Critical Appraisal for Primary Care

Edited by Professor Roger Jones
Editor, *British Journal of General Practice*
Deputy Editor, *BJGP Open*

Royal College of General Practitioners

Royal College of General Practitioners

30 Euston Square, London NW1 2FB
Web: www.rcgp.org.uk

Royal College of General Practitioners is a registered charity in England & Wales (No. 223106) & Scotland (No. SC040430).

The Royal College of General Practitioners is a network of more than 52 000 family doctors working to improve care for patients. We work to encourage and maintain the highest standards of general medical practice and act as the voice of GPs on education, training, research, and clinical standards.

Contents

How to read and appraise a research paper

Roger Jones
Editor, *British Journal of General Practice* & Deputy Editor, *BJGP Open*
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; and
- policy makers and managers, who may need to know how robust the emerging evidence is for new methods of treatment and healthcare delivery.

This document is intended for 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. We have concentrated on six important types of research study:

- surveys;
- research using large databases;
- randomised controlled trials;
- systematic reviews and meta-analyses;
- tools for diagnosis and measurement; and
- qualitative studies.

For each of these, we have provided a citation to papers 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 two of the RCGP curriculum statements:

- 2 02 — Patient Safety and Quality of Care; and
- 2 04 — Enhancing Professional Knowledge.

The competencies involved are summarised in Boxes 1 and 2. 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. Patient Safety and Quality of Care

- The RCGP aims to improve the quality of health care by defining and upholding high standards for general practice education and training, aiming to improve health outcomes for all by promoting high quality general practice at the heart of the health service.
- As a GP you are in a strong position to influence the care of your own patients, that of your practice population and that of the wider healthcare community.
- Understanding how and when to apply tools and metrics to improve the quality of care is a key skill that can and should be learnt during your training, as well as enhanced in lifelong learning.
- Working in partnership with your patients and understanding their needs is vital to improving clinical care and reducing health inequalities.
- Patients, their families and carers have an important role in the assessment of health care; their views are therefore essential for the development of high- quality health care. Patients should be encouraged to be actively involved in planning their care and in the development of services at practice level and beyond.
- How we learn from and share lessons regarding clinical care is an important marker of our personal and collective professional development.

Box 2. Enhancing Professional knowledge

- As a GP you should have the skills to learn, critically appraise, and teach.
- You should be able to appraise research and guidelines critically, understanding their generalisability and validity.
- You should be able to apply evidence in the context of the patient, the community, and the healthcare setting.
- You should be able to audit your own practice and that of your organisation, and develop changes in the light of the findings.
- You should be able to work within a multidisciplinary team so that the views and knowledge of the whole team are applied when discussing the care of a patient.
- You should be able to demonstrate the competences of shared leadership so as to maximise the effectiveness of healthcare delivery.
- You should ensure you are up-to-date in managing the acute care of patients.
- You should, as part of supervising others in your team, be able to teach the need for safer practice and better patient care.
- You should be willing to receive feedback as a teacher from individuals or groups in order to improve and learn from your teaching and educational sessions.
- You should be aware that your own health and that of your colleagues should be optimal to ensure safe 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 email 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, BJGP Open,* and the *BMJ,* and your practice, colleagues, or 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* now print one-page summaries of their research papers, with the full paper that includes the references, tables, and figures available 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. Don't forget that, though you may be most interested in reading about research carried out in primary care, 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 them 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 that 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 3.

Box 3. 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 Method 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 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

Luke Daines* and Aziz Sheikh†

*GP and CSO Academic Clinical Fellow, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh

†Professor of Primary Care Research & Development and Director, Usher Institute of Population Health Sciences and Informatics, 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: https://doi.org/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: https://doi.org/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: https://doi.org/10.3399/bjgp11X613115
- Hall GC, Tulloh LE, Tulloh RMR. Kawasaki disease incidence in children and adolescents: an observational study in primary care. *Br J Gen Pract* 2016; DOI: https://doi.org/10.3399/bjgp16X684325

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 four papers selected for inclusion in this section. Though 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 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 now widely employed to undertake descriptive enquiries.[3,4]

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

1. Were important questions asked?
2. Were the methods appropriate, and thus are the results 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', our aim, when participating in peer review for a journal, is 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.[5]

Critically appraising descriptive studies

We 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*, Cornford *et al,* and Hall *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.[6] 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,[7] the decision by Mathur *et al* 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 socioeconomically disadvantaged populations in the UK (Tower Hamlets, the City, Hackney, and Newham in London). 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.
- Hall *et al* provided data on the incidence and seasonal variation of Kawasaki disease. Although a rare condition, early recognition can limit serious sequelae, making it an important differential to keep in mind in primary care.
- Though 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 four 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 thus are the results 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.[8]
- 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, though 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 our 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. However, we 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. We also find it much more informative to be presented with 95% confidence intervals rather than *P*-values.[9]

As with many such primary studies, the response rate was disappointing at only 46%. Failure to gather data from a high proportion of those in the sample is an important source of bias in surveys.[10] Although a response rate of at least 70% is often sought by investigators and reviewers, there is no agreed cut-off for an adequate response rate.[10,11] Instead, thoughtful analysis, interpretation, and explanation of the findings is needed, particularly if there is a low response rate, because of the imprecision in estimates and the inherent risk of bias in such studies.

The impact of non-response on the results depends not just on the proportion who do not respond, but the degree to which the non-responders are systematically different from the population.[11] Of those selected in a sample, non-response can occur due to the:[11]

- □ method used to collect data not reaching the responder;
- □ responder not wishing to participate; and
- □ responder being unable to provide data, for example, due to barriers such as language, disability, or illness.

These can all be important, as there may be non-random or systematic differences between responders and non-responders. Understanding the effect of non-responders on the summary estimates is challenging, as the characteristics and demography of the non-responders is rarely available. In such cases, it is important to reflect critically on the potential impact of non-responders on the overall findings, and then carefully interpret findings.

Conducting a pilot study, like A'Court *et al*, can be useful to identify the likely response rate and inform the sample size required for the main study.[11] This can also help inform deliberations on the sample size needed for any subgroup analyses, which should be planned *a priori*.

Overall, the low response rate, together with the regional nature of the study, raises important questions about the external validity of the findings from this work.

- Hall *et al* determined the incidence of Kawasaki disease in children aged <20 years using a retrospective, observational design. Data from a large UK-wide database of routinely collected primary care records (The Health Improvement Network [THIN]) were analysed, providing a large sample size, thereby increasing the precision of the estimates. The use of a comprehensive search strategy to identify possible incident cases (including Read Codes, free text searches, and prescription records) was used, minimising the chance of missed cases due to poor data quality. In our judgement, the use of a large dataset, coupled with a carefully considered analysis, contributes to a study with high internal and external validity.
- Cornford *et al* appear to have employed a retrospective cohort design in a single, rather atypical practice, which immediately raises major concerns about the generalisability of this study's findings. The authors aimed to describe the incidence of deep vein thrombosis in opioid users, but we needed a fuller description of the cohort and, given that practice populations are dynamic, this estimate should really have been expressed as a function of person–years at risk, rather than as a percentage. We also wondered how accurate the recording of relevant diagnoses was likely to have been, and about the associated risks of bias. Therefore, we had major concerns about extrapolating the findings from this study to other GP practice populations.
- Overall, we believe that these studies have made reasonable attempts to address the questions at hand, but there were clearly limitations, and these need to be critically reflected on by the authors.

