# How to read and appraise a research paper

Roger Jones
Editor, *British Journal of General Practice*
Emeritus Professor of General Practice, King's College London

## Introduction

Critical reading – the ability to appraise and evaluate the quality of an academic or professional article, generally a research paper – is an important skill in primary care, and critical reading abilities are required by:

- clinicians, in training and in practice, to evaluate the quality of new research and its relevance to their clinical practice
- researchers, to understand the significance of research in their field and to support their own paper-writing
- editors, who have the task of assessing the quality and trustworthiness of research papers submitted to their journal
- reviewers, who are asked by peer-reviewed journals to assess the quality of submitted manuscripts and their suitability for publication
- teachers and trainers, who will need to guide students and trainees through the medical literature
- students, who are increasingly expected to understand the elements of critical appraisal of research papers
- policy-makers and managers, who may need to know how robust the emerging evidence is for new methods of treatment and healthcare delivery.

This article is intended for general practitioners (GPs) in training and in the early stages of their careers whose responsibilities are predominantly clinical and who need to master the skills of critical appraisal to keep abreast of the literature, to inform changes in their practice and to contribute to continuing professional development and other educational activities. I have concentrated on five important types of research study:

- surveys
- randomised controlled trials
- systematic reviews and meta-analyses
- tools for diagnosis and measurement
- qualitative studies.

For each of these I have provided a citation to a paper recently published in the *British Journal of General Practice* (*BJGP*) as an example and as an opportunity to try out the guidance.

## RCGP Curriculum

The relevance of critical appraisal is reflected in three of the RCGP curriculum statements:

- 3.5 – Evidence-based Practice
- 3.6 – Research and Academic Activity
- 4.2 – Information Management and Technology.

The competencies involved are summarised in Box 1. These include an awareness of the place of research and the research literature in providing the evidence base for practice; central to many of the competencies are the skills required to read and evaluate a research paper.

**Box 1 The RCGP curriculum and critical appraisal**

RCGP curriculum statements 3.5, 3.6 and 4.2 include a number of learning objectives related to critical appraisal. In particular, GPs should be able to:

- ask the 'right questions' following a consultation or query from a patient, to enable an efficient search
- apply rigour in appraising the literature
- demonstrate the ability to search the internet for medical and scientific information, including in MEDLINE and the National Electronic Library for Health
- place the answers in the appropriate context.

In addition, these curriculum statements include the need for all GPs to:

- be familiar with essential components of the research process
- be able to develop a research question, identify appropriate methods from a range of designs, draw up a questionnaire, demonstrate basic quantitative and qualitative data analysis skills, draw appropriate conclusions and summarise results
- be familiar with prioritising relevant information, critical appraisal, problem framing, accessing evidence, implementing change in clinical practice, basic statistics, evaluating ethical issues and the need to have projects approved through research governance committees.

They also remind trainees that 'a great deal of research is conducted in secondary care settings; the results are not necessarily applicable in general practice. All GPs must, therefore, be able to judge relevance, applicability and validity of research findings to their own practice'.

They add 'The complexity of undertaking research or implementing research findings should not be underestimated. GPs should use the same holistic approach to such scholarly activity as they would in clinical practice'.

# Approaching the literature

## Be realistic

The volume of medical research literature is enormous and is presented and discussed in increasingly diverse formats, including print and online journals, automatically generated tables of contents and other e-mail alerts, and the blogosphere and other social media. It is easy to feel overwhelmed and to fear drowning. The answer is to be selective and not to feel guilty – decide on what you want to read and how and when you want to read it and, without wishing to undermine my own arguments, keep in mind the fact that a single paper is unlikely to change practice. If something is really going to revolutionise the way that you diagnose or treat a certain condition or organise your practice, the relevant new findings are likely to have been described and confirmed in a number of publications, possibly subjected to meta-analysis and more likely than not summarised in an editorial somewhere.

## Be selective

The chances are that you will receive or have ready access to the *BJGP* and the *BMJ* (*British Medical Journal*) and your practice or colleagues or even family will receive one or two specialist journals related to their areas of interest, along with the GP newspapers and, of course, *InnovAiT.* My advice is to scan and be selective and not to feel oppressed by the need to read everything – see whether there is anything that appeals on first glance, or that relates to

something that has happened in the surgery or is going on in the practice. You might recognise the authors or the institution involved, have a special interest in a particular clinical topic or be looking out for ways of developing an aspect of the services you provide in the practice. Both the *BJGP* and the *BMJ* print one- or two-page summaries of their research papers, with the full paper that includes the references, tables and figures being published online.

Your next step, which is to read the short version (and, if you get interested, move on to the full paper), will frequently be helped by an accompanying editorial. These editorials, which are often a mini-review of the paper, provide an explanation of its significance and implications. Almost every research paper in the *New England Journal of Medicine* has an accompanying editorial, and many of the *BMJ*'s papers do as well. Do not forget that although you may be most interested in reading about research carried out in primary care, other studies conducted in other settings, and meta-analyses of series of papers reporting a variety of studies, may also contain useful material for your work in general practice.

## Peer review

The papers you will read in the major journals will have undergone a fairly rigorous process of peer review, in which two or three reviewers, often including a statistician, will have provided detailed comments for the journal editor to help him or her make a decision about publication and to feed back to the authors. The paper you read will almost certainly have undergone substantial revision since it was originally submitted for publication and it will also have been copy-edited to ensure that the text reads well and conforms with publishing conventions. Publication, however, is still no guarantee of quality or of relevance!

## Generic quality criteria

A few general themes recur in the critical appraisal of a research paper and need to be considered before going on to determine what sort of paper it is and what sort of research it is reporting and to apply a more specific mental or physical checklist to it as you read through. The most important criteria for this initial appraisal, most of which should be satisfied by any research paper, are listed in Box 2.

**Box 2 Critical appraisal criteria**

- Does the paper describe the background to the study and ask a clear research question?
- Are the aims of the study clearly stated?
- Is the methods section sufficiently clear and detailed to allow the research to be repeated by others?
- Are the results clearly presented with good use of appropriate graphics and statistical tests?
- Are the sampling and recruitment methods and inclusion/exclusion criteria clearly stated?
- Are the results relevant to your own practice population/practice setting?
- Is the comparison with existing literature adequate?
- Are the strengths and weaknesses of the study candidly and fully described?
- Is the referencing adequate, with inclusion of relevant previous work and other sources?
- Are potential conflicts of interest stated by the authors?
- Is the funding source identified?
- Is there a statement of ethics committee approval?

In the following sections we will look at the various kinds of research paper you are likely to encounter, and the key criteria that you should have in mind to decide on how trustworthy and useful the results of the study and the conclusions and implications drawn from them are.

Section 1

# Mapping the territory: descriptive studies

Aziz Sheikh
Professor of Primary Care Research & Development and Co-Director, Centre for Population Health Sciences, The University of Edinburgh

Relevant *BJGP* papers:

- Mathur R, Hull SA, Badrick E, *et al*. Cardiovascular multimorbidity: the effect of ethnicity on prevalence and risk factor management. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X572454
- A'Court C, Stevens, R, Sanders S, *et al*. Type and accuracy of sphygmomanometers in primary care: a cross-sectional observational study. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X593884
- Cornford CS, Mason JM, Inns F. Deep vein thromboses in users of opioid drugs: incidence, prevalence, and risk factors. Br J Gen Pract 2011; DOI: 10.3399/bjgp11X613115

## Introduction

Descriptive studies are widely employed in primary care research to answer any of a number of epidemiological, public health and health services research questions, as reflected by the titles of the three papers selected for inclusion in this section. While these studies have historically tended to use survey techniques for data gathering,[1] the considerable proliferation of large-scale repositories of routine healthcare data has meant that descriptive enquiries now increasingly involve secondary analyses of existing datasets (see Section 2).[2] As these datasets continue to mature, and the means and opportunities to link health and other datasets increase, interrogation of routine data is likely to be much more widely employed to undertake descriptive enquiries.[3]

Irrespective of whether surveys or secondary analyses have been undertaken, when critically reviewing such papers I try to ask three key overarching questions, namely:

1. Were important questions asked?
2. Were the methods appropriate and are the results thus likely to be credible?
3. Have the findings from this work been critically reflected on in light of the relative strengths and limitations of the approach employed and the wider body of published evidence?

If the answer to these three questions is 'yes', my aim, when participating in peer review for a journal, is then to try to offer constructive suggestions on how the description of the research and its interpretation can be improved, and to offer the editor reflections on whether the paper is likely to be of interest to the journal's readership.[4]

## Critically appraising descriptive studies

I will consider each of these three questions in turn, focusing in particular on the paper by Mathur *et al*. but also making reference, where appropriate, to the papers by A'Court *et al*. and Cornford *et al*.

## 1. Were important questions asked?

- The UK is an increasingly ethnically diverse society (as indeed are most economically developed and transition countries) but also one with considerable ethnicity-related health variations in disease incidence, prevalence and outcomes.[5] Most of the evidence with respect to these ethnic variations relates to individual long-term conditions, such as cardiovascular disease and asthma. However, given that the majority of adult patients have more than one long-term condition, the authors' decision to focus on cardiovascular multimorbidity is both timely and appropriate. The relevance of this work was heightened by the fact that this question was asked in the context of one of the most ethnically diverse and socio-economically disadvantaged populations in the UK (Tower Hamlets, the City, Hackney and Newham in London). I believe that the study is thus important from both an epidemiological and a public health perspective.
- The study by A'Court *et al*. focused on important health services research questions that are of widespread day-to-day relevance to GPs across the UK and indeed internationally.
- While it is perhaps of more specialist interest, the study by Cornford *et al*. also sought to answer a relevant series of epidemiological questions.
- Overall, therefore, all three studies in this section asked clinically relevant epidemiological, public health or health services research questions. Had this not been the case, there would have been little merit in continuing with the critical appraisal of these studies.

