first human studies of cancer drugs (Phase I trials) are usually tested on volunteers. There is no randomization and incremental doses of the drug are administered so that side effects can be monitored. Once the safety of the drug has been established in humans the drug can then be administered to a small group of patients (approximately 20) with the condition to establish efficacy, i.e., where the aim is to establish if the drug works in people who receive it, with different doses and frequencies. There are very strict inclusion criteria to exclude patients with coexisting disease. These Phase II studies may involve randomization if the outcome measure is appropriate, i.e., pain. Where the end-point is reduction in number of deaths there may not be randomization. Phase III trials are conducted once the drug has been shown to be effective and safe in Phase II studies. Usually Phase III trials are randomized effectiveness studies.

In the cancer field a prominent and well publicized cancer drug trial was an evaluation of the effectiveness of trastuzumab (Herceptin®) for the adjuvant treatment of early breast cancer in HER2-positive cases. This monoclonal antibody against HER2 had proven efficacy in advanced breast cancer and the first interim analysis to be published in early breast cancer trials showed such promising results [52] that there was a desire for clinicians to consider its use in HER2-positive patients with early stages of the disease. The primary outcome in this study was disease free survival with early results showing 92.5% of patients in the

trastuzumab arm free from disease at year 1; as compared with 87.1% in the control arm [53]. These results led to acceptance of the drug for treatment in early stage cancer by the National Institute for Clinical Excellence (NICE) in the UK despite a relatively short follow-up period (median follow-up 2 years) [54].

12.4 Health Economic Assessment and RCTs

Economic assessments in conjunction with RCTs have become increasingly important due to the need to allocate scarce health resources in the most efficient and beneficial way. Economic evaluations deal with both costs and outcomes of activities and the basic purpose of an economic evaluation is to 'identify, measure, value and compare the costs and consequences of the alternatives being considered' [55]. Economic evaluations are comparable in the way they measure costs but differ in the way outcomes or consequences are derived. Essentially evaluations can be divided into three main types [56].

- cost-benefit analysis
- cost-utility analysis
- cost-effectiveness analysis.

Cost-benefit analysis involves the measurement of costs and benefits in comparable monetary terms. An example of the use of a cost-benefit analysis is the evaluation of an intensive follow-up regimen for patients diagnosed with breast cancer. This involves oral history, physical examination, blood tests including biological markers, annual hepatic echography, chest X-ray and a bone scan as compared with a standard clinical follow-up in breast cancer patients to identify early signs of relapse [57]. In this study the authors undertook a simple RCT comparing the two follow-up methods for number of relapses identified during scheduled follow-up appointments. The results identified no difference in the early detection of relapse between the two methods, so no benefit cost but a substantial increase in costs for the intensive follow-up schedule that was three times the cost of the less intensive follow-up regimen [57].

Cost-utility analysis involves the use of a utility based measure such as quality adjusted life years (QALYs). By using a single measure of benefit (QALYs) across RCTs, it is possible to compare the effectiveness of different interventions and hence this type of analysis allows the assessment of the benefit of employing a particular treatment or intervention in one area against the loss in benefit caused by redirecting resources from other programmes, i.e., productive efficiency and allocative efficiency, and is considered as a variation of cost-effectiveness analysis [56].

Cost-effectiveness analysis measures outcomes or benefits in units such as quality of life or improvements in function; in radiotherapy this may be measured as improvements in accuracy of treatment. To illustrate how an economic evaluation can be undertaken, consider the work by Shah et al. [58]. This work compares the cost-effectiveness of:

- standard whole breast irradiation (where patients are treated in 15 treatments with or without a boost at the end of treatment) and
- accelerated partial breast irradiation (APBI, where patients receive 5 treatments in total over 10 days).

The aim of this study was to identify the cost and cost-effectiveness of APBI compared with the current standard whole breast irradiation treatment protocol. The cost-effectiveness evaluation was undertaken from both the health care system perspective and also the societal perspective. A healthcare perspective includes direct costs for staffing, and equipment. Individual staffing costs for each patient attendance can be calculated based on the procedure time and the pay rate for the highest staff grade performing the procedure. In the study by Shah et al. [58] a breakdown of the direct costs is presented in a supplementary file on the journal web site, this is helpful to understand the breakdown of costs and where these differ between the two different approaches. A societal perspective takes in to account the impact on the patient of the treatment regime. In this study the authors calculated lost work time and parking costs for attendance at the hospital for treatments and appointments. Effectiveness in this study was determined by QALYs. Table 12.3 shows the final cost analysis. It can be seen from the table that the effectiveness of the two approaches is similar but the costs (both direct and with indirect costs considered) favour the APBI technique. It is the individual cost per treatment that influences the overall cost-effectiveness outcome in this case; the cost per treatment fraction is lower for whole breast irradiation but as there are 15 treatments compared with only

Table 12.3 Direct and indirect costs from a cost-effectiveness analysis assessing whole breast irradiation versus APBI (reproduced from Shah et al. [58] with permission)

Treatment	Cost	Incremental cost	Effectiveness (QALYs)	Incremental effectiveness (QALYs)	Incremental cost-effectiveness ratio
Direct cost only (in US dollars)					
APBI	2966	–	0.2300	–	
Whole breast irradiation (without boost)	3666	700	0.2289	−0.0011	Dominated
Whole breast irradiation (with boost)	4551	1585	0.2289	−0.001	Dominated
Direct and indirect costs (in US dollars)					
APBI	3569	–	0.2300	–	
Whole breast irradiation (without boost)	4940	1371	0.2289	−0.0011	Dominated
Whole breast irradiation (with boost)	6160	2591	0.2289	−0.0011	Dominated

APBI accelerated partial breast irradiation, *QALYs* quality adjusted life years

5 treatments in the APBI approach the overall cost is higher for the whole breast irradiation technique.

12.5 Systematic Reviews and Meta-Analyses of RCTs

During clinical activities practitioners may come across questions about practice that they do not know the answer to. They may choose to ask an expert who may or may not know the answer; or they may turn to the published literature for an answer. In Chaps. 3 and 4, literature reviews were discussed and the method for searching for literature was presented as an important aspect of the research process. This section focuses on the method for undertaking a systematic review of published literature in relation to HTA: it starts with a discussion of the differences between discussion papers (or narratives), systematic reviews and meta-analyses. Following clarification of the different types of reviews the discussion concentrates on the method for undertaking systematic reviews with particular attention paid to the review process and aspects of the search strategy including the assessment of study quality. The final subsections describes the common principles of meta-analyses and standards required for the presentation of systematic reviews.

12.5.1 Types of Reviews

Literature reviews or discussion papers found in journals are an informal collection of literature on a specific topic and are often invited papers from experts in the field. They are common in journals as they are easy to read and synthesize by practitioners and are often quick to produce. One of the main disadvantages is the variability in the level of detail that is presented about the search strategy employed, making replication of the review difficult. In addition, they may lack rigour and objectivity, with conclusions and recommendations based on a narrow examination of the available data. However, they can provide an opportunity for debate and allow the authors to provide an interesting perspective on a topic of current interest.