3. Have the findings from this work been critically reflected on?

- The hallmark of a well-trained investigator/research team is being self-critical and reflective in positioning their study in relation to the wider literature. The move to structured discussions of research papers that many journals, including the *BJGP*, have now made has been very helpful in ensuring that these issues are systematically considered. Checklists such as the STROBE Statement offer a list of key items that should be included in observational studies, and are particularly valuable when critically appraising descriptive studies.[12]

- Mathur *et al* were correct to highlight the limitation resulting from their collapsing of ethnic groups, and the resulting inability to examine within-ethnic-group heterogeneity. They should, however, also have reflected on the possibility of unmeasured confounding factors, in particular social deprivation. We would also have liked them to explain more fully how this work builds on previous related work, particularly in relation to the relative merits of focusing on softer surrogate measures, rather than harder clinical endpoints. We found the discussion on the implications of this work to be rather underdeveloped.
- Hall *et al* considered some of the challenges of using routinely collected data, and appropriately discussed the potential for under-representation of particular ethnic minority groups due to lower rates of registration with a GP. Implications for clinical practice are suitably presented, and their findings are set in context through a comparison with similar studies from the UK and worldwide.
- Both the A'Court *et al* and Cornford *et al* studies also made welcome attempts to reflect on the strengths and limitations of their work, and to consider the implications of their research for clinical practice and future research.

Conclusion

These studies attempted to answer questions of importance to GPs and their teams. Though not without limitations, the studies by Mathur *et al*, Cornford *et al,* and Hall *et al* are likely to have yielded valid and clinically relevant results, though we are less confident that this is also the case in relation to the study by A'Court *et al.* Nonetheless, each of these papers has, in one way or another, provoked and informed our thinking in relation to clinically relevant questions and design-related considerations.

References

1. Greenhalgh T. *How to read a paper: The basics of evidence-based medicine*. 5th edn. Chichester: Wiley Blackwell Publishing, 2014.
2. Anandan C, Simpson CR, Fischbacher C, Sheikh A. Exploiting the potential of routine data to better understand the disease burden posed by allergic disorders. *Clin Exp Allergy* 2006; **36(7):** 866–871.
3. Cresswell K, Sheikh A. Electronic health record technology. *JAMA* 2012; **307(21):** 2255–2256.
4. Sheikh A. Evidence-based restructuring of health and social care. *PLoS Med* 2017; **14(11):** e1002426. https://doi.org/10.1371/journal.pmed.1002426.
5. Sheikh A, Stephenson P. Next steps in reforming the PCRJ's peer review process. *Prim Care Respir J* 2011; **20(4):** 347–348.
6. Bhopal RS. *Ethnicity, race, and health in multicultural societies: foundations for better epidemiology, public health, and health care.* Oxford: Oxford University Press, 2007.
7. Barnett K, Mercer SW, Norbury M, *et al*. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *Lancet* 2012; **380 (9836)**: 37–43.
8. Coggon D, Rose G, Barker DJP. *Epidemiology for the uninitiated*. 5th edn. London: BMJ Publications, 2003.
9. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. *BMJ (Clin Res Ed)* 1986; **292(6522):** 746–750.
10. Fowler FJ. *Survery research methods*. Thousand Oaks, CA: Sage Publications, 1989.
11. Kelley K, Clark B, Brown V, Sitzia J. Good practice in the conduct and reporting of survey research. *Int J Qual Health Care* 2003; **15(3):** 261–266.
12. Von Elm E, Altman DG, Egger M, *et al*. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *J Clin Epidemiol* 2008; **61(4):** 344–349.

Section 2

Critical appraisal of database studies

Clare R Bankhead,* and Richard J Stevens,†
*Associate Professor, Nuffield Department of Primary Care Health Sciences, University of Oxford
†Course Director, MSc in Evidence-Based Healthcare Medical Statistics
Nuffield Department of Primary Care Health Sciences, University of Oxford

Relevant *BJGP* papers:

- Hippisley-Cox J, Coupland C. Identifying patients with suspected lung cancer in primary care: derivation and validation of an algorithm. *Br J Gen Pract* 2011; DOI: https://doi.org/10.3399/bjgp11X606627
- Lockhart P, Guthrie B. Trends in primary care antidepressant prescribing 1995–2007: a longitudinal population database analysis. *Br J Gen Pract* 2011; DOI: https://doi.org/10.3399/bjgp11X593848
- Stapley SA, Rubin GP, Alsina D, *et al*. Clinical features of bowel disease in patients aged <50 years in primary care: a large case-control study. *Br J Gen Pract* 2017; DOI: https://doi.org/10.3399/bjgp17X690425

Introduction

Large databases of routine clinical data are increasingly widely used in primary care research. In principle, datasets derived from them are just a specific type of observational study, usually a (retrospective) cohort study. The strengths of these datasets are accompanied by specific challenges for the researchers, as we will illustrate below. Correspondingly, when we read such database research papers and try to apply checklists for appraising observational research, we find that underlying principles (population, measurement, follow-up) are the same, but the most pertinent details are quite different. In this article, inspired by the Critical Appraisal Skills Programme (CASP) checklist for cohort studies (http://www.casp-uk.net/find-appraise-act/appraising-the-evidence/), we propose the following bullet points to aid critical appraisal of database studies:

- What is the research question?
- Are the methods valid for this research question? (PROD)
  - □ population studied (P)
  - □ risk factors recorded (R)
  - □ outcomes assessed (O)
  - □ database analysis (D).
- What are the results?
  - □ Are the results clinically as well as statistically significant?

- Do the results apply locally?

What is the research question?

We illustrate these points with reference to three database research papers published in the *BJGP*. Hippisley-Cox and Coupland used the QResearch® database to derive and validate a risk score for detection of lung cancer. Lockhart and Guthrie used a pharmacy prescribing database from Tayside, Scotland, to examine trends in antidepressant prescribing over a 20-year period. Stapley *et al* used the UK Clinical Practice Research Datalink (CPRD) to identify and quantify clinical features of colorectal cancer and inflammatory bowel disease among patients aged <50 years. Two of these papers do not explicitly state a 'research question', preferring instead to use 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 in the Stapley paper, 'Which features are associated with colorectal cancer and inflammatory bowel disease in younger consulting patients?'

The importance of a pre-specified research question in database studies is increasingly recognised, and for this reason database owners typically require a protocol to be submitted for approval before releasing data.[1]

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 — by these criteria, few in the UK are ineligible.

The Lockhart and Guthrie paper illustrates an important consideration in database research — 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. This is the approach used in the Stapley paper, where the researchers restricted the study to patients (cases and controls) who had consulted at least once in the year prior to the diagnosis. In this example, it is a reasonable approach to take, as the research question was to identify clinical features before diagnosis and this information could only be ascertained if a consultation in primary care had occurred.

The study of a population with a particular condition necessitates the development of code lists to identify these conditions. To study people with epilepsy, Ridsdale *et al* identified people who have both diagnostic codes for epilepsy and prescription codes for anticonvulsant drugs:[2] 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, such as diabetes, it may be appropriate to include non-diagnostic codes that indicate attendance at a specialist clinic.[3] This approach would increase the sensitivity of the search, but at the expense of specificity. 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.