## 2. Were the methods appropriate and are the results thus likely to be credible?

- When reviewing the methods of descriptive studies, it is important to assess both internal and external validity. Internal validity always takes precedence: this assessment should focus on considering the role of bias and chance and, in the context of analytical studies attempting to assess causality, confounding and effect modification.[6]
- The study by Mathur *et al*. was a secondary analysis of a large regional database of routinely collected primary care records. Using such data offers considerable advantages in terms of sample size (and hence precision) and substantial cost savings when compared with those incurred in the context of primary data collection. Data quality is, however, a major concern when interrogating routine data sources, although these can often be addressed in expert hands by, for example, triangulating data sources and building in reliability and validity checks and sensitivity analyses. Missing data can also prove to be a major problem, particularly with regard to ethnicity information, which historically has been very poorly recorded. Key strengths of the dataset used included the fact that it covered an area with large numbers of the minority ethnic population of interest and the largely complete recording of ethnicity in this dataset. It was also encouraging that the clustered nature of the data was considered in the data analysis, as this can otherwise result in spurious precision. Both the internal and external validity of this work were, in my judgement, likely to be high.
- A'Court *et al*. conducted a small, regional cross-sectional study that involved trained technicians visiting practices to assess the accuracy of sphygmomanometers. Importantly, the team used a standard protocol to assess these instruments, which should have helped to minimise the risk of bias. I do, however, always struggle with statistical testing in the absence of clearly detailed hypotheses; furthermore, there were no formal sample size calculations assessing whether the study was adequately powered to reliably detect important differences between groups. I also find it much more informative to be presented with 95% confidence intervals rather than *P*-values. As with many such primary studies, the response rate was disappointing at only 46%, which, together with the regional nature of the study, raises important questions about the external validity of the findings from this work.

General Practice Research Database or GPRD) to identify risk factors for death among patients with epilepsy. None of these papers explicitly states a 'research question', preferring instead language such as 'examine trends' and 'identify risk factors'. However, for critical appraisal, the relevant question is that of interest to the reader – 'What motivated you to read this study?' We propose that the reader of the Hippisley-Cox and Coupland paper may ask 'Can a risk score for lung cancer improve my detection of lung cancer?'; of the Lockhart and Guthrie paper, 'Which antidepressants are being increasingly prescribed, and why?'; and of the Ridsdale paper, 'Which patients with epilepsy are at greatest mortality risk?'

# Validity of methods

## Population studied (P)

In database research, considerations about recruitment are largely superseded by considerations about selection, from a general database, of the subset for analysis. These three papers use databases that sample everyone using primary care or everyone eligible to use community pharmacists in a given region: few in the UK are ineligible by these criteria.

To study people with epilepsy, Ridsdale *et al*. identified people who have both diagnostic codes for epilepsy and prescription codes for anticonvulsant drugs: the additional requirement of a relevant prescription was presumably to ensure a highly specific definition of the cohort of interest, guarding against, for example, data entry errors in the diagnostic field. For other conditions it may be appropriate to combine diagnostic codes and prescriptions in another way, for example by using prescriptions for statin medication as well as diagnostic codes to identify people with high cholesterol.[1]

The Lockhart and Guthrie paper illustrates another consideration: identifying the denominator population. The database is everyone using a community pharmacist but the denominator of interest is everyone *eligible* to use a community pharmacist; Lockhart and Guthrie overcome this by incorporating external population estimates from the General Registrar of Scotland. The Hippisley-Cox and Coupland paper also has a potential problem with the denominator, since 'patients registered with practices' may include some who have moved away but not yet notified their practice. Other database analysts may impose criteria such as 'at least one consultation during [some time window]', to reduce this problem at the expense of a slight bias against the healthiest patients.

## Risk factors recorded (R)

In any critical appraisal, the reliability of measurements should be considered. Some things, such as prescriptions, are likely to be recorded with high completeness and accuracy, whereas others, such as indication for treatment, may have to be inferred from other factors such as the recorded reason for the consultation. Authors may help the reader (see, for example, reference 32 in the Hippisley-Cox and Coupland paper)[2] by citing relevant papers from the growing literature on recording accuracy in database research.[3,4]

The methods section of the Ridsdale *et al*. paper describes how epilepsy was 'identified using a list of 186 Read and Oxford Medical Information Systems Codes', rather than by a single tick-box as might be the case in a mainstream observational study. Database researchers invest considerable work in developing such code lists. For the reader appraising a paper without personal experience of database research, it would be useful to consider whether secondary analyses had been conducted on the sensitivity of results to the code lists.

### Outcomes assessed (O)

Lockhart and Guthrie's outcome, antidepressant prescribing, could equally be a risk factor in another study,[5] illustrating that for many outcomes the considerations are the same as for risk factors.

The Hippisley-Cox and Coupland paper illustrates another common feature of primary care database research: lung cancer cases were identified not only through GP records (within the database) but also through death certificates (from an external, linked data source). The authors proposed to extend this in future studies, with further linkage to the National Cancer Intelligence Network. Conversely, other studies of cancer in primary care databases have relied solely on GP records.[3] The reader could refer again to the literature recording validity in widely used databases[3,4] and consider how incomplete or invalid recording might affect results. Under-ascertainment may affect relative risks less than absolute event rates, unless it is in some way differential by exposure. Finally, the usual concerns about outcome ascertainment in observational research must be considered, such as whether there was sufficient length of follow-up, and the nature and extent of individual loss to follow-up.

### Database analysis (D)

Once the population, risk factors and outcomes have been identified and coded, researchers often arrive at a dataset that can be treated as any other observational study: for example, by proportional hazards modelling, as in the Cox regression model of Hippisley-Cox and Coupland, or by the nested case–control approach as used by Ridsdale *et al*.

Some variables may be missing more often in database research than in prospectively designed studies: for example, weight and hence body mass index may be widely unmeasured in primary care databases.[6] Full discussion of missing data handling methods is beyond the scope of this article but we note that there is increasing consensus about the use of multiple imputation, as used by Hippisley-Cox and Coupland. The assumptions on which this is based have been discussed elsewhere[7,8] and an argument made that other methods such as 'complete case analysis' require even stronger assumptions.[9]

Appraisal of relative risks or odds ratios in database studies must consider 'immortal time bias'. This subtle methodological error can have dramatic effects on results. It most often occurs in database studies when information *after* index date is used to define risk factor status *at* index date.[10,11] As is common in nested case–control analysis, the study by Ridsdale *et al*. avoided this bias by defining risk factors strictly in the period before index date. The Hippisley-Cox and Coupland study used risk factor information after entry to cohort but avoided the bias by correct use of time-dependent covariates: see the article by Lévesque *et al*.[12] for an explanation, together with four relevant bullet points to assist identification of this bias.

## What are the results?

The remark 'given the size of the dataset, virtually any comparison is likely to be statistically significant' in the paper by Lockhart and Guthrie applies widely in database research. It is illustrated by Table 6 of that paper, in which all comparisons but one are highly statistically significant, but it remains a matter for discussion by the authors and for judgement by the reader as to whether all the changes are clinically important.

## Do the results apply locally?

The populations in database research can be highly applicable to local practice – within the limitations of the additional selection criteria applied (see 'P' above).

The price paid for the large scale of these databases is that items were, in general, not recorded with research in mind, leading to the considerations listed above (especially under 'R', 'O' and 'D') regarding validity. However, another benefit other than scale is that the data collected in routine practice is by definition highly relevant to routine practice (subject to the caveats under 'P' above). The databases in these UK studies are comprehensive samples of the population they represent. A subjective judgement must be made about the relevance of the study population to local practice, especially across boundaries of countries and healthcare systems, but database research has the potential to achieve a very high relevance to local practice.

## Conclusion

This article cannot be a fully comprehensive guide to the many types of database study and their strengths and weaknesses. For example, we have not had space to address the challenges of cross-database linkage, and it may be too early to consider where database research belongs within 'hierarchies' of evidence for medicine.[13] Here, we have tried to use our recent experience as researchers moving from mainstream epidemiology into database research to guide the reader in critically appraising such studies. Database research is likely to expand, using the large number of cases and person-years available relatively cheaply to test existing hypotheses, generate new hypotheses and in some cases address previously unanswerable questions.

## References


1. Walters K, Rait G, Petersen I, *et al*. Panic disorder and risk of new onset coronary heart disease, acute myocardial infarction, and cardiac mortality: Cohort study using the general practice research database. *Eur Heart J* 2008; **29(24):** 2981–2988.
2. Jick H, Jick SS, Derby LE. Validation of information recorded on general practitioner based computerised data resource in the United Kingdom. *BMJ* 1991; **302(6779):** 766–768.
3. Khan NF, Carpenter L, Watson E, *et al*. Cancer screening and preventative care among long-term cancer survivors in the United Kingdom. *Br J Cancer* 2010; **102(7):** 1085–1090.
4. Herrett E, Thomas SL, Schoonen WM, *et al*. Validation and validity of diagnoses in the general practice research database: a systematic review. *Br J Clin Pharmacol* 2010; **69(1):** 4–14.
5. Walker AJ, Card T, Bates TE, *et al*. Tricyclic antidepressants and the incidence of certain cancers: a study using the GPRD. *Br J Cancer* 2011; **104(1):** 193–197.
6. Sutton M, Elder R, Guthrie B, *et al*. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. *Health Econ* 2010; **19(1):** 1–13.
7. Barzi F, Woodward M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *Am J Epidemiol* 2004; **160(1):** 34–45.
8. Marshall A, Altman DG, Royston P, *et al*. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 2010; **10:** 7.
9. Stuart EA, Azur M, Frangakis C, *et al*. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am J Epidemiol* 2009 ; **169(9):** 1133–1139.
10. Suissa S. Immortal time bias in pharmacoepidemiology. *Am J Epidemiol* 2008; **167(4):** 492–499.
11. Sylvestre MP, Huszti E, Hanley JA. Do Oscar winners live longer than less successful peers? A reanalysis of the evidence. *Ann Intern Med* 2006; **145(5):** 361–363.
12. Lévesque LE, Hanley JA, Kezouh A, *et al*. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ (Online)* 2010; **340(7752):** 907–911.
13. Atkins D, Eccles M, Flottorp S, *et al*. Systems for grading the quality of evidence and the strength of recommendations I: Critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res* 2004; **4(1):** 38.

A useful checklist of things that should be included in a trial report has been produced by the Consolidated Standards of Reporting Trials (CONSORT) Group.[6] Many journals require authors to complete a CONSORT checklist as part of their submission. It should be noted that CONSORT provides a standard for *reporting*, rather than for the *conduct* of a trial; the quality of reporting does not always reflect the quality of the conduct.[7]

The International Committee of Medical Journal Editors (ICMJE) has also compiled guidelines for journal submissions, known as the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*.[8] ICMJE member journals agree only to publish results of trials that were registered in a public trials registry. Trial registration is one way to ensure that a trial protocol specifying details of the trial conduct is published *before* any participants are enrolled.