A systematic review is a formal review of the evidence on a particular topic with a specific research question that is to be addressed and a detailed search strategy that would allow replication. The search strategy includes details about inclusion and exclusion criteria, databases used and the method used to assess the quality of the studies identified by the search, the process for selecting research and the method used for data extraction and synthesis. There is also an attempt to reduce potential bias by using standardized tools for the assessment of study quality as well as using more than one assessor to evaluate selected studies and blinding of reviewers to the authors and journal names of selected studies.

A meta-analysis is a review where the results of RCTs undertaken independently are combined and a statistical analysis produced, usually graphically, to provide an estimate of the effect of an intervention. By combining a number of individual studies it is possible to essentially increase the overall sample size and hence increase

the strengths of the conclusions that can be drawn about an intervention, making meta-analyses a major asset for practitioners needing to make decisions about clinical interventions. However, meta-analyses do have some limitations and these are covered in more detail below.

12.5.1.1 Systematic Reviews

Planning a systematic review is crucial to its success and subsequent quality. Figure 12.6 provides a schematic presentation of the process required to plan and execute a systematic review.

Planning the Review

Before embarking on a systematic review, it is important to be clear about the clinical question that needs to be answered. The research question will be used to define facets of the search strategy and any lack of clarity may reduce the effectiveness of the search. In addition, before any detailed work is undertaken in preparation of the review it is important to identify if:

- a systematic review already exists on the topic area
- sufficient data are available to undertake a systematic review.

Therefore, being clear about the question and the topic of interest is important. Once this has been clarified it is beneficial to undertake a scoping exercise to identify how much literature exists in the field. This takes the form of a small search using the main electronic databases relevant to the topic area; for example, this might include MEDLINE, CINAHL and the Cochrane databases, using key terms. A simple search should allow the opportunity to identify whether any up-to-date systematic reviews already exist and indicate the amount of literature available to answer the proposed question [59]. Once the need for a systematic review in the field has been established a research proposal should be prepared. Box 12.3 highlights the key subheadings that practitioners may find useful to incorporate in a proposal for a systematic review [60].

Once the proposal has been written it may be helpful to gain an independent scientific review (ISR) of the proposal prior to a protocol being implemented, replicating the process undertaken for a primary study. Whereas in a primary study there is a need to gain the relevant research ethics and governance approvals, for systematic reviews there may not be such stringent requirements. However, gaining some peer review of the proposed work prior to the project being initiated is helpful for a number of reasons.

- Reviewers may identify additions to the search strategy that could improve the overall quality of the study.
- Poorly designed reviews will be ineffective and may produce results that are biased or inaccurate leading potentially to an inappropriate technology or treatment being implemented. ISR can identify potentially poor quality reviews and prevent resources being wasted on projects that may not be effective.