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 indications for treatment, may have to be inferred from other factors, such as the recorded reason for the consultation. The latter of these is a similar issue to identification of people living with a particular condition as described above. Authors may help the reader (see, for example, reference 32 in the Hippisley-Cox and Coupland paper[4]) by citing relevant papers from the growing literature on recording accuracy in database research.[5,6]

Outcomes assessed (O)

Lockhart and Guthrie's outcome, antidepressant prescribing, could equally be a risk factor in another study,[7] 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.[8] The reader could refer again to the literature recording validity in widely used databases 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 Stapley *et al*. There has been increasing interest in the use of propensity score methods for database analysis, but we will not give details here because, in general, these methods produce similar results to better known methods of matching or adjustment.[9]

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.[10] 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,[11,12] and an argument made that other methods, such as 'complete case analysis', require even stronger assumptions.[13]

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.[14,15] The studies by Stapley *et al* and Hippisley-Cox and Coupland avoided this bias by defining clinical features or risk factors strictly in the period before index date. See the article by Lévesque *et al*,[16] 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.[17] 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 to relatively cheaply test existing hypotheses, generate new hypotheses, and, in some cases, address previously unanswerable questions.

References

1. Waller P, Cassell JA, Saunders MH, Stevens R. Governance and oversight of researcher access to electronic health data: the role of the Independent Scientific Advisory Committee for MHRA database research, 2006–2015. *J R Coll Physicians* Edinb 2017; **47(1):** 24–29.
2. Ridsdale L, Charlton J, Ashworth M, *et al*. Epilepsy mortality and risk factors for death in epilepsy: a population-based study. *Br J Gen Pract* 2011; DOI: https://doi.org/10.3399/bjgp11X572463.
3. Tate AR, Dungey S, Glew S, *et al*. Quality of recording of diabetes in the UK: how does the GP’s method of coding clinical data affect incidence estimates? Cross-sectional study using the CPRD database. *BMJ Open* 2017; **7(1):** e012905.
4. 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.
5. 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.
6. Nissen F, Quint JK, Wilkinson S, *et al*. Validation of asthma recording in electronic health records: a systematic review. *Clin Epidemiol* 2017; **9:** 643–656.
7. Walker AJ, Card T, Bates TE, Muir K. Tricyclic antidepressants and the incidence of certain cancers: a study using the GPRD. *Br J Cancer* 2011; **104(1):** 193–197.
8. Khan NF, Carpenter L, Watson E, Rose PW. Cancer screening and preventative care among long-term cancer survivors in the United Kingdom. *Br J Cancer* 2010; **102(7):** 1085–1090.
9. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58(6)**: 550–559.
10. Sutton M, Elder R, Guthrie B, Watt G. 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.
11. 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.
12. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Med Res Methodol* 2010; **10:** 7.
13. Stuart EA, Azur M, Frangakis C, Leaf P. Multiple imputation with large data sets: a case study of the Children’s Mental Health Initiative. *Am J Epidemiol* 2009; **169(9):** 1133–1139.
14. Suissa S. Immortal time bias in pharmaco-epidemiology. *Am J Epidemiol* 2008; **167(4)**: 492–499.

15. 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, discussion 92.
16. Levesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010; **340:** b5087.
17. 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.

Section 3

Finding the best answer: randomised controlled trials

Richard Hooper* and Melanie Smuk†
*Reader in Medical Statistics, Queen Mary University of London
†Lecturer in Medical Statistics, Queen Mary University of London

Relevant *BJGP* paper:

- Little P, White P, Kelly J, *et al*. Randomised controlled trial of a brief intervention targeting predominantly non-verbal communication in general practice consultations. *Br J Gen Pract* 2015; DOI: https://doi.org/10.3399/bjgp15X685237

What is a randomised controlled trial?

> Little *et al* — Reviewer 1: *'A randomised controlled trial is the most rigorous way to test a clinical intervention, and the authors are to be commended on their efforts over a long period of time to consolidate a feasible study design.'*

Randomised controlled trials are often seen as the 'gold standard' for evaluating the effectiveness of an intervention. A randomised controlled trial takes participants and randomly assigns them either to an explorative intervention (there might be more than one) or to a control treatment (which might be routine care, or a placebo, or sham therapy). The groups are followed up in parallel using similar methods, allowing outcomes to be compared over the same time period. The element of a *control group* in the design (along with randomisation) is what makes the trial a randomised controlled trial.

Guidelines for clinical trials

Starting with the publication of the *Nuremberg Code* after the Second World War,[1] and the subsequent *Declaration of Helsinki*,[2] there have been attempts to articulate principles for the ethical conduct of clinical trials. At a local (that is, national) level, these have now become statutory requirements. In the UK, clinical trials are governed by the 2001 European Union *Clinical Trials Directive*[3] (which will be replaced by the EU Clinical Trials Regulation in 2019[3]) and the *Medicines for Human Use (Clinical Trials) Regulations 2004* and its subsequent amendments.[4] Regulations across Europe, Japan, and the US are harmonised under the aegis of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH), whose guidelines constitute a quality standard for clinical trials often referred to as Good Clinical Practice.[5]

Before any trial begins, it must be approved by a local research ethics committee or institutional review board. Any subsequent journal article reporting on the trial should include details of how ethical approval was obtained. Most journals also require authors to include a statement about how (and whether) participants consented.

> Little *et al*: *'Participants were any adult patient attending their GP who had agreed to participate in the study and were able and willing to consent to study procedures. Excluded were those who were unable to consent or complete questionnaires (for example, because of severe mental illness, severe distress, very unwell generally, and difficulty reading or writing).'*

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 *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly work in Medical 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.

> Little *et al*: *'At the time of developing the initial questionnaires to measure patient-centredness, an associated randomised trial was approved by the ethics committee, and prior to the requirements to register trials.'*

The intervention

> Little *et al*: *'The consultation is central to all medical encounters, and patient-centred communication is highlighted as the core of good practice, yet the evidence to inform training needs of health professionals of what they need to do in consultations to maximise effective verbal and non-verbal communication is limited.'*

Authors should be clear about what question the trial was intended to answer. Some trials are intended to show that a novel intervention *can* work under tightly controlled, ideal conditions (explanatory trials or trials of efficacy). Some are intended to show that it *will* work when rolled out into routine practice (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]

> Little *et al*: *'A strength of this study, which investigated a simple intervention to improve communication, is that it could be rolled out relatively easily.'*

The active and control interventions should both 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]

> Little *et al* — Reviewer 2: *'Do you know how many videos each of the GPs in the intervention group watched? I think this is important, because it impacts on how time consuming the intervention would be — videoing and reflecting on 15 consultations would take quite some time.'*

The comparator or control group

> Little *et al* — Reviewer 1: *'As the authors describe, this is a complex intervention, consisting not just of a training intervention but also of reflective video feedback on performance. While this is noted in the strengths and limitations section, the distinction between the two could perhaps be more detailed since, as the authors state, other studies have assessed the impact of training interventions alone. The authors might wish to comment on whether the control group could have been given video feedback of their existing consultation style without the training intervention. My hunch would be that it is the video reflective prompt which is the more powerful component, since most of these GPs would have had prior training in the elements of good consultation style which were contained in the KEPe warm intervention.'*

Authors need to make the nature of the control intervention clear. The choice between routine care, a placebo, a sham therapy, or an attention control (or something else) needs careful consideration. The decision of which to use will be contextual to the study.