> Llor *et al*.: '*The trial was approved by the Ethical and Clinical Research Committee of the Jordi Gol Institute of Research in Primary Health Care (Certificate number: P06/03) … clinical trial registration trial number ISRCTN23587778*' [accessible at http:/www.controlled-trials.com/isrctn/pf/23587778]

## What is being compared with what?

> From the *Uniform Requirements for Manuscripts Submitted to Biomedical Journals*:[8] '*The ICMJE defines a clinical trial as any research project that prospectively assigns human subjects to intervention or concurrent comparison or control groups to study the cause-and-effect relationship between a medical intervention and a health outcome.*'

Authors should explain what the comparison or control was – this is usually:

1. an established treatment or 'routine care', or
2. a placebo, sham, or inactive control.

They should also clarify what question the trial was trying to answer. Often this will be 'is the active intervention *better* than the control?' (a superiority trial) but, depending on the circumstances, it could be 'is the active intervention *no worse* than the control?' (a non-inferiority trial), or even 'does the active intervention have *the same* effect as the control?' (an equivalence trial).

Some trials are intended to show that the intervention *can* work under tightly controlled, ideal conditions (explanatory trials, or trials of efficacy); some are intended to show that it *does* work in routine use (pragmatic trials, or trials of effectiveness). Trials in primary care are often pragmatic but it is helpful to make the intention clear, as it influences who should be eligible for the trial, who should deliver the intervention, what the control should be, and more. A useful guide to distinguishing explanatory from pragmatic trials is provided by the Pragmatic-Explanatory Continuum Indicator Summary, or PRECIS tool.[9]

> Paterson *et al*.: '*The study aimed to answer the question "In patients who attend frequently in primary care with MUPS* [medically unexplained physical symptoms] *that have persisted for more than 3 months, does the addition of classical five-element acupuncture to usual GP care, compared to usual care alone, improve self-reported health and reduce conventional medication and general practice consultation rates?".*'

The active and control interventions should be described in enough detail to allow suitably qualified people to replicate them. In the case of a drug treatment, it may be enough to give the name of the drug, route of administration, dose, timing and duration. Many interventions trialled in general practice are, however, non-pharmacological.[10] These sometimes incorporate

a number of components, targeted at more than one level – for example, at the level of the GP and at the level of the patient. Understandably, such interventions are described as complex.[11] It can be a challenge to describe a complex intervention in enough detail to permit its replication: the description should include the setting, mode, intensity and duration of each intervention component, and information about who delivered it.[12]

> Paterson *et al*.: '*Individualised classical five-element acupuncture was delivered in the GP surgeries by eight five-element acupuncture practitioners who were members of the British Acupuncture Council. Twelve sessions, on average 60 minutes in length, were provided over a 6-month period at approximately weekly, then fortnightly and monthly intervals. These timings were adjusted to individual patients' needs.*'

## Randomisation

> Paterson *et al*. – Reviewer 2: '*You state that you used block randomisation and then in the next sentence stated you used minimisation. I am unclear how both of these can be the case.*'

Randomisation is, of course, a key component of a randomised controlled trial, but the terminology and practice of randomisation can be confusing. Simple randomisation involves making a random choice of group allocation for each new recruit into the study. Block randomisation means that allocations in a consecutive sequence of, say, eight new recruits are constrained so that exactly half are to the active intervention and half to the control. Block randomisation helps to ensure roughly equal group sizes. Block randomisation is often stratified – that is, done separately in different strata, such as in men and in women. Stratified randomisation helps ensure the intervention groups are balanced with respect to the stratifying variables, although it will not work if individual strata are small compared with the block size. A more general approach to achieving balance is minimisation, which looks at a number of characteristics of each new recruit and makes an allocation that minimises the overall imbalance between the intervention groups. Minimisation might involve no randomness at all (other than in the first allocation), but a random element is usually incorporated to make allocations harder to predict – indeed, a random element is required under some regulatory frameworks.

Cluster randomisation is where each group allocation is applied to an entire cluster of individuals. This might employ simple, block or stratified randomisation, or minimisation. Cluster randomisation is used where the intervention is targeted at a higher level than an individual participant (for example, an intervention aimed at the participants' GP), or to prevent benefits of the active intervention being shared or distributed within a cluster that also includes control participants (contamination). The CONSORT extension to cluster-randomised trials suggests that the title or abstract of the manuscript should specify that randomisation was in clusters.[13]

Whatever method is used, the manuscript should describe how allocations were generated and how they were distributed to recruiters. Traditionally, recruiters were given a series of numbered, opaque, sealed envelopes containing the allocations, but there is some concern about the integrity of such methods,[14] and more secure electronic methods and randomisation services now tend to be favoured.

> Paterson *et al*.: '*Group allocation was undertaken by the Institute of Psychiatry Clinical Trials Unit, using their web-based service.*'

## Blinding

> Llor *et al*. – Reviewer 2: '*Could one use a placebo RADT* [rapid antigen detection test] *kit to reduce bias? This might help blind GPs as well as patients.*'
>
> Paterson *et al*. – Reviewer 2: '*Blinding is clearly difficult in this situation. However, there is the potential for a sham intervention and this could be considered in more detail.*'
>
> Paterson *et al*. – Reviewer 2: '*The abstract suggests a blinded intention-to-treat analysis adjusted for baseline outcomes has been undertaken. … I also feel that using blinding here is misleading for the reader as only the statistician was blinded.*'

Blinding refers to steps taken to conceal group allocations. Other terminology, such as 'masking', may be used. A number of key players in the trial could potentially be blinded:

- senior investigators
- those who interact with the participants or assess outcomes
- statisticians who analyse the data.

There are likely to be different practical challenges to blinding different groups of people, and failure to blind different sorts of people can lead to different kinds of bias. Trials are sometimes referred to as 'single blind' if participants do not know which intervention group they are in but investigators do, and 'double blind' if participants and investigators alike are blinded. Given the number of distinct players in a trial, however, these phrases are ambiguous.[14] It is best for the manuscript to be explicit about which groups of people were blinded. In some situations it will be impossible to blind some groups; nevertheless, blinding should be as complete as possible, and blinded roles should be distinguished in the manuscript in as much detail as possible.

## Accounting for all the participants

> Paterson *et al*. – Reviewer 2: '*I found the CONSORT flowchart a little difficult to follow. At some points it took me a bit to work out how many we had left!*'

Triallists have an ethical responsibility to plan the size of their clinical trials,[15] and a report of a trial should include a justification of the sample size. This is usually expressed in terms of the statistical power achieved (that is, the chance of finding evidence for an effect of the intervention, if a clinically important effect is present). Conventional targets for power are 80% or 90%.

The statement of power should include all the information necessary to replicate it,[16,17] including the minimal clinically important effect that was assumed. Crucially, the calculation of power should have been included in the original protocol, and should be based on the primary outcome measure as advertised in the manuscript.

If follow-up is likely to be incomplete, the sample size calculation should include an allowance for drop-outs. (Incidentally, this is frequently done incorrectly: if you want to analyse data from 100 people and you anticipate a 20% drop-out rate then you need to recruit $100/0.8 = 125$ people, not $100 \times 1.2 = 120$ people.)

The Results section should include a flow diagram recording numbers of people who were enrolled, randomised, allocated to different treatments, followed up, and analysed. This diagram is often called a CONSORT flowchart because it is one of the requirements on the CONSORT checklist.[6] It is vital for readers to be able to understand where and why enrolled participants were lost to the final analysis. Readers may be suspicious of bias if a large proportion of participants are lost, or if this proportion differs noticeably between treatment groups.

There remains a question of what to do with people who were followed up but who had not complied with their treatment. Investigators should explain their strategy, which often comes down to doing either an intention-to-treat analysis (in which all participants who were followed up are analysed according to the treatment to which they were allocated, irrespective of whether they complied) or a per-protocol analysis (in which only those participants who complied are analysed). Per-protocol analysis may be useful in an efficacy trial, but for pragmatic trials intention-to-treat is often preferred as it addresses effectiveness in practice. More sophisticated alternatives to per-protocol analysis, taking better account of non-compliance, are available.[18] As with blinding, it is best for authors to be explicit about which participants are included and how their data are used, rather than to rely on general terms.

An issue often confused with intention-to-treat is the question of how to cope with missing data, for example as a result of loss to follow-up. In many situations, under reasonable assumptions, unbiased results can be obtained in a trial by analysing only the non-missing outcomes, adjusting for participants' baseline characteristics if necessary. 'Intention-to-treat' does not mean that data must have been collected for everyone (although it does mean that investigators should have tried).[19] Researchers will, however, sometimes use a variety of methods to impute – or fill in – missing outcome data. Methods such as carrying forward the outcome observed on an earlier occasion are still common but should be discouraged, as they lead to bias. Better computational methods for imputation, and general strategies for coping with missing data, are available.[19]

## Other issues

In this summary I have limited myself to issues that are specific to clinical trials, although these will form only a fraction of the points you may want to raise in a review of a particular trial report.

Concerns about clarity, precision and consistency are common to manuscripts of all kind. In quantitative research, reviewers may have to accept that analyses are impossible to confirm without access to the data; but they can and should check that when the same number appears in different parts of the manuscript it is quoted consistently, and that numerical results are sensible – for example, that estimates of effect size are within their quoted confidence intervals. Guidance on how quantitative results should be presented can be found in the ICMJE's *Uniform Requirements*.[8]

Finally, note that even if all methods are correct and reporting guidelines followed to the letter, it counts for nothing if the trial answers a different question to the one that was initially posed, or claims to answer a question it cannot answer.

## References

1. Shuster E. Fifty years later: the significance of the Nuremberg Code. *N Engl J Med* 1997; **337(20):** 1436–1440.
2. World Medical Association. *Declaration of Helsinki*. http://www.wma.net/en/20activities/10ethics/10helsinki [accessed 16 Nov 2012].
3. European Parliament and Council. *Clinical Trials Directive*. http://www.cardiff.ac.uk/racdv/resgov/Resources/CT%20Directive%202001.20.ec.pdf [accessed 16 Nov 2012].
4. HM Government UK. *The Medicines for Human Use (Clinical Trials) Regulations 2004*. http://www.legislation.gov.uk/uksi/2004/1031/contents/made [accessed 16 Nov 2012].
5. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH). *ICH Guidelines*. http://www.ich.org/products/guidelines [accessed 16 Nov 2012].