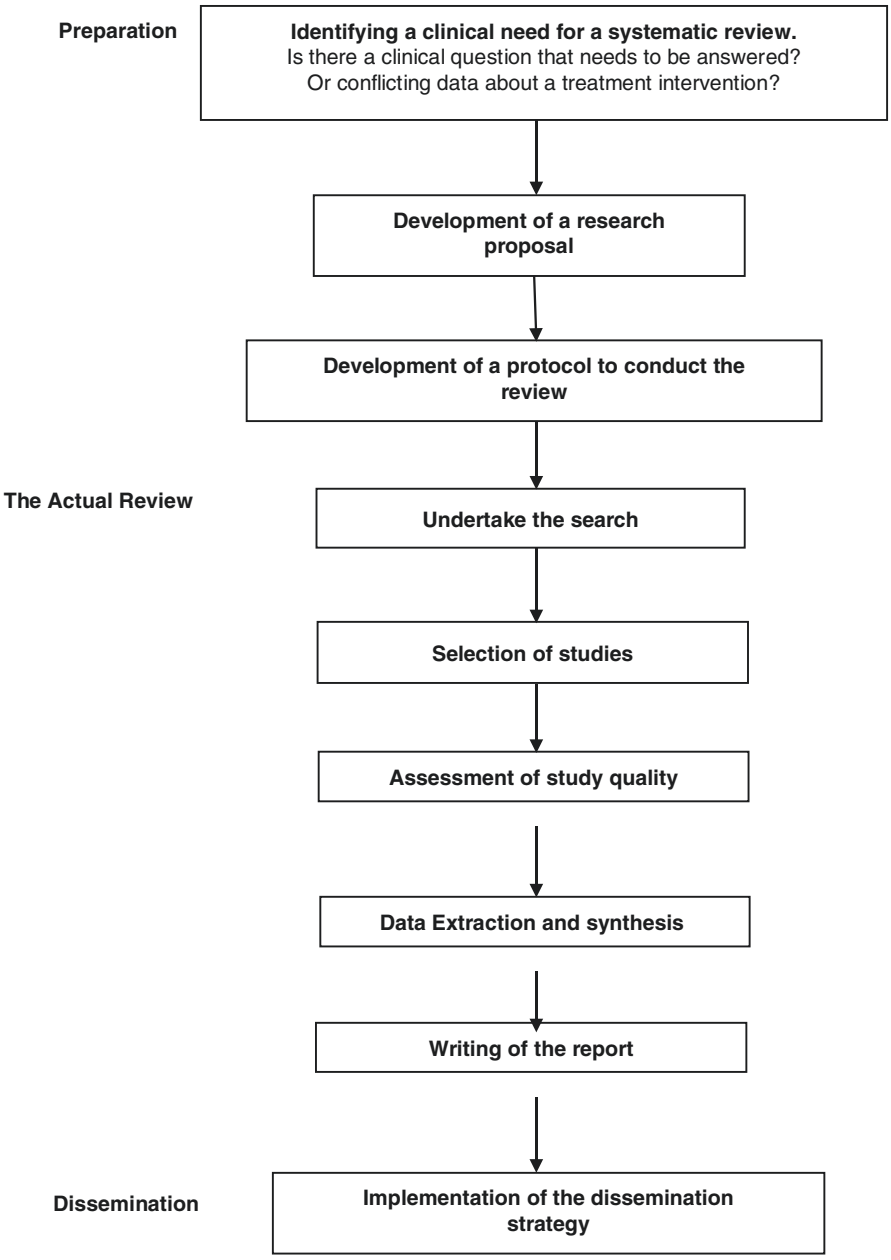

Fig. 12.6 The process for undertaking a systematic review

Box 12.3 Headings (and Content) for a Proposal for a Systematic Review

- Title.
- Summary—a brief synopsis of the aims of the review and the significance of the work will give readers an instant understanding of the importance of the proposed project.
- Aims—detail of the study aims and the research questions that the review is aiming to answer as well as the end-points for the study. End-points may include development of key research questions that remain unanswered and need further primary study or identification of a specific intervention to apply in practice.
- Background—this section should include a brief review of the literature to place the proposed work in some context, it might focus on the political, economic or social drivers for the project or give a historical perspective to current treatment or imaging rationales. Evidence identified from the scoping exercise may be beneficial in this section.
- Method—this section should include the search strategy, databases to be used, key terms, inclusion/ exclusion criteria, search limits, data extraction method, approach to be taken to quality assessment of the individual studies, how data will be synthesized (including information on any quantitative analysis), how reliability of the review will be determined and how bias will be minimized.
- Timeline—a detailed breakdown of the key milestones for the study.
- Project management—how the study will be managed and the key roles of members of the project team.
- Dissemination strategy—details of how the results will be disseminated should be multifaceted and practitioners may find it useful to consider how to measure impact, this work by Cruz Rivera et al. [71] helpful.
- Costs—identified costs including time for researchers undertaking the review, costs of searching databases (there may be a cost for access to some databases), costs of retrieving articles as well as costs to disseminate the results.

Funding bodies provide ISR during the application approval process but practitioners may wish to seek peer review prior to a funding application and this may be available locally through a university or via the local research and development department of the employing organization. Undergraduate and postgraduate students can use the experience of their supervisors to review the quality of their proposal.

The Search Strategy

Developing a multifaceted search strategy should ensure the review identifies as much of the available evidence as possible. The search strategy should detail the databases to be searched, the key terms for the search, inclusion and exclusion criteria and any limits placed on the search. Box 12.4 provides an example of a search

Box 12.4 A Sample Search Strategy

1. *Databases*
   a. MEDLINE
   b. CINHAL
   c. EMBASE
   d. Cochrane Reviews database,
   e. National Research Register including the ongoing reviews database (CRD Register of reviews)
   f. LILACS Latin American and Caribbean Literature in Health Sciences
   g. ISI Web of Knowledge to search Science Citation Index to follow citations from key papers.
   h. ScienceDirect to search for articles from journals not listed on MEDLINE
2. *Websites- to identify professional reports*
   a. National Institute for Clinical Excellence (NICE)
   b. National library for Health
   c. TRIP (Turning Research into Practice)
   d. Intute: Health and Life Sciences Medicine http://www.intute.ac.uk/healthandlifesciences/medicine/
   e. UK Society and College of Radiographers www.sor.org
   f. UK College of Radiologist www.rcr.ac.uk
3. *Key Journal hand Searches*: these will vary according to the topic area but common journals of relevance may include:
   a. *Radiation Therapy*
      i. Radiotherapy and Oncology
      ii. International Journal of Radiation Oncology Biology Physics
      iii. Journal of Radiotherapy in Practice
      iv. European Journal of Cancer
      v. Clinical Oncology
   b. *Imaging*
      i. British Journal of Radiology
      ii. Clinical Radiology
      iii. Radiology
   c. *Imaging and Radiation Therapy*
      i. Radiography
      ii. Journal of Medical Imaging and Radiation Sciences
4. *Author Searching*: searching databases by author may be beneficial where an author is known to publish or is a known expert in the topic area, this may be identified from literature retrieved in the original scoping exercise.
5. *Grey Literature*
   a. Index to Theses
   b. Index to Scientific and Technical Proceedings (via ISI web of knowledge)

c. Conference Papers Index
d. British Library Integrated Catalogue
e. COPAC–merged online catalogue of major university and national libraries in the UK and Ireland
f. Clinical trials databases here are the UK and US web addresses https://bepartofresearch.nihr.ac.uk https://clinicaltrials.gov

6. *Key Words*: for each facet of the research question key words and MEDLINE subject headings should be identified, for example, if a facet of the question included 'patients with cancer', then keywords might include:
   a. Carcinoma, tumour, tumour, cancer, invasive carcinoma
   b. MEDLINE subject heading—neoplasms
7. *Inclusion/Exclusion Criteria*: these may be specific to the topic area, for example, factors in a review to identify the effectiveness of partial breast irradiation inclusion criteria may be studies that consider external beam methods as well as brachytherapy (including balloon catheter methods). Alternatively, the focus for inclusion may be on the types of studies to be included. For example, in effectiveness reviews it may be relevant to include RCTs or quasi-experimental studies (trials without randomization).
8. *Search Limits*: For studies in HTA it is sensible to limit the review to data produced once the technology under question was implemented. For practical reasons undergraduate and postgraduate students often choose to limit studies to those published in the English language but possible bias needs to be considered where this is adopted.

strategy with common databases, websites and other strategies that may be useful for those working in imaging or radiation therapy. The Cochrane Collaboration provides a useful starting point for a search strategy along with the other major electronic databases. The Cochrane Library contains the Cochrane Database of Systematic Reviews (CDSR) and the Cochrane Central Register of Controlled Trials (CENTRAL). It may be beneficial to also search the Cochrane Methodology Register (CMR), the National Institute for Health Research (NIHR) Health Technology Assessment Database (HTA) and the NHS Economic Evaluation Database (NHS EED).

A detailed protocol using the main databases listed in Box 12.4 will go some way to helping retrieve as many of the relevant research studies as possible. However, in complex reviews it is possible that the protocol itself may only identify a proportion of the available data, and researchers should try to broaden their approach to include a range of strategies that develop as the review progresses.

For example, use of snowballing, which is using the reference lists of retrieved articles and forward tracking from a selected article to identify articles that have subsequently cited this paper—citation tracking, can increase the yield of relevant articles, and has been shown to account for approximately 53% of articles used in a

complex systematic review [61]. Other strategies to consider include using personal networks to contact individuals who may know of relevant research. This type of informal approach has been found to increase the proportion of relevant articles for a review by approximately 60% [61].

Searching the grey literature is also of importance as this may limit the effect of publication bias [62]. In a systematic review of studies including grey literature as well as published trials, it was identified that published trials tended to show a greater treatment effect than grey literature. This may be due to differences between published and unpublished trials such as sample size differences, and grey literature studies finding the intervention has no effect, which is a less interesting result and less likely to be published [62]. Grey literature refers to studies not yet formally published and may be found in conference proceedings, indexes to theses or on trial registers.

A common problem with using electronic databases as the primary search strategy is their lack of sensitivity in some cases to identify all the relevant RCTs that have been published. The Cochrane Collaboration has developed a sensitive search strategy that should allow greater search precision and using this database to identify effectiveness trials should be a fundamental part of the search strategy. Other strategies to maximize retrieval of all relevant trials is the use of electronic databases searches that contain journals not registered with Medline. Some new journals may not be registered with electronic databases such as MEDLINE or CINAHL so individual hand searching of these journals and other key journals known to publish research in the field of interest should be considered. Hand searching has been shown to identify between 92% and 100% of the total number of trials identified from both hand searching and electronic searching [63], with MEDLINE identifying 55% of the total trials identified [63]. While hand searching is a useful additional strategy it is time-consuming, involving review of each article, review and letter published in each issue of the chosen journal to identify relevant work.