If there is an existing standard beneficial treatment, on ethical grounds this should be offered to participants in preference to no treatment at all. As a control treatment this is usually described as routine care.

In a drug trial, a placebo is a treatment which has the same appearance as the active intervention but is pharmacologically inactive. Placebos must look, smell, feel, and taste the same as the active intervention to ensure that participants and clinicians remain blind to the treatment allocation (see below).[13] The Latin word 'placebo' means 'I please', and authors should also be aware that the act of taking a placebo may itself have some benefit to the participant. A well-documented example of this placebo effect is within hypertension trials: the participant is often reassured by taking an intervention, and this results in reducing their anxiety, which may have been partially responsible for raising their blood pressure. Treatment effects may be harder to see when placebo effects are large.[14,15]

When the intervention is a procedure of some kind, the equivalent of a placebo is a sham therapy which lacks the supposed active component but is designed to mimic the procedure convincingly enough to blind participants to their allocations. An example of a sham therapy would be an acupuncture trial, where the control group still receives needle therapy but in the incorrect locations. A related idea is the attention control, in which participants receive the same amount of contact time (attention) from health professionals as intervention participants.

Lastly, one of the most difficult placebo effects to reduce is the beneficial effect of just being in a trial. This is especially problematic if the trial is meant to be pragmatic.

Randomisation

> Little *et al*: '*The trial was originally designed as an individually randomised trial where patients would be randomised to a more empathic or less empathic encounter, and a more positive or less positive approach. Two issues forced a change in design …'*

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, for example, eight new participants 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, though 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 help ensure allocation concealment (see below).

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 participant's 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.[16]

> Little *et al*: *'Given the difficulties experienced in randomising individual consultations, the modified trial design was a cluster-randomised trial: GPs still available who had agreed to the initial observational phase of the study were randomised to receive the brief training intervention or no training intervention.'*

Blinding and allocation concealment

> Little *et al*: *'The open nature of the trial meant that neither GPs nor patients could be blinded to the intervention, although patients in both groups were simply told that this was a study assessing communication.'*

Blinding refers to steps taken to conceal group allocations once they have been made. 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; and
- 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.[17]

Allocation concealment is a technique which conceals the sequence of upcoming allocations from those allocating participants to groups. For example, steps may be taken to prevent a GP who is recruiting to a trial from knowing what group the next participating patient will be allocated to, though the GP might have to become unblinded once the allocation is made. Allocation concealment prevents the allocator from unconsciously or consciously influencing participant allocation, or even patient recruitment, into the trial. Traditionally, allocation concealment was achieved by giving recruiters a series of numbered, opaque, sealed envelopes containing the allocations, but there are concerns that methods like this are easily subverted,[17] and more secure electronic methods and randomisation services now tend to be favoured.

It is best for the manuscript to be explicit about which groups of people were blinded, and whether allocation concealment has been implemented. 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

> Little *et al* — Reviewer 1: *'198 participants with complete data were required.'*

Triallists have an ethical responsibility to plan the size of their clinical trials,[18] and a report of a trial should include a justification of the sample size. This is usually expressed in terms of the statistical power: the chance that the trial will find statistical evidence for an effect of the intervention, if a clinically important effect is present. Conventional targets for power are 80% or 90%. Power should be calculated at the design stage, not after the results are in.[19] The sample size calculation, the primary outcome measure, and the minimal clinically important treatment effect should all have been specified in the original protocol. The statement of power in the trial report should include all the information necessary to replicate it.[20,21]

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 X 1.2 = 120 people.)

The Results section should include a flow diagram recording numbers of people who were screened for eligibility, consented and 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.

> Little *et al* — Reviewer 1: *'The authors state that many patients could not be consented; it would be helpful to know to what extent the authors think this produced a bias in the final sample. There is an overall good response rate from those who were recruited.'*

There remains the 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.[22] 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.[23]

> Little *et al*: *'The data were analysed on an intention-to-treat basis, that is, patients were analysed according to their randomisation group, using complete data with no imputation of missing values.'*

As with the sample size calculation, all decisions about analysis should have been pre-specified before seeing the results, to avoid any conscious or unconscious bias.

Validity and generalisability

> Little *et al* — Reviewer 2: *'A further limitation of the study is that all of the GPs who took part had >10 years' experience. Could these results be generalised to younger, newly qualified GPs?'*

The validity of a trial is often judged on two components, internal and external. Accuracy of the estimated intervention effect for participants in the trial itself is referred to as 'internal validity', while the accuracy of the estimated intervention effect for the population in which the study's findings are to be applied is known as the 'external validity' (generalisability).

The internal validity is often threatened by systematic error (bias), and/or random error. Maximum effort should be made to minimise these elements in the trial conduct, design, and analysis. Within the trial design, careful attention should be paid to the randomisation, blinding, and sample size choices.

The external validity is often threatened by the trial setting, the type of intervention implemented, the study population (either by choice, availability, or ethics), and assessment design (for example, follow-up duration or type of outcome measured). In medicine, the evaluation of external validity often requires clinical rather than statistical expertise, as it 'depends on a detailed understanding of the particular clinical condition under study and its management in routine clinical practice'.[24] Having a clear specification of the target population will increase external validity.

In the past, trials often employed stringent eligibility criteria which reduced their external validity — excluding participants with psychological, psychiatric, and physical comorbidities, for example. Such restrictive eligibility criteria may be implemented 'in the hopes of ensuring the safety of vulnerable patients, reducing study costs, increasing study feasibility and/or interpretability, decreasing heterogeneity in response (and thereby increase statistical power), and complying with guidelines by regulatory agencies'.[25] However, these benefits do not always come to fruition and there is now a move towards greater inclusiveness in trials.[26] Researchers should think carefully about the balance of external validity and practical constraints, especially within effectiveness trials.

Other issues

In this summary, we have limited ourselves to issues that are specific to clinical trials, though 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 kinds. 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 *Recommendations*.[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*. https://www.wma.net/what-we-do/medical-ethics/declaration-of-helsinki/ (accessed 22 Feb 2018).
3. European Commission for Public Health: Medicinal products for human use. https://ec.europa.eu/health/human-use/clinical-trials_en (accessed 22 Feb 2018).
4. HM Government UK. *The Medicines for Human Use (Clinical Trials) Regulations 2004*. http://www.legislation.gov.uk/uksi/2004/1031/contents/made (accessed 22 Feb 2018).

5. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. *ICH Guidelines*. http://www.ich.org/products/guidelines (accessed 22 Feb 2018).
6. Consolidated Standards of Reporting Trials (CONSORT) Group. *The CONSORT statement*. http://www.consort-statement.org/ (accessed 22 Feb 2018).
7. Mhaskar R, Djulbegovic B, Magazin A, *et al*. Published methodological quality of randomized controlled trials does not reflect the actual quality assessed in protocols. *J Clin Epidemiol* 2012; **65(6):** 602–609.
8. International Committee of Medical Journal Editors. The recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. http://www.icmje.org/urm_main.html (accessed 22 Feb 2018).
9. Thorpe KE, Zwarenstein M, Oxman AD, *et al*. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. *J Clin Epidemiol* 2009; **62(5):** 464–475.
10. Boutron I, Moher D, Altman DG, *et al*. Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Ann Intern Med* 2008; **148(4):** 295–309.
11. Craig N, Dieppe P, Macintyre S, *et al*. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ* 2008; **337:** a1655.
12. Abraham C, Albarracín D, Araújo-Soares V, *et al*. *WIDER recommendations to improve reporting of the content of behaviour change interventions*. https://static-content.springer.com/esm/art%3A10.1186%2F1748-5908-7-70/MediaObjects/13012_2011_537_MOESM4_ESM.pdf (accessed 22 Feb 2018).
13. Kaptchuk TJ. Intentional ignorance: a history of blind assessment and placebo controls in medicine. *Bull Hist Med* 1998; **72(3):** 389–433.
14. Moerman DE. Cultural variations in the placebo effect: ulcers, anxiety, and blood pressure. *Med Anthropol Q* 2000; **14(1):** 51–72.
15. Kaptchuk TJ, Goldman P, Stone DA, Stason WB. Do medical devices have enhanced placebo effects? *J Clin Epidemiol* 2000; **53(8):** 786–792.
16. Campbell MK, Elbourne DR, Altman DG. CONSORT statement: extension to cluster randomised trials. *BMJ* 2004; **328(7441):** 702–708.
17. Herbison P, Hay-Smith J, Gillespie WJ. Different methods of allocation to groups in randomized trials are associated with different levels of bias: a meta-epidemiological study. *J Clin Epidemiol* 2011; **64(10):** 1070–1075.
18. Altman DG. Statistics and ethics in medical research: III How large a sample? *BMJ* 1980; **281(6251):** 1336–1338.
19. Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Stat* 2001; **55(1):** 19–24.
20. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995; **311(7013):** 1145–1148.
21. Kerry SM, Bland JM. Sample size in cluster randomization. *BMJ* 1998; **316(7130):** 549.
22. Hewitt CE, Torgerson DJ, Miles JN. Is there another way to take account of non-compliance in randomized controlled trials? *CMAJ* 2006; **175(4):** 347.
23. White IR, Horton NJ, Carpenter J, *et al*. Strategy for intention-to-treat analysis in randomised trials with missing outcome data. *BMJ* 2011; **342:** d40.
24. Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials* 2006; **1(1):** e(9).
25. Hoertel N, Lopez S, Wang S, *et al*. Generalizability of pharmacological and psychotherapy clinical trial results for borderline personality disorder to community samples. *Personal Disord* 2015; **6(1):** 81–87.
26. Van Spall HG, Toren A, Kiss A, *et al*. Eligibility criteria of randomized controlled trials published in high impact general medical journals: a systematic sampling review. *JAMA* 2007; **297(11):** 1233–1240.

Section 4

Measuring health and illness: development and validation of tools

Sarah F Moore,* Kevin Barraclough,† and William Hamilton‡
*Academic Clinical Fellow in Primary Care, University of Exeter
†GP, Hoyland House Surgery, Painswick, Gloucestershire
‡Professor of Primary Care Diagnostics, University of Exeter

Relevant *BJGP* papers:

- Snoeker BA, Zwinderman AH, Lucas C, Lindeboom R. A clinical prediction rule for meniscal tears in primary care: development and internal validation using a multicentre study. *Br J Gen Pract* 2015; DOI: https://doi.org/10.3399/bjgp15X686089
- Haasenritter J, Bösner S, Vaucher P, *et al.* Ruling out coronary heart disease in primary care: external validation of a clinical prediction rule. *Br J Gen Pract* 2012; DOI: https://doi.org/10.3399/bjgp12X649106
- van Ierland Y, Elshout G, Berger MY, *et al.* Translation of clinical prediction rules for febrile children to primary care practice: an observational cohort study. *Br J Gen Pract* 2015; DOI: https://doi.org/10.3399/bjgp15X684373

What are clinical prediction rules?

It is likely that you use clinical prediction rules (CPRs) frequently; scores are calculated and applied to diseases ranging from depression (PHQ-9)[1] to atrial fibrillation (CHA_2DS_2-VASc),[2] osteoporosis (FRAX™)[3] to cardiovascular disease (QRISK3),[4] and sore throats (Centor criteria)[5] to thromboembolism (Wells *et al*).[6] These and other similar scores can be used at all stages of the patient journey. They are most commonly used by UK GPs to aid diagnosis, assess severity, guide referral, comply with guidelines or Quality and Outcomes Frameworks (QOF) targets, and sometimes to educate patients.[7]

However, CPRs are not universally accepted by the clinical community. Current evidence suggests that a lack of familiarity and a consideration that the CPR is unnecessary, coupled with a preference for one's own clinical judgement, are key reasons that GPs do not use CPRs.[7] It is certainly true that clinicians have long used personal experience to make decisions, but the nature of this experience and its interpretation in differing contexts can lead to inequities in patient care. Many of these judgements have now been translated into predictive tools to aid clinical judgement. Indeed, some tools have already been shown to reduce inconsistency and improve accuracy of decision making.[8]

As medicine becomes ever more complex, the relevance of these tools for supporting clinicians in measuring health and illness grows. Over the past 20 years, the percentage of CPR-related papers on PubMed has expanded more than 10-fold, from 0.87/100 000 papers in 1996 to 10.2/100 000 papers in 2016.[9] Though CPRs are commonly recommended (for example, for primary prevention of cardiovascular disease, depression, transient ischaemic attack [TIA]/stroke, and breast cancer), there is little consensus on *which* CPR should be used when more than one exists.[7] For example, current National Institute for Health and Care Excellence (NICE) guidance on identification of depression simply suggests: 'Consider using a validated measure (for example, for symptoms, functions, and/or disability) to inform and evaluate treatment.'[10] There is, however, no subsequent evaluation of these measures, or recommendation on which to use. We must, therefore, equip ourselves to make these selections in the interests of providing the best possible patient care.

How do clinical prediction rules reach clinical practice?

In order to critically appraise CPRs, the first step is to have a basic understanding of how they reach clinical practice. There are three key stages that should be undertaken before a CPR is ready for clinical use:[11]

Development — identification of appropriate contributing factors and their subsequent weighting through statistical analysis (see Adams and Leveson for an approachable summary of these methods).[12]

Validation — testing to ensure the rule works in an external population, that is, does it still function outside the population that the tool was developed on?

Impact analysis — measuring the utility of the rule in clinical practice. This is a crucial step and one often overlooked.[8] Factors addressed include practicality for patient, clinician, and healthcare system, as well as clinical and cost-effectiveness.

These stages are reflected in the three papers that we have chosen at the beginning of this section. The first, by Snoeker *et al*, develops and shows internal validation (that is, in the development cohort) of a rule to aid detection of meniscal tears in primary care. The second, by Haasenritter *et al*, externally validates a CPR aiming to assist GPs in ruling out coronary disease as a cause of chest pain. The final paper, from van Ierland *et al*, compares several CPRs to identify serious infections in febrile children. Though it does not carry out full impact analysis, it does present some suggestions for how their impact might be increased in primary care.