Section 4

# Measuring health and illness: development and validation of tools

Kevin Barraclough* and William Hamilton†
*General Practitioner, Hoyland House Surgery, Painswick, Gloucestershire
†Professor of Primary Care Diagnostics, University of Exeter

Relevant *BJGP* papers:

- McCowan C, Donnan PT, Dewar J, *et al*. Identifying suspected breast cancer: development and validation of a clinical prediction rule. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X572391
- Cameron IM, Cardy A, Crawford JR, *et al*. Measuring depression severity in general practice: discriminatory performance of the PHQ-9, HADS-D, and BDI-II. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X583209
- Smith LF. Postnatal care: development of a psychometric multidimensional satisfaction questionnaire (theWOMBPNSQ) to assess women's views. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X601334

Doctors have traditionally been taught the clinical skills involved in diagnosis and illness management on the basis of overall clinical impression and anecdotal 'rules'. We may thus well have been taught that any discrete breast lump in a woman over the age of 45 needs to be biopsied to exclude cancer. However, such 'rules' are relatively crude and ignore the subtleties of medicine. In particular, they are usually binary, in that certain symptoms require investigation whereas others do not. Patient perspectives rarely enter into this decision making.

In the last couple of decades, researchers have started developing far more complex tools, particularly for diagnosis but also for quantifying severity, prognosis, or response to treatment. These methods integrate several separate features (clinical variables of some sort), which are weighted according to evidence, to come up with a single parameter that measures either the probability of disease or the severity of disease. Irrespective of the apparent complexity of the statistical procedures, we understand enough if we merely accept that these techniques take multiple variables and integrate them to get a single parameter or score. These tools are more complex to use than clinical intuition but they have the potential to improve the accuracy of diagnosis and the assessment of severity or treatment response significantly.

The development of a particular rule or score will seem like an arcane statistical art to the clinician, but the essentials of being able to assess the tool, or of assessing the paper that puts forward the derivation and validation of the rule, are less complex than they at first appear.

For the practising clinician, it is necessary to hold on hard to the arms of the chair and keep chanting the mantra that you are a practising clinician and not a statistician. However, in that higher state of mind it is possible to appraise critically this type of paper using basic clinical common sense and a minimum of statistical knowledge.

However, it would seem clinically odd if a CPR for breast cancer or an acute coronary syndrome did not include age.

The methodology of the paper may be faultless but if the study does not include data on a key clinical parameter that most doctors would take account of instinctively then the results are of little use. It is surprising how many CPRs seem to ignore age when there is a steep dependence of the condition of interest on increasing age. If one of the variables that clinical common sense would suggest is important is missing then it is necessary to check at the outset whether the variable was included in the original derivation and subsequently discarded as being non-discriminatory. If an obviously discriminatory clinical variable was merely ignored then the study is largely useless.

### 4. Could you ever see yourself using the rule or tool yourself in clinical practice?

If you picture yourself in the consulting room, would you use the rule or tool if the study suggests it is valid? It may be that you could see yourself finding the tool useful, but not in the way suggested by the paper. The '7-point score' for detecting malignant melanoma[1] may be useful as a reminder of the key clinical features but, given that 40–60% of all benign pigmented lesions assessed in practice score 3 or more points (and it is impractical to refer the majority of pigmented lesions seen in primary care), it may be that the tool is potentially useful but not in the way suggested in the paper.

However, a simple numerical score that would help to differentiate patients with symptomatic breast disease into those at high risk of breast cancer (say a 5–8% risk) and those with low risk (say 1% or less) would be considered extremely useful for most GPs. (Of course, this still neglects the most difficult area of the 1–5% risk group.)

## More detailed questions

If the study seeks to answer a clinical problem that is relevant to general practice, it was carried out on a relevant population and it appears to make clinical and practical sense then the paper can be examined in more detail.

### 5. Is the study new, or have other relevant studies been published?

A quick PubMed, MEDLINE or Google Scholar search will establish whether there have been other relevant studies in the field. Some clinical areas indeed have so many CPRs published that systematic reviews and meta-analyses of these have been published. Theoretically, the authors should have mentioned other CPRs on the subject, but alternative scoring systems have a sneaky way of slipping out of the authors' minds when it comes to writing their papers. A study that might otherwise be of relevance or interest will be of little additional value if an earlier, far larger study has addressed the same clinical question. For example, a small pilot study in 50 patients that looked at the predictive value of a particular physical sign would be of little interest if a major US journal had published a study of 700 similar patients 10 years previously.

It may also be that related studies have rendered the clinical question irrelevant. If a large study has shown that asymptomatic patients with carotid stenosis do not benefit from surgery then the question of whether a carotid bruit is predictive of significant carotid stenosis becomes clinically irrelevant (at least from a therapeutic view point; it retains prognostic significance, of course). The clinical question has been superseded.

## 6. Precisely how was the study population recruited and how complete was the recruitment?

If the study population was derived from several practices, do the numbers suggest that recruitment was complete? If the study population covered 10 large practices for a year but only 70 patients with back pain were recruited then it is clear that recruitment was incomplete. Is the recruited study population biased in some way? By selection bias we mean that, if we could identify those patients who presented with back pain but were not recruited into the study, would they differ from the study population in important characteristics such as symptom severity, age, socio-economic class, and so on? If the 70 patients recruited into the study were only those with more severe pain, or those in whom the GPs were anxious about underlying disease, then the cohort is not representative of the whole population of interest. The sample is biased and the conclusions cannot safely be extrapolated into clinical practice.

Does the paper make any attempt to identify the characteristics of the population on whom data were not initially obtained? If those who agreed to participate in the study were all in professional classes and non-smokers then the non-participants probably differ significantly from those in the study, and the study population is again biased. Some studies make efforts to identify the characteristics of non-participants.

In a paper that is both deriving and validating a clinical prediction rule (such as the paper by McGowan *et al*.) the study will involve recruiting two populations (two 'cohorts'). The first population will be used to derive the rule (the 'derivation cohort') and the second will be used to test the rule prospectively (the 'validation cohort'). Clearly, if the rule is ever to be used clinically, the circumstances under which the cohorts were recruited have to be as close as possible to the clinical circumstances under which the rule would be used.

## 7. Were the data collected in a way that was easily understandable and reproducible, and was the data gatherer blinded to the outcome?

For example, there would clearly be scope for subconscious bias if the individual collecting the data already knew whether the patient did or did not have disease. The data have to have been collected in a way that could be easily reproduced. If collecting some of the clinical data requires clinical judgement (for example, 'Is another diagnosis more or less likely than DVT?') then this weakens the reproducibility and objectivity of the study.

## 8. How were the data analysed?

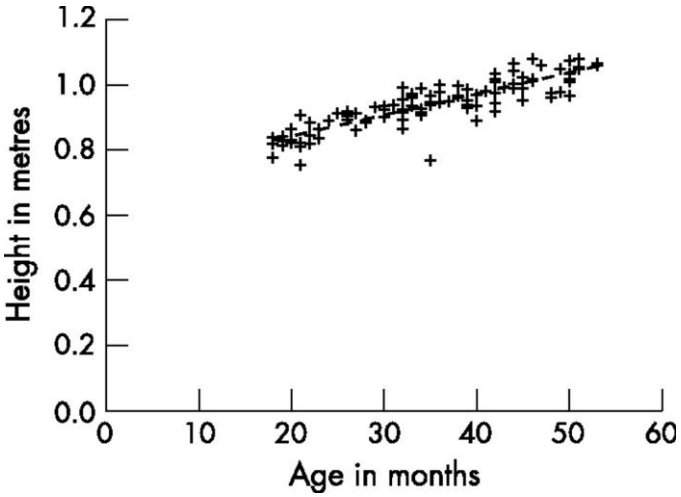
This is the stage at which it does become necessary to have a rough idea of the statistical methods employed. However, it is not necessary to have a detailed understanding and it is as important to maintain a common sense distance from the data and 'eyeball' it.

If, for example, the paper is deriving and/or validating a clinical tool (as in the paper by McGowan *et al*.) then the key process will involve the statistical technique called regression. This comes in two main types: linear and logistic regression.

### *Linear regression*

Most of us will be vaguely familiar with linear regression when a single continuous outcome variable of interest (say height) relates to a second explanatory variable (say age). We are attempting to find out how much the variable we are interested in (height) varies when the 'explanatory' variable changes (the explanatory variable 'explains' – to some extent – the variation in the outcome we are interested in).

If the paper is carrying out regression analysis on just two continuous variables (say height and age) then it should really show a 'scatter diagram' for the data, as in the example in the figure below of height against age in young children.




If there is a true linear correlation then, without any mathematics, it should be clear that one could draw a 'best fit' line through the data points. However, if the linear correlation is not clear to the eye then really no fancy mathematics should be carried out on the data. The graphical message that the correlation is poor or absent should be respected.

However, if there does appear to be a line correlating the data then a statistical software package can produce the 'best' line through the data points (using the process of minimising the square of the vertical distances of the points from the line) and come up with a '$y = ax + b$' straight line. The spread of the data points from the line will be measured by the $R^2$ statistic, which is 1 for perfect correlation (all the variation in the outcome data is 'explained' by the explanatory variable) and zero if there is no line or pattern at all.

### *Logistic regression*

More often, the outcome of interest is binary (yes/no, for example 'has cancer' or 'does not have cancer') and the outcome variable (cancer or no cancer) is plotted as a logarithm of the (probability of disease/the probability of no disease) against the explanatory variable. This is called logistic regression ('logistic' because of using logs). Logistic regression can also accommodate binary explanatory variables (breast lump/no breast lump), ordered variables (age 40–49, 50–59, 60–69, etc.) and continuous variables (such as body mass index).

If there are two explanatory variables (say age and daily calorie intake) then the same process can be imagined using a three-dimensional graph with two axes for the two explanatory variables (though the two axes would only be perpendicular to each other if the two explanatory variables were truly independent of each other). If there are multiple explanatory variables then the same tools are used (multivariable regression) but the process cannot be visualised (as we are confined in a three-dimensional world). (A complexity for the statistical software is whether the multiple variables are truly independent or not. Age and weight would not be independent variables because weight goes up with age, but age and sex would be independent variables.)