Another aspect of the search strategy that practitioners need to consider is the restriction of the search to English language journals. This is often undertaken for simplicity in undergraduate and postgraduate studies and where funding is not available for translation. There is a possibility that limiting the search in this way may bias the outcome of the review, but evidence about the impact of such a strategy is unclear. It has been identified that the quality of English language versus non-English language articles is the same [64, 65], but it is possible that research published in non-English journals is less likely to demonstrate a significant result [64], so by their exclusion may alter the outcome of any meta-analysis. However, in a review of language-restricted and language-inclusive meta-analyses, no difference in estimates of benefit was identified [66, 67]. It is therefore difficult to predict the overall impact of excluding non-English language studies.

Quality Assessment

Chapter 4 highlighted the importance of critical appraisal of the published literature and identified a range of tools that can be used to help in the appraisal process. A number of tools are reported in the literature and these include checklists, as well

as scales, with many different quality assessment tools available. Quality is a difficult construct to define for the range of research that a practitioner is likely to come across and no one tool may be appropriate for a range of topic areas. The QUADAS-2 tool for the assessment of the quality of studies of diagnostic accuracy is a validated tool that has built on the original QUADAS tool based on a consensus Delphi study [68, 69]. All quality assessment tools should be developed using formalized methods of development with assessments of face, content and construct validity and tested for reliability across different raters [27]. A tool developed initially to assess the quality of RCTs in pain research (the Jadad scale) was also based on a Delphi consensus method of agreement of experts and has been proposed for use across a range of clinical trials [70]. This tool uses a scale from 0 to 5 with reviewers scoring the answers to three questions as either yes (scores 1 point) or no (scores no points), with additional points awarded where blinding and randomization were appropriate [70].

In contrast, the Cochrane Collaboration recommends a domain-based approach to quality assessment of RCTs including assessment of the following [71].

- sequence generation
- allocation concealment
- blinding of participants, personnel and assessors
- incomplete outcome data
- selective outcome reporting
- other sources of bias.

Assessment tools often consider the internal validity of the study as reported but published trials judged by assessment tools to be low quality may actually reflect poor reporting rather than poor design quality, resulting from a lack of understanding on how to report a clinical trial, a problem of under-reporting. To overcome problems associated with poor reporting, it is now possible to publish trial protocols in peer reviewed journals; allowing readers to see greater detail of study designs and allowing better assessment of study quality of the final published results manuscript.

A quality assessment threshold should be identified to exclude weak studies from the review and can be achieved by applying a cutoff level for study selection. This may be based on quality assessment criteria identified above, as well as using a hierarchy of study designs. For example, in effectiveness studies the primary research question is based on an assessment of one intervention over another, which is best studied using a RCT with concealment of allocation. Where these are not available the next best design should be chosen, i.e., quasi-experimental studies where there is no randomization or cohort studies [59].

The ROBINS-I (Risk Of Bias In Non-Randomized Studies of Interventions) tool has been designed to facilitate researchers in assessing the quality of research involving non-randomized cohorts. Details of the ROBINS-I can be found here https://sites.google.com/site/riskofbiastool/welcome/home, and a useful guide to using the tool was published by Sterne et al. [72]. For reviews considering test

accuracy the hierarchy of study designs differs and the method at the top of the hierarchy is a blind comparison where there is a reference standard and a broadly defined sample of consecutive patients. Similarly, where these do not exist or are limited for the test under review it may be necessary to include studies where there is a narrow population or differential use of a reference standard [59].

When attempting to assess trial quality it is helpful to use a data collection/extraction form that includes details of the bibliographic reference, description of study characteristics and the quality assessment. This can then be used to develop a table of evidence comprising all the included trials. Examples of such forms can be found on the Scottish Intercollegiate Guidelines Network (SIGN) website (https://www.sign.ac.uk/assets/sign50_2015.pdf).

Regardless of the chosen assessment tool or threshold level chosen it is important that the quality assessment is not only integrated into the selection of studies for inclusion in the review but also incorporated within the results that are presented. However, in many published systematic reviews, while quality assessment is apparent in the selection of included trials, the quality of the selected studies is not always transparent in the final reporting of the results [73]. Quality assessment should be incorporated into the systematic review process at the selection of studies phase, in the interpretation of conflicting trial results, in the weight apportioned to trials within a meta-analysis and in the conclusions and recommendations of the review [59]. This can be achieved in its simplest form by a description of the results with a review of any risks of bias within the individual studies included. It can also be achieved by listing the quality score, or using the method adopted by SIGN, where ++ refers to high quality, + refers to acceptable quality and '-' refers to low quality (see the SIGN checklists https://www.sign.ac.uk/checklists-and-notes.html) against the tabulation of the individual trial characteristics so that readers can instantly see how the study quality may influence the overall outcomes of the review. See Table 12.4 for an example of where the quality assessment has been included in an evidence table in a published systematic review [74]. In this systematic review only research that scored '+' or '++' was used to draw conclusions.

The Cochrane handbook [71] provides good guidance on how to report the risk of bias in studies included in a systematic review and it is worth taking a look at the Cochrane handbook available here https://training.cochrane.org/handbook.

Meta-Analysis

Where individual studies allow, a formal quantitative analysis of the results may be undertaken in the form of a meta-analysis. This quantitative analysis provides a precise estimation of intervention effects and can indicate heterogeneity between studies where this exists. Including inappropriate studies in the meta-analysis can lead to misleading results, hence care needs to be taken in the execution of the analysis. For systematic reviews that include meta-analysis inclusion criteria need to prescribe the characteristics of studies that allow them to be combined in the meta-analysis; this may be trials studying the same intervention with the same outcome measures, undertaken on patients with similar characteristics (such as age or disease type).