Quick review of statistical concepts

The second critical step before we dive into appraising CPRs is a quick review of the terminology commonly used in these papers.

Sensitivity is the ability of the rule to identify correctly patients with the disease (true positives).

Specificity is the ability of the rule to identify correctly patients who do not have the disease (true negatives). These are usually presented as percentages. The ideal model will have high sensitivity and specificity.

Positive predictive value (PPV) is the chance that a patient with a positive test result truly has the disease. **Negative predictive value (NPV)** is the chance that a patient with a negative test result does not have the disease. These are also presented as percentages and, ideally, both will be high.

Positive likelihood ratio (LR+) describes how much the chance of having a disease increases when a test result is positive. **Negative likelihood ratio (LR–)** describes how much the chance of having a disease decreases when a test result is negative. If the test did not make any useful prediction the LR would be 1. If it makes a positive prediction it would be >1, and a negative prediction would be <1.

For the formulae behind these statistics, Table 1 contains more detail and a worked example using the data from Haasenritter *et al.* The example also helps to illustrate the important concept of cut-points. A

cut-point is the value chosen (usually by the developer of the CPR) to provide the *best* possible outcome for a test. One might expect that a cut-point would automatically be chosen to maximise the number of patients with the target condition who are identified, *and* minimise the number of patients falsely identified. However, there is often a real-world trade-off in identifying the best cut-point. This is seen in the example from Haasenritter *et al*, where the cut-point for the test has been set to make the test as safe as possible — by making sure not to miss patients with a cardiac chest pain (that is, choosing a high sensitivity of 89%). However, ensuring you capture as many as you can with the disease means you capture many without the disease (so the chance of having the disease if you are test-positive (PPV) is only 23%). In situations where the consequences of a missed diagnosis might be less serious it may be possible to tolerate more missed cases for higher accuracy.

In this case, a reasonable question would be whether a sensitivity of 89% is useful in this clinical context. A clinician relying on this CPR would miss 11% of cardiac chest pain. Ideally, for a serious condition one would like a higher sensitivity so that a negative result made the condition highly unlikely (a so-called ‘SnOUT’ with a high NPV. SnOUT means a high Sensitivity makes the test useful for ruling OUT disease. Its converse is SpIN, which you will be able to work out for yourself).

A receiver operating characteristic (**ROC**) **curve** is often used to illustrate this trade-off between the numbers of cases identified and numbers of cases missed. The area under the curve (**AUC**) is used as a numerical illustration of the overall accuracy of a test. In general, the closer the AUC is to 1, and the closer the line on the ROC curve is to the top left corner, the more accurate the rule. This is further illustrated in Figure 1, and a very approachable and more detailed article on this is available from Luke Oakden-Rayner.[13]

How to evaluate clinical prediction rules for primary care

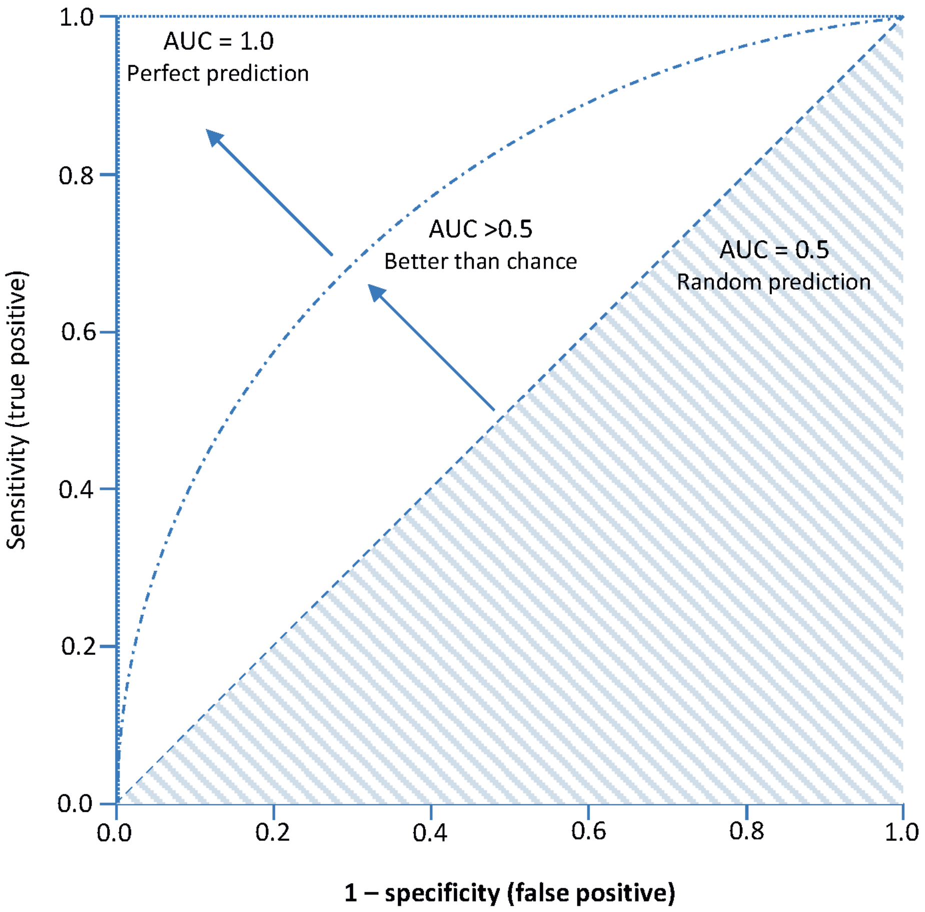
There are already many published approaches to evaluating CPRs, for example, from the Critical Skills Appraisal Programme (CASP).[14] Our approach has been to combine existing strategies with knowledge of primary care to provide a practical list of questions that will help you appraise whether a CPR is applicable to your practice.

Throughout this, we provide an explanation of the question, and then in italics an example of our appraisal of the clinical prediction rule for meniscal tears from Snoeker *et al*.

Table 1. Explanation of test results, with worked example using data from Haasenritter *et al*.

		Patients with coronary heart disease (CHD) as cause of chest pain		
		Condition positive	Condition negative	
Marburg Heart Score prediction of CHD as cause for chest pain	Predicted positive	True positive TP = 82	False positive FP = 270	**Positive predictive value (PPV)** = TP/(TP + FP) = 82/(82 + 270) = 23%
	Predicted negative	False negative FN = 10	True negative TN = 470	**Negative predictive value (NPV)** = TN/(FN + TN) = 470/(10 + 470) = 98%
		Sensitivity = TP/(TP + FN) = 82/(82 + 10) = 89%	Specificity = TN/(FP + TN) = 470/(270 + 470) = 64%	
		Positive likelihood ratio = sensitivity/ (1 – specificity) = 0.89/(1 – 0.64) = 2.44	Negative likelihood ratio = (1 – sensitivity)/ specificity = (1 – 0.89)/0.64 = 0.17	

Figure 1. Receiver operating characteristic (ROC) curve illustration.



Is the tool relevant?

Is this a clinically important problem to your patients?

The CPR must address a relevant question to your clinical practice.

Given how common knee pain is in primary care and the potential impact of a meniscal tear, as well as the possibility of good treatment, then the answer is 'yes'.

Has the model been developed for or applied in relevant settings?