### *Odds ratios*

From our point of view we merely remember that the correlation of each explanatory variable with the outcome variable is calculated with the other variables held fixed. Most papers (for

example, the PISAPED study)[2] give the correlation (regression) coefficient for each explanatory variable, and most convert the log odds ratios into normal odds ratios and present these. If the coefficient is close to 1, the explanatory variable is playing very little part in determining the outcome variable (and may well be insignificant). The further the odds ratio (OR) is from 1, the larger the contribution that the explanatory variable makes.

If, for example (as in Table 1 in the paper by McGowan *et al*.), there are 122 people in the age range 50–59 and eight have cancer, then 114 do not have cancer. The odds of having cancer to not having cancer in that age range are 8:114 (or 0.070:1). The lowest risk age range is 25–35 and in that category the odds of having cancer versus not having cancer are 3:269 (or 0.011:1). The OR is the odds of having cancer in the higher risk age category (50–59) as compared with the odds in the lowest risk group. This is 0.070⁄0.011 = 6.29. The OR is roughly equal to relative risk. If you are aged 50–59 in this population, you are roughly 6 times as likely to have cancer than if you are in the age range 25–39.

Because the numbers studied are small there will be an uncertainty in the OR, caused by random chance (as in any study). Repeating the same study with different individuals may have given another value of the OR. This uncertainty is expressed as the 95% confidence interval. If we repeated the same study 100 times with different participants then we would expect the measured value of the OR to fall within the quoted range 95 out of 100 times. The larger the number of patients in the study, the narrower the confidence interval will be.

Much of the commonsense interpretation of a study can be derived from looking at a table of ORs such as Table 1 in the paper by McGowan *et al*. In this table, it is clear that the ORs increase steeply across age categories and the confidence intervals (from the age of 50 upwards) do not include the value 1. On the other hand, the ORs fluctuate up and down with different weight categories and are relatively close to 1. It is thus unlikely that weight has much effect on risk and the variation (from an OR of 1) is likely to be due to chance. This is also reflected in the fact the confidence intervals are wide and cross the value of 1 (for example 0.07 to 4.93). (However, other, larger, studies have shown that breast cancer risk increases with obesity.)

From the table of ORs for all the explanatory variables considered, the authors can identify those variables with high ORs that independently predict the outcome of interest (in this case, cancer). These variables are listed in Table 2 of the paper. If you are a patient with a discrete lump in this population, you are roughly 15 times more likely to have breast cancer than if you do not have a discrete lump. If the lump is larger than 2 cm, you are roughly 5 times more likely to have breast cancer than if it is smaller than 2 cm. Your risk (in fact, odds) of breast cancer goes up by a factor of 1.1 with each additional year of age.

From the ORs, the authors can pick the variables from which to construct the score for a CPR and then (approximately) weight them (the number of points associated with each feature that is awarded) and calculate roughly what percentage risk of having the condition each patient has with a certain number of points. This percentage risk can be calculated in the derivation cohort and should then be measured in the validation cohort, with the expected value versus the actual value being compared (as in Figure 1 in the paper).

If the score is going to be used as a decision rule (rather than as a guide to the probability of disease) then further studies should be done to demonstrate that in a real clinical setting clinicians do use the score and that it leads to an improvement (on some particular measure) over clinical judgement alone. This is the ‘impact analysis’ of a CPR – which is rarely done. One example where this has been done is with the Ottawa ankle rules (with measured outcomes being the number of X-rays ordered and the number of fractures missed).

### 9. Do the authors realistically assess the strengths and weaknesses of their study?

If the numbers are small and half of those who were asked to participate declined it does not mean that the study is not valuable, but such weaknesses need to be realistically discussed.

### 10. Would you agree with the authors' assessment of the significance of their study?

At this stage it is necessary to sit back and re-examine the study in the whole, and the wider clinical context from which it takes its significance. If you were writing the conclusion, would your views be similar to those of the authors about what the reader can take from this paper?

## References

1. Whited JD, Grichnik JM. The rational clinical examination. Does this patient have a mole or a melanoma? *JAMA* 1998; **279(9):** 696–701.
2. Miniati M, Monti S, Bottai M. A structured clinical model for predicting the probability of pulmonary embolism. *Am J Med* 2003; **114(3):** 173–179.

Section 5

# Bringing it all together: systematic reviews

Niek J de Wit
Professor of General Practice, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands

Relevant *BJGP* papers:

- Astin M, Griffin T, Neal RD, *et al*. The diagnostic value of symptoms for colorectal cancer in primary care: a systematic review. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X572427
- Jefferis J, Perera R, Everitt H, *et al*. Acute infective conjunctivitis in primary care: who needs antibiotics? An individual patient data meta-analysis. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X593811
- Mugunthan K, McGuire T, Glasziou P. Minimal interventions to decrease long-term use of benzodiazepines in primary care: a systematic review and meta-analysis. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X593857

## Introduction

Medical research generally aims at assessing the incidence of disease, the diagnostic value of tests or the effectiveness of a new therapeutic intervention. If the result of a research project is clear then it is relatively straightforward to implement the new knowledge in clinical practice. However, if the result is questionable, the study is often repeated by another research group. A series of such studies may lead to conflicting results that vary in direction and magnitude.

In such cases, it is useful to consider the collected evidence in a systematic review, which is a powerful instrument in evidence-based medicine that allows conclusions to be reached about the current state of the art regarding a clinical dilemma and is solidly based on the current evidence in the scientific literature.

Systematic reviews are highly relevant for clinical practice. For example, the evidence on the effectiveness of antiviral and corticosteroid therapy in patients with shingles is conflicting and, for clinicians to make evidence-based treatment decisions for patients with herpes zoster infection, they need up-to-date pooled study results. Systematic reviews, such as those provided by the Cochrane Collaboration, provide these results.

The very sensitive outcome analysis in a systematic review requires robust methodology and this means that a systematic review is a research project in its own right, and one that requires a firm scientific effort.

There are various types of systematic review, differing in their aims, data selection and outcome assessments, but in all cases the studies to be included are selected based on a detailed clinical question and using predefined eligibility criteria. In the next step, the methodology and outcomes of the studies are assessed. As long as the inclusion criteria for subjects and the methodologies used in the eligible studies do not differ too much from each

other (that is, there is only limited heterogeneity among the studies), the outcomes of the individual studies can be integrated into a sum score, a so-called pooled result. In this case, the systematic review quantifies the collected results of the individual studies into a new study outcome, and this is known as a meta-analysis. Although a meta-analysis is most commonly done with randomised controlled trials (a therapeutic meta-analysis), it can be carried out for other kinds of relevant clinical study, for example a diagnostic meta-analysis.

If the methodologies and outcomes of the studies included in the systematic review are comparable but the patient populations are not, an individual patient data (IPD) analysis can often be used to calculate a pooled result. In this case, the overall outcomes of the various studies are not used, and instead the individual patient data from all studies are entered into a new dataset, leading to a new pooled outcome. In contrast to the outcome of a meta-analysis of individual studies, that of an IPD is the summed result of all individual patients participating in the various studies. The advantage of the IPD method is that it makes the overall result less dependent on differences between the study populations. However, an IPD analysis is a major undertaking as it requires collecting all the original research data from all studies involved.

The validity of the outcome of systematic review/meta-analysis depends on the quality of the review process. For 'state-of-the-art' conduct of a systematic review, the following steps are required:

- defining the detailed research question for the review
- defining the eligibility criteria for the studies to be included
- specifying the search terms for the literature search strategy
- searching the literature, selecting the papers and excluding those not meeting the eligibility criteria
- quality assessment of the studies identified
- data extraction from the selected studies
- synthesis of results.

**Example of a diagnostic meta-analysis (Astin *et al*.)**

The authors conducted a systematic review of the diagnostic value of symptoms associated with colorectal cancer. They searched MEDLINE, Embase, the Cochrane Library and CINAHL for diagnostic studies of symptomatic adult patients in primary care. Studies of asymptomatic patients, screening, referred populations or patients with colorectal cancer recurrences, or with fewer than 100 participants, were excluded. The target condition was colorectal cancer. The data were extracted to estimate the diagnostic performance of each symptom or pair of symptoms. The data were pooled in a meta-analysis. The quality of studies was assessed with the QUADAS tool.[1]

Twenty-three studies were included. Positive predictive values (PPVs) for rectal bleeding from 13 studies ranged from 2.2% to 16%, with a pooled estimate of 8.1% (95% confidence interval [CI] = 6.0% to 11%) in those aged 50 years and older. The pooled PPV estimate for abdominal pain (three studies) was 3.3% (95% CI = 0.7% to 16%) and for anaemia (four studies) it was 9.7% (95% CI = 3.5% to 27%). For rectal bleeding accompanied by weight loss or by change in bowel habit, the pooled positive likelihood ratios (PLRs) were 1.9 (95% CI = 1.3 to 2.8) and 1.8 (95% CI = 1.3 to 2.5), respectively, suggesting higher risk when both symptoms were present. Conversely, the PLR was 1 or less for rectal bleeding accompanied by abdominal pain, by diarrhoea or by constipation.

The authors concluded that investigation of rectal bleeding or anaemia in primary care patients is warranted, irrespective of whether other symptoms are present. The risks from other single symptoms are lower, though multiple symptoms also warrant investigation.

For clinicians reading a systematic review, the following aspects are important to consider when judging its quality and the potential impact of its conclusions for their practice.

## Clinical scope of the review

The first question for clinicians is 'Is the research question of the review relevant for my practice?' Early identification of colorectal cancer is important for all physicians. The traditional alarm symptoms were often only identified in retrospect in secondary care. Diagnostic characteristics of signs and symptoms are specific to the level of care – primary, secondary or tertiary. So a systematic review focusing on the diagnostic value of symptoms in patients in primary care does have additional value, despite the fact that several similar projects have been conducted in secondary care and in mixed patient populations. In the case of the review by Jefferis *et al*., one could question the need for pooling results, if all randomised controlled trials demonstrate that antibiotics do work in cases of bacterial conjunctivitis. However, the review focuses on patients in whom bacterial cultures had not been undertaken, which makes it a very relevant clinical question.