Table 12.4 Example of evidence table-immobilization literature (supine position) from Probst et al. [74]

Author+ year	Description	Accuracy	*n*	Materials used on the breast	Skin reactions	Advs/disad	QA
Latimer et al. (2005) [13]	A micro-shell vs two other breast rings	Not measured	8	Polyacrylic micro-shell shaped into a horse-shoe	Micro-shell increased surface dose by 9%, other devices increased by 22%	• Shaped to reduce skin dosage • Reusable • Expandable capacity	+
Carter et al. (1997) [26]	Retrospective review	CLD variability average = −1.2 mm	20	Alpha Foam cradle	Not applicable	• No patient demographic available so unable to assess impact of patient size on reproducibility • No control group for comparison	-
Thilmann et al. (1998) [18]	Comparison between a positioning support cushion and no immobilization	Mean error without support 8.4 mm vs 6.1 mm	55	Foam	Not observed	Accuracy significantly improved with support (72% more comfortable)	+
Graham et al. (2000) [19]	Randomization to armrest or vacuum bag immobilization	Lung exposure (mean SD): Vac-bag 0.21 cm (95% CI 0.17–0.26) Arm-rest 0.21 cm (95% CI 0.17–0.24)	30	None thorax stabilization	Less skin folds present in armrest	Armrest more comfortable, vacuum bag allowed less lung exposure, no difference in stability and setup time	+
Nalder et al. (2001) [20]	Comparison of standard breast board and vacuum bag attached to a breast board	mean and SD of the systematic errors (mm): *With VB* AP −1.8 (2.9) *No VB* AP −1.7 (2.8) SD of the random errors: *With VB* AP 2.6 *No VB* AP 2.2	17	Not stated	n/a	• Minimal improvements found using the VB • Majority found the VB more comfortable	+

Bentel et al. (1999) [21]	Patients with large and/or pendulous breasts underwent radiotherapy using a breast ring; comprised of a hollow tube and fitted around the breast in contact with the skin	n/a	56	PVC tube (other material of tube tested was nylon)	Moist desquamation in 60.7% Surface dose under the ring approximately 85% of D_{max} dose. Without ring surface dose 35%	• Reduce skin folds and lateral movement in supine position no quantitative data • Good cosmetic outcome reported	-
Strydhorst et al. (2011) [22]	Assessment of the effect of a thermoplastic immobilization device on minimizing breast/chest wall movement during chest wall/breast irradiation	Inter-fraction motion: average random error Left/rt = 4 mm Sup/inf = 12 mm and AP = 4.5 mm Intra-fraction motion: av = 1 mm	$N = 8$	Thermoplastic shell	Not measured	Inter-fraction motion appears large which would indicate this method of immobilization does not work well	-
Cross et al. (1989) [23]	Feasibility study to assess the usefulness of the lateral decubitus position for women with very large breasts	Not measured	$N = 4$	Styrofoam block plus alpha cradle	all developed moist desquamation inferiorly due to contact with styroform foam, surface dose increased from 40 to 80%	Conclude lateral decubitus position feasible for women (cup size EE). technique does not allow matching of an scf	-

(continued)

Table 12.4 (continued)

Author+ year	Description	Accuracy	*n*	Materials used on the breast	Skin reactions	Advs/disad	QA
Goldsworthy et al. (2010) [24]	RCT comparing positioning on a breast board with either both arms abducted (intervention group) or single arm abducted. (control group)	*CLD* systematic error mean = −1.7 mm vs −1.9 mm $p = 0.06$, population systematic error 4 mm vs 2.3 mm $p = 0.005$ in favour of intervention. Population random error 2.1 mm vs 1.6 mm p=0.055	50	Traditional breast board with armpole device	not measured	The use of bi-lateral arm abduction resulted in smaller set up errors than the single arm positioning, although differences small	+
Zierhut et al. (1994) [25]	A repeated measures design to test the usefulness of a thermo plastic immobilization device. Patients were treated in the thermoplastic but simulation data available with and without the device	AP mean deviation = 3 mm with the device. sup-inf 4.1 mm	7	Thermoplastic	Surface dose increased from 47% to 64% on patients, on the phantom the surface dose was increased from 51 to 64% (of the maximum dose). The increase in skin dose was 17%	The increase in skin dose was 17%	-
Chopra et al. (2006) [29]	A case series	Displacements: Sup-inf = 1.3 mm Med-lat = 1.3 mm Ant-post = 4.4 mm	5	Vacuum bag immobilization	Not measured/Not applicable	Patient demographics not reported, no control group for comparison	-

Creutzberg et al. (1993) [27]	Non-randomized trial 1. patients lying flat with plastic mask (n = 17) 2. patients no mask (n = 14) 9 on inclined wedge, 5 lying flat	Ventral-dorsal displacement: With mask = 3.2mm Without mask = 4.6 mm	31	Plastic mask vs no mask And flat vs inclined on a wedge	Not measured	Not clear the criteria for allocation (except for those with additional nodal fields), no patient demographic data	-
Valdagni and Italia (1991) [28]	Case series	Ventral-dorsal shift = 2.7 mm (± 2.2 mm) Craniocaudal shift = 1.9 mm (± 1.8 mm)	20	Plastic mask immobilization		No control group for comparison Patient demographic data, no information on observer reliability	-
Keller et al. (2013) [30]	A commercially available bra/bustier compared with no bra	Not measured	N = 246	Commercial bra using thin plastic stays	Bra: 90% of cases grade 2 dermatitis No bra: 70% (p = 0.003)	Baseline characteristics were uneven across control and intervention (i.e., more cases with larger breast cup size in the intervention group), no randomization between control and intervention	-

++ = high quality low risk of bias, + = acceptable quality with some risk of bias, - = high risk of bias

The meta-analysis itself involves combining the results of all included studies that are combinable, i.e., have the same outcome measure. The individual trial results are weighted according to trial size although weighting based on trial quality has been proposed [75]. The methods used to combine the data are defined by two models, 'fixed effects' and 'random effects'. The choice of model depends on the presence of heterogeneity or variability between studies. Variability across studies, i.e., between-studies heterogeneity, can be assessed using either a Q statistic or an I^2 index [76]. The Q statistic produces a binary outcome identifying whether heterogeneity is present or absent. The I^2 index has been proposed as it gives a better indication of the level of heterogeneity that is present. Studies with an I^2 index of 25%, 50% and 75% would be classified as having low, medium or high variability, respectively [76].

The 'fixed effects' model combines the results of studies assuming that the effect of the intervention is constant across studies so only within-study variation is included in the analysis. In contrast, the 'random effects' model is based on the premise that the true treatment effect is different across individual studies [77] and this method is preferred when variability across studies is high [75, 76].

The results of combining data are often presented in graphical form; traditionally, this has been using a forest plot like the one in Fig. 12.7.

Figure 12.7 is a forest plot from the independent review of breast cancer screening trials [78].

Each trial is described by one line. Squares indicate the relative risk of death from breast cancer from screening versus the non-screened population for each trial. The horizontal line on forest plots usually defines the 95% confidence intervals. The solid vertical line indicates a ratio of 1.0 (i.e., 1.0 indicates no difference between screened and non-screened populations) trials that fall on the solid line would indicate no benefit from screening. For each category of trial the total ratio (relative risk) are shown as a diamond. The overall results of this meta-analysis identified a benefit from screening; the reduction in breast cancer mortality in those invited for screening was estimated to be 20% (95% CI 11–27). There was some heterogeneity between the trials but this was not statistically significant. However, the confidence interval around the RR of 0.8 is reasonably large (i.e., 0.73–0.89).

This meta-analysis serves to highlight an important dilemma in HTA primarily that when mature data are available for analysis, the technology and treatments for the condition may have moved on substantially, making the outcomes difficult to interpret within a new context. In some of the breast cancer screening trials included in this analysis there has also been discussion about the internal validity and therefore the accuracy of the predicted benefits of screening programmes.