Check the setting of the development and application, for example, primary versus secondary care, and which healthcare system in which country.

In this case, the setting was in primary care or physiotherapy. This seems relevant to our practice. We also noted that the study took place in the Netherlands. We performed a quick online search for information on the Dutch primary care system and were sufficiently happy that it was similar to our own.[15]

Is the tool practical?

Are the metrics required sensible?

Check if there are any predictors you think would be important that are missing, and whether they were considered.

At first glance, the selected list of variables seems sensible. Our only concern was that it misses out on the history of 'locking' that we might have expected, but a quick review showed that this was considered and found to be non-contributory.

Is it easy to collect the metrics required?

Look at what the tool requires you to collect. You aren't going to be performing positron emission tomography (PET) scans, highly specialised blood tests, or spending 3 hours on psychological tests.

We felt the metrics were easy to collect. Though we did not recognise the 'deep squat test', further reading suggested it would be easy to perform.

Could you see yourself using this tool in practice?

Probably the most important question. If it is impractical, expensive, or unacceptable to you or the patient, then it probably doesn't merit further review.

This tool seems acceptable to us as clinicians, as it supports a clinical decision that we might not make regularly and therefore need extra support with (and the consequences of over- or under-investigation are not minor). The feasibility of collecting the data seems good, though we are not keen on the score card provided. For it to be practical, it would need to be provided either as an online tool, or integrated into our clinical systems. We decided this problem was surmountable.

Now we have reviewed these screening questions, we can decide whether to proceed to a more detailed analysis.

Is the tool the best one for the job?

Check if subsequent analyses have been published, and whether there any competing CPRs.

A quick search reveals a systematic review and meta-analysis which suggests no single physical test is able to identify a torn tibial meniscus accurately.[16] *In addition to this, there do not appear to be any other potential CPRs. If there were, then we would want to review their relative predictive abilities, as is done for children with a fever by van Ierland* et al.

How were predictors and cut-point identified and weighted?

Potential predictor identification

Methods commonly used include systematic literature review and expert opinion. Check that the original pool of predictors is comprehensive.

This paper used a systematic literature review and other specific literature. We might have hoped to also see an expert review of the predictors identified to ensure no key variables were missing from the analysis.

Final predictor identification

This is the section that usually contains the most complex statistics. It is beyond our remit to provide details of all the possible methods used, but a reputable journal should have peer reviewed the statistics correctly. For those interested in further reading, a short accessible summary of techniques can be found in Adams and Leveson.[12]

The statistical methods here were particularly recondite (we had to look them up), but it's reasonable to assume the journal has done its job.

How was the cut-point chosen?

Look at the reasoning behind selection of a cut-point for prediction, and whether the emphasis has been on minimising false positives or false negatives (see statistics above in this section for more detail).

In this case, the decision was taken to choose a cut-point of 150, resulting in 46% specificity and 86% sensitivity, that is, 14% false negatives. There seems little justification for this except for a sentence in the Method suggesting a false negative rate of maximum 15% would be acceptable, as it is to be used as a screening tool. It would have been good to see some data on the acceptability of this number to clinicians and patients.

Have the results been interpreted correctly?

Look at the claims made by the authors, and whether you think they are justified. Consider whether the tool has been externally validated, and whether there has been any impact analysis.

The authors suggest that the tool should be a first step in selecting patients for magnetic resonance imaging (MRI) referral in primary care. They do not make it explicit that external validation and impact analysis are needed before it is put into practice, but they recognise the need for both of these in future research.

References

1. Kroenke K, Spitzer RL, Williams JBW. The PHQ-9: Validity of a brief depression severity measure. *J Gen Intern Med* 2001; **16:** 606–613.
2. Olesen JB, Torp-Pedersen C, Hansen ML, Lip GYH. The value of the CHA2DS2-VASc score for refining stroke risk stratification in patients with atrial fibrillation with a CHADS2 score 0–1: a nationwide cohort study. *Thromb Haemost* 2012; **107(6):** 1172–1179.
3. Kanis J A, Johnell O, Oden A, et al. FRAXTM and the assessment of fracture probability in men and women from the UK. *Osteoporos Int* 2008; **19(4):** 385–397.
4. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ* 2017; **357:** j2099. DOI: 10.1136/bmj.j2099.
5. Aalbers J, O'Brien KK, Chan WS, *et al*. Predicting streptococcal pharyngitis in adults in primary care: a systematic review of the diagnostic accuracy of symptoms and signs, and validation of the Centor score. *BMC Med* 2011; **9:** 67. DOI: 10.1186/1741-7015-9-67.
6. Wells PS, Anderson DR, Bormanis J, *et al*. Value of assessment of pretest probability of deep-vein thrombosis in clinical management. *Lancet* 1997; **350(9094):** 1795–1798.
7. Pluddemann A, Wallace E, Bankhead C, *et al*. Clinical prediction rules in practice: review of clinical guidelines and

survey of GPs. *Br J Gen Pract* 2014; DOI: https//doi.org/10.3399/bjgp14X677860.

8. Wallace E, Uijen MJK, Clyne B, *et al*. Impact analysis studies of clinical prediction rules relevant to primary care: a systematic review. *BMJ Open* 2016; **6(3):** e009957. DOI: 10.1136/bmjopen-2015-009957.
9. Search results. Medline trend. http://dan.corlan.net/cgi-bin/medline-trend?Q=%22clinical+prediction+rule%22+OR+%22clinical+decision+tool%22+OR+%22clinical+decision+rule%22+OR+%22Clinical+prediction+tool%22+ (accessed 22 Feb 2018).
10. National Institute for Health and Care Excellence. *Depression in adults: recognition and management*. CG90. London: NICE, 2016. https://www.nice.org.uk/guidance/cg90/chapter/1-guidance (accessed 22 Feb 2018).
11. McGinn TG, Guyatt GH, Wyer PC, *et al*. Users' guides to the medical literature: XXII: how to use articles about clinical decision rules. *JAMA* 2000; **284(1):** 79–84.
12. Adams ST, Leveson SH. Clinical prediction rules. *BMJ* 2012; **344:** d8312.
13. Oakden-Rayner L. Do machines actually beat doctors? ROC curves and performance metrics. 2017. https://lukeoakdenrayner.wordpress.com/2017/12/06/do-machines-actually-beat-doctors-roc-curves-and-performance-metrics/ (accessed 22 Feb 2018).
14. Critical Appraisal Skills Programme (CASP). CASP tools & checklists. http://www.casp-uk.net/casp-tools-checklists (accessed 22 Feb 2018).
15. Faber MJ, Burgers JS, Westert GP. A sustainable primary care system: lessons from the Netherlands. *J Ambul Care Manage* 2012; **35(3):** 174–181.
16. Hegedus EJ, Cook C, Hasselblad V, *et al*. Physical examination tests for assessing a torn meniscus in the knee: a systematic review with meta-analysis. *J Orthop Sports Phys Ther* 2007; **37(9):** 541–550.