## Selection criteria for the studies

Before performing the literature search, the criteria for the studies to be selected need to be properly defined. What study design (for example, only diagnostic studies for the review by Astin *et al*.), which patients (in this case, patients with abdominal symptoms) and in which practice (in this case, from primary care only). Another important issue is the outcome under study, and how it was diagnosed. In the review by Astin *et al*. this may seem obvious as it focuses on colorectal cancer, but it is important to realise that large polyps, being pre-malignant disorders, may also be considered as relevant disease outcomes in diagnostic studies. Even though they are not always symptomatic, they might be the cause of rectal bleeding in some cases.

**Example of an individual patient data (IPD) meta-analysis (Jefferis *et al*.)**

The authors' aim was to determine the benefit of antibiotics for the treatment of acute infective conjunctivitis in primary care and to identify which subgroups benefit most.

Three eligible trials were identified. Individual patient data were available from 622 patients. Eighty per cent (246/308) of patients who received antibiotics and 74% (233/314) of controls were cured at day 7. There was a significant benefit of antibiotics versus control for cure at 7 days in all cases combined (risk difference 0.08, 95% confidence interval [CI] = 0.01 to 0.14). Subgroups that showed a significant benefit from antibiotics were patients with purulent discharge (risk difference 0.09, 95% CI = 0.01 to 0.17) and patients with mild severity of red eye (as opposed to those with moderate or severe red eye) (risk difference 0.10, 95% CI = 0.02 to 0.18), while the type of control used (placebo drops versus nothing) showed a statistically significant interaction ($P$ = 0.03).

The authors concluded that patients with purulent discharge or a mild severity of red eye may have a small benefit from antibiotics. Acute conjunctivitis seen in primary care can be thought of as a self-limiting condition, with most patients getting better regardless of antibiotic therapy. Prescribing practices need to be updated, taking into account these results.

## Process and outcome of the literature search

Authors need to state what their search criteria were and in which databases the search was conducted. In addition, the outcome of the literature search needs to be specified in detail:

- how many studies were identified
- how many were eligible for inclusion, and on what grounds some were excluded
- finally, how many of the eligible studies could be used in the final analysis.

Search criteria need to be predefined in a search string, so that it can be repeated. The selection process needs to be done by two reviewers who independently scrutinise all the abstracts. The results of this search are usually reported in a *flowchart*.

## Quality assessment of the eligible papers

A valid systematic review is based on high-quality studies only. Various sets of criteria have been developed for systematic assessment of the methodological quality of studies. In the review by Astin *et al*., the QUADAS criteria[1] were used. These, and other sets of criteria,[2] capture the most relevant methodological aspects of the study design:

- Was the randomisation procedure described?
- Was the intervention blinded for patient and physician?
- Was the outcome assessment done using validated scales?
- Was the sample size pre-calculated and was the loss to follow-up described?
- Was the analysis done on all patients who started the intervention (intention-to-treat) or only on those who completed the follow-up (per-protocol)?

For every study that is included in the review, the score on these quality items needs to be assessed (again by two independent reviewers) and reported in a separate table so that this information is accessible for the reader.

**Example of a therapeutic meta-analysis (Mugunthan *et al*.)**

The authors aimed to systematically review randomised controlled trials that evaluate the effectiveness of minimal interventions to reduce the long-term use of benzodiazepines (BZDs) in UK general practices.

From 646 potentially relevant abstracts, only three studies (with 615 patients) met all the inclusion criteria. The pooled risk ratio showed a significant reduction or cessation in BZD consumption in the minimal intervention groups compared with usual care (reduction: risk ratio [RR] = 2.04, 95% confidence interval [CI] = 1.48 to 2.83, $P < 0.001$; cessation: RR = 2.31, 95% CI = 1.29 to 4.17, $P = 0.003$). Two studies also reported a significant proportional reduction in consumption of BZD from baseline to 6 months in intervention groups compared with the control group. The secondary outcome of general health status was measured in two studies, with both showing a significant improvement in the intervention group.

The authors concluded that a brief intervention in the form of either a letter or a single consultation by GPs, for long-term users of BZD, is an effective and efficient strategy to decrease or stop their medication.

## Data extraction from the studies

In the next step, the results of the remaining studies, which are both eligible and of sufficient quality, are extracted. In the review by Astin *et al*., the predictive value for every individual abdominal symptom was taken from every single study, and summed. This generates an

overall pooled estimate of the positive and negative predictive values of the various symptoms for colorectal cancer. Not all studies did consider all symptoms, so the pooled results may be based on different patient numbers. The significance level of both the results of the individual studies as well as of the pooled result is expressed in the 95% confidence level. The pooling process is often visualised graphically in a so-called forest plot, which demonstrates the pooled result at a glance.

In some cases it may not be possible to quantify the summed result of the studies. If the studies differ too much in patient population or in outcome assessment, they may be too heterogeneous to pool the result (heterogeneity can be assessed using a specific test). Another reason may be that the studies used different types of measurement to assess the outcome, such as dichotomous (yes/no), ordinal (five-point scale) or continuous (0–100 scale).

In many reviews, additional subgroup analyses are performed to assess outcomes in different age groups, differences between males and females or differences in clinical presentation. These may help to explain the result of the review.

## Interpretation and clinical impact of the result

The final step is the interpretation of the effect size, in relation to the clinical question of the review. In a meta-analysis of intervention randomised controlled trials, this may be simple: if the confidence interval of the pooled risk ratio does not include 1, there is a difference between the intervention and the control group. However, the impact needs clinical interpretation. Although the review by Jefferis *et al*. found a statistically significant benefit of antibiotics in patients with mild (as opposed to moderate or severe) red eye and in those with purulent discharge, the difference is small – too small to justify widespread use in clinical practice.

In a diagnostic review, the interpretation is more complex, and the result also needs to be considered from a clinical perspective. In the review by Astin *et al*., the conclusion was that the diagnostic value of anaemia and rectal bleeding for colorectal cancer warrants systematic assessment.

It is also important that consideration be given to the risk of bias, such as publication bias (the fact that negative outcomes do have a lower chance of being reported).

## References

1. 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 Meth* 2003; **3:** 25.
2. Jadad AR, Moore RA, Carroll D, *et al*. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Control Clin Trials* 1996; **17(1):** 1–12.

Section 6

# Getting under the skin: qualitative methods in primary care research

Ann Griffin
Deputy Director, UCL Medical School, London

Relevant *BJGP* papers:

- Blakeman T, Chew-Graham C, Reeves D, *et al*. The Quality and Outcomes Framework and self-management dialogue in primary care consultations: a qualitative study. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X601389
- Mitchell C, Dwyer R, Hagan T, *et al*. Impact of the QOF and the NICE guideline in the diagnosis and management of depression: a qualitative study. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X572472
- Wainwright E, Wainwright D, Keogh E, et al. Fit for purpose? Using the fit note with patients with chronic pain: a qualitative study. Br J Gen Pract 2011; DOI: 10.3399/bjgp11X613133

## What is qualitative research?

Qualitative research is interpretative. It encompasses a broad range of approaches though all methodologies share one common aim, which is to explore multifaceted phenomena looking in detail at and reporting on the particular features that help us develop a deeper, more nuanced understanding of the object of our enquiry. They aim to portray the everyday, acknowledging that context is vital in shaping what individuals experience, understand and create. Qualitative research is a multidisciplinary field that is informed by a wonderfully rich range of influences – philosophical, sociological, psychological, anthropological and linguistic, to name a few. The diversity of theoretical perspectives and methodological approaches involved in qualitative work means that the production of a universally applicable checklist, or tick box, is not a straightforward matter; and some would contest the very notion.[1,2]

What constitutes robust practice in qualitative research will be governed, in part, by the guiding principles embedded in the specific methodological approach that the researcher has chosen to take.[3] Judgements about the quality of qualitative research should thus be mindful of, and reflect, the methodological steps that define these different approaches. For example, reviewing a paper that uses grounded theory raises different sorts of issues for the appraiser than reviewing a paper that utilises a phenomenological methodology. Some of these specific methodological matters that are contained in the sample texts will be highlighted in this section but, for further detailed guidance on common methodological approaches in health care, see the BEME Collaboration guides (http://www.bemecollaboration.org). Despite the unique attributes of the various approaches, the primary and unifying factor that marks out superior, robust qualitative research is a strategy, or research design, that has methodological integrity. How to demonstrate this coherency and rigour will be discussed and illustrated by reference to three key texts.

This section will focus on the common qualitative approaches used in healthcare research and provide guidance for those writing and appraising in this particular field. In qualitative healthcare research, there are three principal sources of enquiry:

- people, their views and experiences
- interactions, either interpersonal or interactions with artefacts
- structures such as the workplace, its organisation and its culture.

Typically, qualitative research uses first-person accounts, interviews, focus groups and (sometimes) observations as its main supply of data. This chapter will present 12 tips on how qualitative research can be shown to be both systematic and sensitive to the issues it claims to report.

## 1. Is qualitative research the correct approach?

A qualitative methodology attempts to explore phenomena in depth, to 'get under the skin' and to explore an issue in its fullest sense. It wants to hear, or see, people's everyday experiences and, through interpretative work, generate meaning. Research in this paradigm should:

- aim to reveal features of the real world that up till now have been hidden or tacit
- be conscious of, and explicit about, the context in which the empirical work is carried out
- go beyond just reporting, but explore, reinterpret, reflect and, where appropriate, challenge.

## 2. Is the research question pertinent to primary care?

The nature of qualitative work is open and the research question(s), or aim, should provide enough scope to interrogate the field. The use of the words 'explore', 'discover', 'understand' and 'reveal' are commonly seen in qualitative research questions.

Additional considerations include:

- Is the question answerable by a qualitative approach?
- Is it clear?
- Does it include the context as well as the line of enquiry?
- Is it worth asking; is it relevant to primary health care?

## 3. What is the role of theory?

The use of theory in qualitative research has varied from one discipline to another. Health sciences research is sometimes criticised for its insufficient use of theory but increasingly researchers are recognising its importance.[4] A theory can be used in the initial phases of the study to define the nature or essence of the phenomena being studied. It can also permeate through the research process and be used to analyse data. Qualitative research can also be used to generate a hypothesis (see the paper by Blakeman *et al*.). Importantly, theory can be used to extrapolate local empirical findings and inform a wider context, thereby providing conceptual generalisability.

- Would a theoretical perspective add value to this study?
- Has the theory been sufficiently described?
- How has it been woven throughout the rest of the study?