Meta-analyses do have limitations which may be ascribed to the quality of the original RCTs available for analysis. As described in the previous section, inadequate sample sizes or opportunities for bias, such as inadequate concealment, may reduce the quality of the research which may then lead to inaccuracies in subsequent meta-analysis. Furthermore, research with a positive result is more likely to be published than a study showing no treatment or intervention benefit. Therefore,

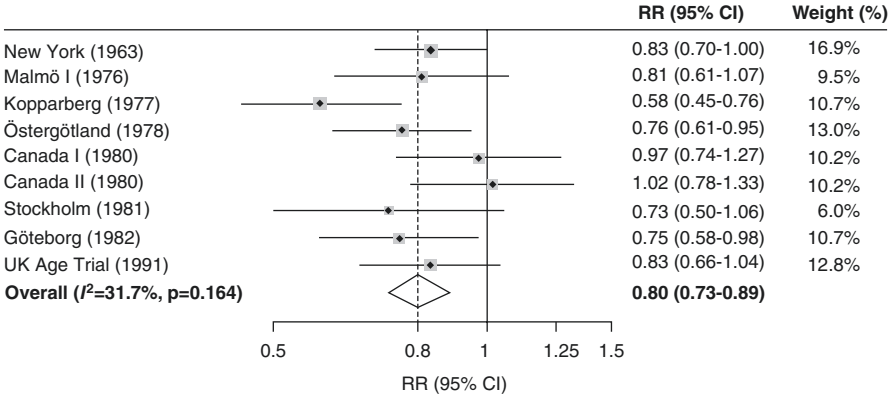

Fig. 12.7 Forest plot example—Meta-analysis of breast cancer mortality after 13 years of follow-up in breast cancer screening trials. Adapted from the Cochrane Review [78]. *RR* relative risk. Malmö II is excluded because follow-up of about 13 years was not available; the Swedish Two County (Kopparberg and Östergötland) and Canada I and II trials are split into their component parts; the Edinburgh trial is excluded because of severe imbalances between randomized groups. Weights are from random effects analysis.

meta-analyses may suffer the effects of publication bias if search strategies to identify eligible studies exclude the grey literature. In addition, meta-analyses suffer the risk of bias that may occur from the process of undertaking a systematic review including bias in the selection of studies, the assessment of study quality by the reviewers and problems with poor reporting of study results or errors in the data of the published reports [77]. A method proposed to identify publication bias in meta-analyses is the use of a simple graphical presentation of the individual trials estimate of treatment effect plotted against the trial sample size (funnel plot). If there is no bias, the plot should be symmetrical, depicting an inverted funnel with greatest dispersion of effects among trials of small sample sizes and a less marked dispersion in trials with larger sample sizes [75, 77], with meta-analyses that contain bias demonstrating asymmetrical funnel plots [77].

Reporting the Results of a Systematic Review

In a review of the methods of reporting of systematic reviews of diagnostic tests in cancer, Mallett et al. [79] identified significant variability in reporting of critical criteria such as defining the target condition where 51% failed to report if tumours were primary, recurrent or metastatic, with equal failings when it came to reporting tumour stage. To improve the quality of reporting of systematic reviews a consensus report (by the QUOROM group) proposed a checklist of items and a flow diagram that should be included in systematic reviews and meta-analyses [80]. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist and flow diagram can be downloaded from the author section of most leading journal websites or from

here http://www.prisma-statement.org, and consists of 27 headings and subheadings to guide authors in the reporting and quality assessment of this type of research [81]. The PRISMA guidance covers the detail that is needed in the reporting of the search strategy, selection of studies for inclusion in the review, quality assessment of the selected trials, method of data extraction, details of the study characteristics, the quantitative data analysis (if there is any) the discussion of the results and the reporting of funding of the review. PRISMA also suggests the use of a flow diagram to indicate the number of trials identified, those included, and information about trials that were excluded. Figure 12.8 is an example of a PRISMA diagram from a published systematic review [74].

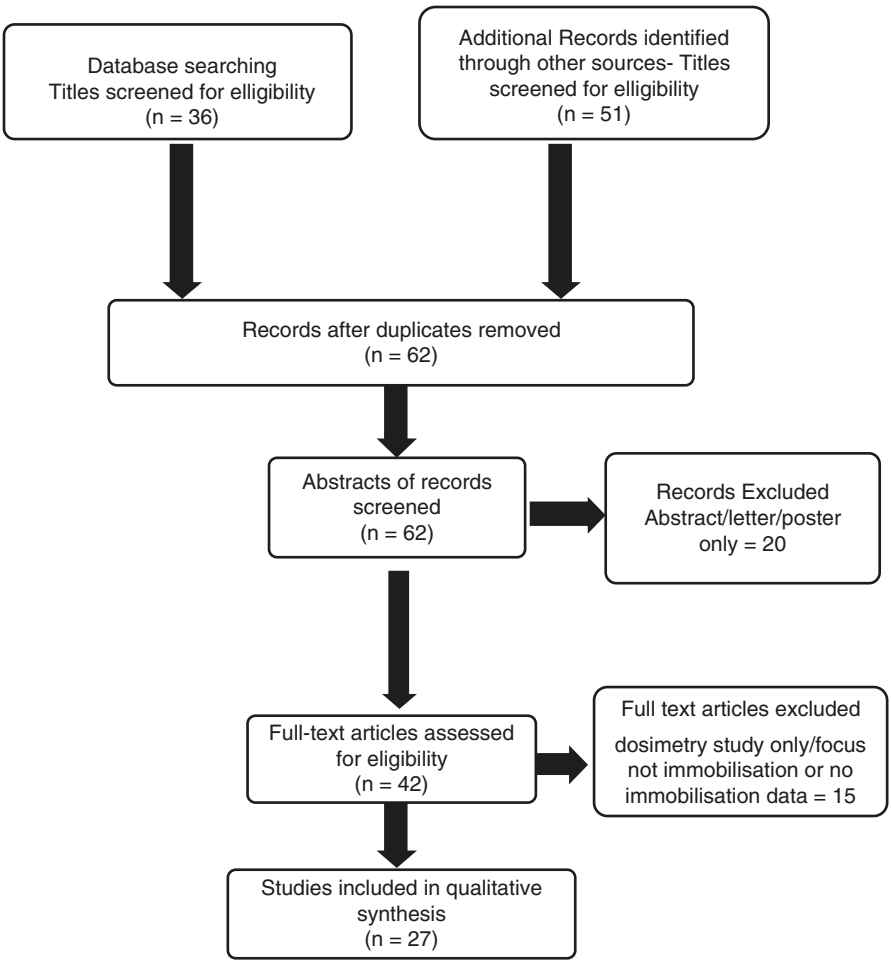

Fig. 12.8 An example flowchart demonstrating how articles were included and excluded from a systematic review by Probst et al. [74]

12.6 Conclusion

This chapter has provided an insight into the research design approaches that are useful for assessing new diagnostic imaging technologies and radiotherapy interventions. The use of healthcare technology has medical, social, ethical and economic considerations. Although in diagnostic imaging and radiotherapy this technology tends to be complex, many healthcare approaches can rely on quite simple devices. The full evaluation of a diagnostic test requires assessment at every level of the evaluative framework to demonstrate how good quality images contribute to accurate diagnoses, beneficial changes to diagnoses and management plans and improved patient outcomes, at acceptable costs.

Randomization is important for ensuring a balance in the characteristics of patients between groups and should be performed remote from clinical practice to help ensure adequate concealment in treatment allocation. The sample size calculations should be conducted based on clinically significant improvements in the primary outcome measure. The recruitment of patients is a major challenge in clinical trials. The methods to facilitate recruitment must include careful consideration of the participant consent process, inclusion of important members of the multidisciplinary team to encourage recruiting participants and a realistic timeframe to recruit a sufficient sample size and adequate funding. The attrition in the follow-up of patients should be limited by considering methods to reduce participant burden such as questionnaire length and minimizing the collection of missing data. With the increasing emphasis on resource allocation it is important to consider the economic implications of any new technology or new process. In HTA a cost-effectiveness analysis maybe appropriate and can be considered alongside the design of the RCT. Systematic reviews differ from the conventional type of review in that they adhere to strict scientific design to make them more comprehensive, to minimize bias and errors thus providing more reliable results to support evidence-based decision-making in policy and practice. It is important to examine variation or heterogeneity across studies to inform the choice of statistical model (‘fixed effects’ or ‘random effects’) for pooling the results of studies. ‘Healthcare technology’ is a broad term and encompasses a variety of instruments and techniques which promote health, prevent and treat disease, and enhance rehabilitation.

References

1. Fineberg HV, Bauman R, Sosman M. Computerised cranial tomography: effect on diagnostic and therapeutic plans. JAMA. 1977;238:224–7.
2. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Mak. 1991;11:88–94.
3. Mackenzie R, Dixon AK. Measuring the effects of imaging: an evaluative framework. Clin Radiol. 1995;50:513–8.
4. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard-lessons from the history of RCTs. N Engl J Med. 2016;374:2175–81.
5. Robinson PJA. Radiology’s Achilles’ heel: error and variation in the interpretation of the Röntgen image. Br J Radiol. 1997;70:1085–98.

6. Brealey S, Scally AJ. Methodological approaches to evaluating the practice of radiographers' interpretation of images: a review. Radiography. 2008;14(1):e46–54.
7. Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. p. 19–38.
8. Sackett DL, Haynes RB. Evidence base of clinical diagnosis: the architecture of diagnostic research. BMJ. 2002;324:539–41.
9. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, Irwig L, Levine D, Reitsma JB, de Vet HCW, Bossuyt PMM. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.
10. Whiting P, Westwood M, Rutjes AWS, et al. Evaluation of QUADAS, a tool for the quality assessment of diagnostic accuracy studies. BMC Med Res Methodol. 2006;6:9.
11. Kelly S, Berry E, Roderick P, et al. The identification of bias in studies of the diagnostic performance of imaging modalities. Br J Radiol. 1997;70:1028–35.
12. Deeks J. Systematic reviews of evaluations of diagnostic and screening tests. In: Egger M, Smith GD, Altman G, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Publishing Group; 2001. p. 248–82.
13. Hajjan-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627–35.
14. Habbema JDF, Eijkemans R, Krijnen P, et al. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Books; 2002. p. 117–44.
15. Brealey S, Scally AJ. Bias in plain film reading performance studies. Br J Radiol. 2001;74:307–16.
16. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–46.
17. Lee W. Technology assessment: vigilance required. Int J Radiat Oncol Biol Phys. 2008;70(3):652–3.
18. Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. Br Med J. 1998;317:1185–90.
19. Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014;43(4):1272–83.
20. Scottish Intercollegiate Guidelines Network Health Improvement Scotland. SIGN 50: a guideline developers handbook. Quick Reference Guide. 2015.
21. Herst PM, Bennett NC, Sutherland AE, Peszynski RI, Paterson DB, Jasperse ML. Prophylactic use of Mepitel Film prevents radiation-induced moist desquamation in an intra-patient randomised controlled clinical trial of 78 breast cancer patients. Radiother Oncol. 2014;110(1):137–43.
22. Goldsmith C, Haviland J, Tsang Y, Sydenham M, Yarnold J. Large breast size as a risk factor for late adverse effects of breast radiotherapy: is residual dose inhomogeneity, despite 3D treatment planning and delivery, the main explanation? Radiother Oncol. 2011;100(2):236–40.
23. Noble-Adams R. Radiation induced reactions 2: development of a measurement tool. Br J Nurs. 1996;8(18):1208–11.
24. Noble-Adams R. Radiation induced reactions 3: evaluating the RISRAS. Br J Nurs. 1999;8(19):1305–12.
25. Radiation Therapy Oncology Group. Cooperative Group Common Toxicity Criteria. Minimize. 2019. https://www.rtog.org/ResearchAssociates/AdverseEventReporting/CooperativeGroupCommonToxicityCriteria.aspx.
26. Neal A, Torr M, Helyer S, et al. Correlation of breast dose heterogeneity with breast size using 3D CT planning and dose volume histograms. Radiother Oncol. 1995;34(3):210–8.
27. Jadad A. Randomised controlled trials: a user's guide. London: BMJ Books/Wiley; 2004.
28. Moss S, Thomas I, Evans A, Thomas B, Johns L. Randomised controlled trial of mammographic screening in women from age 40: results of screening in the first 10 years. Br J Cancer. 2005;92(5):949–54.