The paper by Blakeman *et al*. took a constructivist perspective, a theoretical position that, broadly speaking, means that people's understanding of the world around them is gained

through active engagement in a specific, contextually bound, location which involves complex social practices. The meanings people make are therefore unique, multiple and dynamic. How does this view permeate the rest of their research?

## 4. Is there methodological integrity?

Getting the design right is the critical step in making a qualitative piece credible. Methodological integrity means that theory resonates in the methodology, or research recipe,[5] and that these perspectives, in turn, inform the choice of methods by which data are gathered, analysed and presented. Each choice is crucial and affects the very essence of the study and the knowledge claims the researchers are entitled to make.

- Is there a logical progression between theory, methodology, methods, results and discussion?

Blakeman *et al*. were informed by a social constructivist perspective. Their chosen methodological approach was ethnographic, qualitative and emergent. By using observation as well as interviews to explore this diversity, they engaged a constant comparative interpretative method, an a priori approach, drawn from grounded theory. A constructivist perspective was even applied to the transcription of their research materials.

## 5. What are the characteristics of the participants?

This is an important area in qualitative research. Samples are typically not 'representative' of the population as a whole, like they often aspire to be in quantitative research, but informants are usually chosen deliberately (purposefully) because they have the 'knower's' perspective; that is, they have a particular characteristic and/or first-hand experience of the phenomena being studied. Sampling has important implications for the results and a study's wider applicability.[5,6]

Mitchell *et al*. and Blakeman *et al*. used maximum variation sampling to try to capture the fullest possible range of views:

- Is the rationale for the sampling strategy clearly stated?
- Does it resonate with the theoretical and/or methodological perspective?
- Was this strategy applied to all informants and, if not, has this been explained?
- Did all the participants have the 'knower's' perspective?
- Does the sample give the research greater credibility and/or the ability to draw wider conclusions about the results?
- Is the variation of views presented and discussed in the article?

Wainwright *et al*. used a purposive (non-random) sampling strategy:

- In what way was it purposive?
- Who was likely to have come forward using this strategy and what effects would that have had on the results?
- How have the authors dealt with these limitations?

Mitchell *et al*. showed the demographics of the four participating practices and this revealed some intriguing differences between them. Presenting these data on the sample allows the reader to make their own decision about whether what is reported will be applicable to other settings.

How many informants, interviews, focus groups or observations is reasonable? This is a difficult question and supposedly inadequate sample sizes as well as their lack of universal 'representativeness' are criticisms frequently levelled at qualitative work. However, the aim of this sort of research is not to produce objective generalisable results but to explore subjective accounts. The in-depth nature of this work means that it can be legitimate to study very small numbers but most researchers will choose larger numbers, and the numbers of interviews and focus groups presented in these studies certainly appear to be sufficient – 86 observations, as in the study by Blakeman *et al*., is unusually high. In the study by Wainwright *et al*., their sample was flexibly controlled by their data analysis, another prudent design feature.

## 6. Which methods should be used?

Different methods will reveal different things. Interviews, visual images, policy texts and observations represent the potential range of methods that can be used in qualitative research. However, most healthcare qualitative researchers commonly work with texts; these are commonly transcripts of what people have reported at interview or during a focus group. Observations are less frequently the primary source of data gathering, although they can be used to supplement interviews and to record non-verbal cues, group dynamics, and so on. Observations have the advantage of being able to report not only what people say they do but observing what they actually do in practice, adding an extra complexity to the interpretation.

**Focus groups versus interviews**

Focus groups are a means of bringing together a specific group of individuals to have a targeted conversation. They allow people to speak out, or to remain quiet, and generate socially constructed responses. Interviews, particularly one-to-one interviews, are used when information from just an individual is required and they are particularly deployed for interviewing about sensitive areas.

- Is the rationale behind the choice of interview versus focus group stated?
- Is the rationale for observing clear?
- Is there an interview/observation schedule and how does that affect the data?
- Has the researcher shown their interview/observation schedule?
- Does the schedule translate the research aim into appropriate operationalising questions?
- Has the interview/observation schedule been piloted and what changes were made?

Most qualitative research published in healthcare journals uses semi-structured interviewing; this allows the researcher to pose their own questions and to probe intriguing responses, and it provides an opportunity for informants to present their perspectives. However, structured and open formats are also possible for interviews and observations; again, if used, the rationale and subsequent implications for the data need to be discussed.

## 7. How should data be analysed?

Interpretation can be inductive or deductive and is often both. If you ask interview questions you are already shaping the sort of data generated and therefore some of the themes that emerge from your data will be those you wanted to gather. Inductive themes will be those that arise *de novo*.

- Is the method of analysis stated?
- Were the transcripts checked for accuracy?
- How were the themes or codes generated?
- Does this fit with the methodological approach?
- If a particular approach was used (thematic, framework, phenomenological, etc.), was it applied correctly?
- Was software used and, if so, how did it contribute to the process?

Grounded theory has been used in the studies by Blakeman *et al*. and by Wainwright *et al*. Grounded theory, which was devised by Glaser and Strauss,[7] is an approach typically used when very little is known about the empirical field and the research is exclusively exploratory. The researcher invites informants to talk freely about the area of study and tries not to control the outcome, by not asking specific or leading questions. The method of constant comparison is a specific feature of this sort of analysis where meaning emerges entirely from the data. In healthcare research, the grounded theory methodology is often slightly adapted, as we do usually have some prior knowledge and wish to ask certain questions.

## 8. How should data analysis be presented?

Presenting qualitative work in a meaningful way and demonstrating the complexity of issues and of consensus as well as outlying voices is a major challenge given the restrictions of word count. Tables of coding themes and hierarchies or the use of software to generate summaries of coding information or models may help the reader to quickly get a better idea of the scope of the data analysis. However, many would argue that this is quantifying qualitative work and is therefore inappropriate. The way forward for qualitative researchers and for publishers may be to increase flexibility with regard to word count limits and to use the capacity of online publishing.

General points to consider when assessing the data analysis:

- Have the results been presented in a way that is consistent with the research design?
- Are there enough data to explain the theme?
- Do the data appropriately illustrate the themes?
- Does the analysis show outlying voices too?
- Do the quotes identify the source?
- Are all sources represented in the data?

The paper by Blakeman *et al*. showed interactions and those by Mitchell *et al*. and by Wainwright *et al*. showed quotes from individuals that were consistent with the research design.

If the recruitment strategy was to capture the broadest range of views, this should be reflected in the results. If there are no outlying voices and no contrary views despite an active search, this is an important finding in itself.

## 9. What makes data analysis robust?

Validity, reliability and generalisability tend to be words associated with a positivistic paradigm. In contrast, many qualitative researchers use the words authenticity, credibility and trustworthiness to demonstrate rigour in their research process. A range of methods can be deployed, including:

- multiple coding (including inter-rater coding agreements available in qualitative software) – data independently analysed by more than one researcher

- peer audit – review by colleagues or supervisors
- triangulation with other data sources
- participant validation – asking informants to corroborate data analysis
- data saturation – gathering and analysis continue until no new themes arise
- a logical research design that permeates the entire research process.

What approaches have these three papers taken?

See chapter 30 in the book *Researching Society and Culture* by Seale for further guidance.[8]

## 10. Has the researcher addressed their role in the research process?

Different methodologies take different approaches to acknowledging the effect of the researcher on the research process. Some methodologies work on the principle that the researcher and his or her biases and assumptions can be eliminated from the research process while others consider the researcher as an active agent who constructs and co-constructs the research at all stages. If the latter, intersubjective stance is part of the methodological approach, the researchers should usually go to some lengths to demonstrate their reflexivity.

- Have research bias and assumptions been declared?
- Has the research team explained how they have dealt with this?
- Does their approach fit with any methodological guidance?

All three papers discussed in this section presented ways in which researcher bias may be present in their study and how they attempted to mitigate against it overly influencing their analysis. Mitchell *et al*. used field notes, regularly challenging the coding in the thematic analysis as well as employing independent scrutiny. Blakeman *et al*. accepted that they were active agents in the research process but that through a reflexive approach they could attempt to minimise their ‘preconceived notions’.

## 11. What are the ethical issues?

What are the ethical issues involved in asking people for their opinions, exploring their views or observing them? Clearly, they are different from therapeutic interventional studies, but there are just as many issues that may need addressing. The ethical issues raised are largely dependent on the question and context: the more sensitive the issue and the more research intrudes into clinical work and involves patients, the more likely it is to need formal NHS permissions. However, all work should make an explicit statement about ethics.

- Has the paper made a statement about ethical clearance?
- Has it said how the participants were informed and gave their consent?
- Has it declared how it will maintain confidentiality?

What are the various ethical issues raised by the three research papers discussed in this section? How is the ethics of observing different from the ethics of asking? What are problematic data and what responsibilities does the researcher have?

The website *Research Ethics Guidebook: a resource for social scientists* provides a comprehensive overview (http://www.ethicsguidebook.ac.uk).

## 12. What is the meaning of this research?

The results should be meaningfully developed into a discussion and conclusion, showing where and in what way they are related to the bigger picture and other disciplines, and this may be from empirical and theoretical fields. In order to be worthwhile, research has to contribute to building upon our existing knowledge base. It is prized if it can impact upon our own work and change practice.

- Is the research useful to you and your practice?
- Do the results justify the conclusions?
- How does it add to what we already know?
- How wide is its impact?
- What does it contribute to the next steps?
- Can it develop our conceptualisation of the topic?

All three papers discussed in this section investigate the impact of artefacts on clinical practice and reveal their profound effects on the way healthcare practitioners work. Qualitative research has the capacity to theorise further about the nature of these socially constructed tools. Local empirical findings can be held in relation to broader theoretical frameworks and, by doing so, conceptual generalisability can be shown.

## In summary

Primary care research frequently uses qualitative methodologies to explore issues in detail, to 'get under the skin' and investigate complex phenomena in depth. Qualitative methodologies include a broad range of approaches. Each theoretical perspective has its own particular way of conducting the investigation and this, naturally, influences every step in the research process. Judging what counts as high-quality qualitative primary care research is also sometimes difficult because of the word count limitation of many healthcare journals. This restriction frequently prohibits a detailed methodological discussion and sometimes leaves reviewers with queries. However, the following pointers are crucial for all qualitative work:

- Rigour and quality are demonstrated through the methodological approach to the research, the methodological integrity and coherent research design.
- The methodological approach influences the choice of methods, analysis, presentation of results, and so on, and these need to be compatible with the particular methodological approach stated.
- Theory should be used to help healthcare research develop a greater understanding of its practices and facilitate generalisability.