29. Hendrick RE, Smith RA, Rutledge JH, et al. Benefit of screening mammography in women aged 40–49: a new meta-analysis of randomised controlled trials. J Natl Cancer Inst Monogr. 1997;1997(22):87–92.
30. Probst H, Griffiths S. Increasing the work speed of radiographers: the effect on the accuracy of a setup of a complex shaped cranial field, part of a matched cranio spinal junction. Radiother Oncol. 1996;38(3):241–5.
31. Norrman E, Persliden J. A factorial experiment on image quality and radiation dose. Radiat Prot Dosim. 2005;114(1–3):246–52.
32. Probst H, Dodwell D, Gray JC, et al. An evaluation of the accuracy of semi-permanent skin marks for breast cancer irradiation. Radiography. 2006;12(3):186–8.
33. Cancer Research UK. Breast cancer incidence 2018. Accessed May 2019. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/incidence-invasive#heading-One.
34. Roberts C, Torgerson D. Understanding controlled trials randomisation methods in controlled trials. Br Med J. 1998;317:1301.
35. Torgerson DJ. Contamination in trials: is cluster randomisation the answer? Br Med J. 2001;322:355–7.
36. Kennedy ADM, Torgerson DJ, Campbell MK, Grant AM. Subversion of allocation concealment in a randomised controlled trial: a historical case study. Trials. 2017;18(1):204.
37. Torgerson DJ, Roberts C. Understanding controlled trials randomisation methods: concealment. BMJ. 1999;319(7206):375–6.
38. Pocock SJ. The size of a clinical trial. Clinical trials: a practical approach. Chichester: Wiley; 2008. p. 123–41.
39. Probst H, Dodwell D, Gray J, Holmes M. Radiotherapy for breast carcinoma: an evaluation of the relationship between the central lung depth and respiratory symptoms. Radiography. 2005;11(1):3–9.
40. Corrie P, Shaw J, Harris R. Rate limiting factors in recruitment of patients to clinical trials in cancer research: descriptive study. Br Med J. 2003;327:320–1.
41. Welton A, Vickers M, Cooper J, et al. Is recruitment more difficult with a placebo arm in randomised controlled trials? A quasi-randomised, interview based study. Br Med J. 1999;318:1114–7.
42. Hancock BW, Aitken M, Radstone C, et al. Why don't cancer patients get entered into clinical trials? Experience of the Sheffield Lymphoma Group's collaboration in British National Lymphoma Investigation studies. BMJ. 1997;314(7073):36.
43. Mitchell G, Abernethy AP, Investigators of the Queensland Case Conferences Trial, Palliative Care Trial. A comparison of methodologies from two longitudinal community-based randomized controlled trials of similar interventions in palliative care: what worked and what did not? J Palliat Med. 2005;8(6):1226–37.
44. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: cochrane systematic review and meta-analysis. BMJ Open. 2013;3(2):e002360.
45. Torgerson DJ, Sibbald B. Understanding controlled trials. What is a patient preference trial. BMJ. 1998;316(7128):360.
46. Torgerson DJ, Roland M. What is Zelen's design? BMJ. 1998;316(7131):606.
47. Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ. 2006;332(7547):969–71.
48. Pocock SJ. Protocol deviations. Clinical Trials: a practical approach. Chichester: Wiley; 2008. p. 176–86.
49. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670–4.
50. Altman DG, Bland JM. Missing data. BMJ. 2007;334(7590):424.
51. Fergusson D, Aaron SD, Guyatt G, et al. Post-randomisation exclusions: the intention to treat principle and excluding patients from analysis. BMJ. 2002;325(7365):652–4.

52. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353(16):1659–72.
53. Smith I, Procter M, Gelber RD, et al. 2-year follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer: a randomised controlled trial. Lancet. 2007;369(9555):29–36.
54. National Institute for Clinical Excellence. Trastuzumab for the adjuvant treatment of early-stage HER2-positive breast cancer. London: NHS Department of Health; 2006 NICE Technology Appraisal Guidance 107.
55. Drummond MF, O'Brien B, Stoddart GL, et al. Basic types of economic evaluation: methods for the economic evaluation of health care programmes. 2nd ed. Oxford: Oxford University Press; 1997. p. 6–26.
56. Palmer S, Byford S, Raftery J. Economics notes: types of economic evaluation. BMJ. 1999;318(7194):1349.
57. Amparo O, Santaballa A, Munarriz B, et al. Cost-benefit analysis of a follow-up program in patients with breast cancer: a randomized prospective study. Breast J. 2007;13(6):571–4.
58. Shah C, Ward MC, Tendulkar RD, Cherian S, Vicini F, Singer ME. Cost and cost-effectiveness of image guided partial breast irradiation in comparison to hypofractionated whole breast irradiation. Int J Radiat Oncol Biol Phys. 2019;103(2):397–402.
59. Centre for Reviews and Dissemination. Undertaking systematic reviews of research on effectiveness: CRD's guidance for those carrying out or commissioning reviews. Centre for Reviews and Dissemination, University of York; 2001. Centre for Reviews and Dissemination Report 4.
60. Cruz Rivera S, Kyte DG, Aiyegbusi OL, Keeley TJ, Calvert MJ. Assessing the impact of healthcare research: a systematic review of methodological frameworks. PLoS Med. 2017;14(8):e1002370. https://doi.org/10.1371/journal.pmed.1002370.
61. Greenhalgh T, Peacock R. Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources. BMJ. 2005;331(7524):1064–5.
62. Hopewell S, McDonald S, Clarke M, et al. Grey literature in meta-analyses of randomised trials of healthcare interventions. Cochrane Database Syst Rev. 2007;(2):MR000010.
63. Hopewell S, Clarke M, Lefebvre C, et al. Handsearching versus electronic searching to identify reports of randomised trials. Cochrane Database Syst Rev. 2007;(2):MR000001.
64. Egger M, Zellweger-Zahner T. Language bias in randomised controlled trials published in English and German. Lancet. 1997;350(9074):326.
65. Moher D, Fortin P. Completeness of reporting of trials published in languages other than English: implications for conduct and reporting of systematic reviews. Lancet. 1996;347(8998):363.
66. Moher D, Pham B, Klassen TP, et al. What contributions do languages other than English make on the results of meta-analyses. J Clin Epidemiol. 2000;53(9):964–72.
67. Juni P, Holenstein F, Sterne J, Bartlett C, et al. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002;31(1):115–23.
68. Whiting P, Rutjes A, Reitsma J, et al. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003;3(1):25.
69. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JAC, Bossuyt PMM. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
70. Jadad RA, Moore D, Carroll C, et al. Assessing the quality of reports of randomized clinical trials: is blinding necessary? Control Clin Trials. 1996;17:1–12.
71. Higgins JPT, Green S, editors. Cochrane handbook for systematic reviews of interventions version 5.1.0 [updated March 2011]: The Cochrane Collaboration; 2011. http://handbook.cochrane.org.
72. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

73. Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the meta-quality cross sectional study. BMJ. 2005;330(7499):1053.
74. Probst H, Bragg C, Dodwell D, Green D, Hart J. A systematic review of methods to immobilise breast tissue during adjuvant breast irradiation. Radiography. 2014;20(1):70–81.
75. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. BMJ. 1997;315(7121):1533–7.
76. Huedo-Medina TB, Sanchez-Mecca J, Bottela J, et al. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? Psychol Methods. 2006;11(2):193–206.
77. Egger M, Davey SG, Schneider M, et al. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629–34.
78. The benefits and harms of breast cancer screening: an independent review. Lancet. 2012;380(9855):1778–86.
79. Mallett S, Deeks JJ, Halligan S, et al. Systematic reviews of diagnostic tests in cancer: review of methods and reporting. BMJ. 2006;333(7565):413.
80. Felson DT. Bias in meta-analytic research. J Clin Epidemiol. 1992;45(8):885–92.
81. David M, Cook D, Eastwood S, et al. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Lancet. 1999;354:1896–900.