## References


1. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ* 2001; **322(7294):** 1115–1117.
2. Mays N, Pope C. Qualitative research in health care. Assessing quality in qualitative research. *BMJ* 2000; **320(7226):** 50–52.
3. Park S, Griffin A, Gill D. Working with words: exploring textual analysis in medical education research. *Med Educ* 2012; **46(4):** 372–380.
4. Reeves S, Albert M, Kuper A, *et al*. Why use theories in qualitative research? *BMJ* 2008; **337:** a949.
5. Clough P, Nutbrown C. *A student's guide to methodology: justifying enquiry*. London: Sage: 2002: 21–39.
6. Robson C. *Real world research: a resource for social scientists and practitioner-researchers. 2nd edn.* Oxford: Blackwell Publishing, 2002.
7. Glaser B, Strauss A. *The discovery of grounded theory*. Chicago: Aldine, 1967.
8. Seale C. *Researching society and culture. 3rd edn*. London: Sage, 2012.

Section 7

# Can we afford it? Costs and benefits of interventions

Anne Boyter* and Douglas Steinke†
*Senior Lecturer, Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow
†Senior Lecturer, School of Pharmacy and Pharmaceutical Sciences, University of Manchester

Relevant *BJGP* papers:

- Edwards RT, Neal RD, Linck P, *et al*., on behalf of the CHARISMA study group. Enhancing ventilation in homes of children with asthma: cost-effectiveness study alongside randomised controlled trial. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X606645
- Bryan S, Dormandy E, Roberts T, *et al*. Screening for sickle cell and thalassaemia in primary care: a cost-effectiveness study. *Br J Gen Pract* 2011; DOI: 10.3399/bjgp11X601325

## Introduction

NHS resources are limited and thus economic analyses are gaining increasing prominence in the health service. There are four common types of economic analysis:

- cost-minimisation (CMA)
- cost-effectiveness (CEA)
- cost–utility (CUA)
- cost–benefit (CBA).

Each type of analysis has a defined role in health economics. CMA and CEA are generally used to assess technical efficiency and answer questions related to how something should be done. CUA and CBA are generally used to answer allocative efficiency questions relating to whether something should be done compared with another, often unrelated, option. In the NHS, related to medicines and services, it is usually a technical efficiency question that is to be answered and thus the most common analysis used is CEA. Each analysis design uses a different outcome measure depending on what the results of the analysis are to be used for. For example, CEA results are used to identify the most effective option for the money spent, while CUA results are used to identify the quality-adjusted life years (QALYs) gained or lost by using one option over another.

Cost-effectiveness is seen as the fourth hurdle, after safety, efficacy and quality, to the approval of new medicines and services for the NHS. The National Institute for Health and Clinical Excellence (NICE) in England and Wales and the Scottish Medicines Consortium (SMC) in Scotland require a CEA as part of the submission for approval of new medicines or new licence indications.

It is therefore important that clinicians have a basic understanding of economics to allow them to evaluate the ever-increasing economics literature. Basic critical evaluation techniques are

demonstrated in this section that may help the reader understand what should be identified in an economic analysis.

## Paper 1 (Edwards *et al.*)

### Is the economic evaluation valid?

The study by Edwards *et al*. aims to evaluate the cost-effectiveness of improvements in housing, specifically related to improvements in heating and ventilation. The outcome used is improvement in asthma control; the patients' medication remained unchanged. As this is an intervention that has implications for both the health budget and local authority budgets, the perspective (or viewpoint) of the evaluation is valid. The economic analysis was carried out alongside a pragmatic clinical trial of housing modification. In general, this is a better design than basing the economic analysis on results of other papers. When information from other studies or papers is used, not all the information required may be available or the population may be inappropriate. In this paper, a good population size of 177 was used – this is large enough to provide a good clinical trial result but may be smaller than required for an economic analysis, where there is often a need for larger numbers in the study owing to the uncertainty associated with the estimation of some of the costs and consequences.

The intervention is well described – the modification of the house to bring the heating and ventilation to a defined standard, although this standard is not defined in the paper but in a companion paper. In addition, the companion paper gives the full details to explain that the intervention is clinically effective. The companion paper also gives the full details of the patient population, which is needed to assess whether the outcome is applicable to other populations. When reading the economic analysis, the companion paper should be accessed so that you can be well acquainted with the methodology and the applicability of the results to your setting.

In this paper, the effects of the intervention in terms of improvement in asthma control are well defined. However, the distinction between moderate and severe asthma was determined (arbitrarily) by the authors as being at the median of the PedsQL scores for the sample at the start of the study, and this may be different in another population. Therefore, this may limit the applicability and generalisability of the study. The single outcome of the PedsQL score collected by parental self-report, along with the lack of clinical indicators or analysis of mould spores and the short-term follow-up, also limit this study. Having a quantitative outcome measure in the trial, such as spirometry measurements, would have increased the robustness of the study and the economic analysis.

### How were the consequences and cost assessed and compared?

The perspective of the intervention is stated as 'multi-agency, public sector' and thus all the NHS and social costs should be taken into consideration. All the costs to the NHS and to the council for the modifications to the houses are taken into account, but none of the costs borne by the family are considered in this analysis. If family costs were included then the study would be from the patients' perspective. The authors also confirm the problem of using either patient or GP information on the costs of health care. These differ in this paper and the authors make the decision to use the GP resources to estimate the NHS costs and they thus include the costs of prescribing. This means that the same resources are used for the estimation of costs and could be validated by sensitivity analysis. The costs are well defined in the tables so you can compare how these reflect local costs.

As this analysis took place over the limited period of 1 year, there was no need for discounting of either the costs or the benefits. Appropriate sensitivity analysis was undertaken to investigate the costs that could vary; these are the building costs but not the costs of medicines, which

remain constant across the UK. The use of the PedsQL score as an outcome measure is interesting as it was used as an arbitrary severity measure for distinguishing between moderate and severe asthma: those patients with asthma below the median PedsQL score were defined as having severe asthma and those on the median and above as having moderate asthma. This may not be transferable to other populations as the median PedsQL score will be different. Sensitivity analysis was carried out on each of the subpopulations to investigate whether this definition had an effect. If the study were to investigate the effects of housing modification on asthma over a 3-year period then discounting of costs and benefits would have to be included, increasing the complexity of the analysis.

The results of the evaluation are well described and thus it would be simple to compare the costs from the study to local costs. In terms of prescribed medication, differences in the use of bronchodilators and inhaled steroids in the two groups were not statistically significant. This may be an effect of the small sample size or may reflect better prescribing in the intervention group. It could also show that medication use does not change between severe and moderate asthma as defined by this group, but the quality of life of the patient does change. Perhaps a change in medication use would be seen in a longer follow-up period, but discounting would then have to be considered. The main outcome used for cost-effectiveness was the parent-reported PedsQL score but this is parent-reported data rather than independent assessment.

### Will the results help purchasing for local people?

The sensitivity analysis gives a good indication that the intervention would be applicable in other areas: as the costs for the heating and ventilation modifications were varied based on the costs from other areas (both higher and lower costs than the setting for the study), the analysis still indicated that the intervention would be cost-effective.

## Paper 2 (Bryan *et al.*)

### Is the economic evaluation valid?

The paper by Bryan *et al*. has a companion paper that was published in the *BMJ*.[1] The two papers should be read together so that you can understand the clinical trial on which the economic analysis is based.

The study explored the cost-effectiveness of offering antenatal screening for sickle cell disease and thalassaemia in a primary care setting, during the pregnancy confirmation visit. The question ‘How should the screening be carried out?’ is one of technical efficiency but this assumes that the screening is worth doing, which is determined by asking an allocative efficiency question. Three alternative options of screening were compared using cost-effectiveness. The effectiveness measure was the number of women screened by 12 weeks’ gestation.

The three alternatives are clearly described in the Introduction section and, based on this, the effectiveness measure of the number of women screened is appropriate. This does not take into account the screening of the fathers, which is also mentioned in the Method section but which only occurs in the parallel model for all fathers. The effectiveness measure used was thus the only one applicable to all three models.

### How were the consequences and costs assessed and compared?

The costs for this model are well defined – these are the costs relating to screening and testing, and the costs of antenatal diagnosis and termination of pregnancy. The main analysis was carried out from the NHS perspective but the service users’ costs were also included in the sensitivity analysis. The sources of the NHS cost information are clear and detailed and

the sources are clearly identified. As the time frame of the intervention is less than 1 year there is no need for discounting to be included in the analysis.

The patient information collected was based on questionnaires completed by a sample of the participants in the study. However, the size of the sample was not stated so it is difficult to decide whether this sample was representative of the population as a whole.

Details of the economic analysis are clearly defined in the methodology with a clear flow diagram. The costs used in the model are clearly detailed so that comparison can be made with costs in different areas. The one cost that seems to be missing is the cost of training the GPs in screening and counselling for sickle cell disease and thalassaemia – this may have an impact on the outcome and could have been included in the sensitivity analysis. The probabilistic analysis indicates that the intervention is both more costly and more effective (more women are screened)

The analysis is limited in that, although the costs of termination are included, the costs that would be associated with the continuation of the pregnancy are not included. The study appears to assume that all parents will choose termination of the pregnancy. For the parents who opt to continue with the pregnancy, the future costs may be offset by the parents' ability to plan. In general, this is an issue with many economics analyses of screening, where many undiagnosed and asymptomatic patients may be identified and then treated, resulting in an increased burden on the NHS. To undertake an analysis that included all the long-term costs and consequences would be complex and involve the use of discounting.

The incremental analysis indicates that the incremental cost-effectiveness ratio is £13 per additional woman screened by 10 weeks, for the move away from the midwife care service to a primary care sequential programme.

### Will the results help purchasing for local people?

The results of this analysis could be transferred to other areas of the UK based on the analysis undertaken but care should be taken when considering the applicability for areas with lower potential rates of sickle cell disease and thalassaemia. Additional comments could have been made about the applicability of the screening to areas where the target group are a minority, as the study was carried out in an area rich in the target group thus giving economies of scale.

## Reference

1. Dormandy E, Gulliford M, Bryan S, *et al*. Effectiveness of earlier antenatal screening for sickle cell disease and thalassaemia in primary care: cluster randomised trial. *BMJ* 2010; **341:** c5132.