Section 5

Bringing it all together: systematic reviews and meta-analyses

Marie-Louise E L Bartelink* and Niek J de Wit†
*Associate Professor of General Practice, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands
†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: https://doi.org/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: https://doi.org/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: https://doi.org/10.3399/bjgp11X593857
- Pritchett RV, Daley AJ, Jolly K. Does aerobic exercise reduce postpartum depressive symptoms? A systematic review and meta-analysis. *Br J Gen Pract* 2017; DOI: https://doi.org/10.3399/bjgp17X692525

Introduction

As clinical research aims at assessing knowledge relevant to clinical practice, relevant study questions for general practice provide answers about the diagnostic value of tests, the aetiology of disease, the prediction of prognosis of disease, and the effectiveness of a therapeutic intervention (DEPTh). Though results of repeated studies on such a research topic provide more information than a single one, they can also produce conflicting results that vary in direction and magnitude.

Therefore, it is useful to consider the collected evidence in a systematic review. A systematic review is a powerful instrument that allows conclusions to be reached about existing evidence in the scientific literature regarding a clinical dilemma. This process comprises finding all relevant studies, appraising their content, and summarising the measures of effect of the selected studies. Pooling data enables estimating the true association between the determinants and outcome solidly, as compared to a single study.

Systematic reviews are highly relevant to 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. A systematic review requires robust methodology, and this also means that conducting a systematic review is a research project on its own that requires a firm scientific effort and a systematic approach.

In a systematic review, studies are methodologically retrieved, appraised, selected, and summarised. The studies to be included are primarily selected on the basis of a detailed and pre-specified clinical question containing a well-defined and specific population, specific determinants, and a specific outcome. With this in mind, selection from the literature takes place using predefined eligibility and exclusion criteria. In the next step of the process, the methodology and outcomes of the retrieved studies are critically appraised, and a definite selection of studies to be included for further analysis is made.

As long as the inclusion criteria for subjects and the methodologies used in the eligible studies, particularly the measurement of outcomes, 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 'pooled result'. In this case, the systematic review quantifies the collected results of the individual studies into a new study outcome. This process is known as a meta-analysis. Although a meta-analysis is most commonly performed with randomised controlled trials (a therapeutic meta-analysis), it can also be used in other kinds of relevant clinical studies, for example, in a diagnostic meta-analysis.

By combining information from all relevant studies, meta-analyses can provide more precise estimates of the effects of health care than those derived from the individual studies included within a review. Meta-analyses facilitate investigations of the consistency of evidence across studies, and the exploration of differences across studies.

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 be used to calculate a pooled result. In this case, it is not the calculated outcomes of the various studies that are used. Instead, the individual patient data from all studies are entered into a new dataset, enabling an outcome to be calculated for this new population. In contrast to the outcome of a meta-analysis, which is based on pooling the results of the individual studies, the outcome 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, and that it enables pooling of subgroups, which are often not present in sufficient numbers to analyse within the individual datasets. An IPD thus provides a unique opportunity to identify subgroups that benefit more or less from an intervention. However, an IPD analysis is a major undertaking, as it requires collecting all the original research data and information from all studies and investigators involved.

It is important to mention that this section focuses on quantitative reviews. However, there is also an increasing interest in methods for synthesising qualitative studies, and many qualitative meta-synthesis with different approaches have been published.

For a 'state-of-the-art' systematic review, the following steps are required:

- defining a detailed research question;
- defining the eligibility criteria for the studies to be included;
- specifying the search terms and sources for the literature search strategy;
- searching the literature thoroughly, selecting relevant papers, and excluding those not meeting the eligibility criteria;
- critical appraisal and quality assessment of the studies identified;
- data extraction from the included studies; and
- synthesis of results and systematic presentation of the characteristics and findings of the included studies.

In writing and reporting, the authors should follow the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, www.prisma-statement.org), an evidence-based set of items for reporting in systematic reviews and meta-analyses. PRISMA focuses on the reporting of reviews evaluating randomised trials, but can also be used as a basis for reporting systematic reviews of other types of research. The AMSTAR-2 checklist (https://amstar.ca) provides a formal critical appraisal tool for systematic reviews that includes randomised or non-randomised studies of healthcare interventions, or both. For clinicians reading a systematic review, the following aspects are the most important when judging its quality and the potential impact of its conclusions for their practice. These will be discussed in more detail, using examples of the studies mentioned at the beginning of the section.

Clinical relevance of the review

The first question for clinicians is: 'Is the research question of the review relevant for my practice?' Astin *et al* conducted a systematic review of the diagnostic value of symptoms associated with colorectal cancer. Early identification of colorectal cancer is important for all physicians. Traditional alarm symptoms were mostly only identified in retrospect, and in secondary care. But the diagnostic importance of signs and symptoms are largely domain specific, and differ for patients presenting in primary, secondary, or tertiary care. Therefore, a systematic review focusing on the diagnostic value of symptoms and signs in patients suspected of colorectal cancer in *primary care* does have additional value, despite the fact that systematic reviews have been conducted in secondary care or mixed patient populations.

In the case of the systematic review by Jefferis *et al* of therapeutic trials in patients with conjunctivitis, one could question the need for pooling results, given the fact that all randomised controlled trials demonstrate that antibiotics work in cases of bacterial conjunctivitis. However, the IPD analysis enables a focus on the subgroup of patients in whom bacterial cultures had not been undertaken, which makes the systematic review very relevant for clinical practice.

The systematic review of Pritchett *et al* is of clinical importance, as current advice in guidelines to prescribe antidepressants or provide cognitive behavioural therapy is not well implemented for various reasons, and exercise may be an effective and easy alternative.

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, 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]

In all, 23 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. 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 for rectal bleeding accompanied by abdominal pain, diarrhoea, or 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.

Selection criteria for the studies

Before performing the literature search, the criteria for the studies to be selected need to be properly defined. Decisions about a relevant research question guide the choice for the patients under study (the study domain), the determinants (which intervention and comparison to be included, for example, in therapeutic trials, or which diagnostic tests in diagnostic studies), and the outcome of interest.

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; 80% (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% 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 which sources were explored, and for what period. Not only bibliographic databases such as MEDLINE, but also other sources, such as congress abstracts, or personal contacts with investigators or the pharmaceutical industry, could give access to relevant additional information. Relevant papers can also be identified from the references of retrieved papers (snowballing). If possible, authors should also try to identify registered trials of which the results were not published (which might be because of negative outcomes). This whole search process should be described in detail. As, for instance, in the paper of Astin *et al*:

> 'Further searches of ongoing studies included the European Organisation for Research and Treatment of Cancer, National Cancer Institute of Canada Clinical Trials Group, Cancer Research UK Directory of Funded Research, and the National Cancer Research Network. Reference lists of included studies were screened for relevance; personal literature collections and contacts of authors were also used.'

In addition, the outcome of the literature search needs to be specified in detail:

- How many studies were identified initially?
- How many were eligible for inclusion, and on what grounds were some excluded?
- Finally, how many of the eligible studies could be included to be further analysed in full text?

Search criteria need to be predefined and presented as a search string, so that the search is transparent and can be repeated by other researchers. The selection process needs to be done by two reviewers who independently scrutinise all the abstracts, thus avoiding bias. 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, such as the Jadad-score[2] or the risk of bias criteria by the Cochrane Collaboration (as in Pritchett *et al*, www.cochrane.org). In the diagnostic review by Astin *et al*, the QUADAS criteria1 were used. All these sets of criteria capture the most relevant methodological aspects and pitfalls of the study design. For example, important issues in RCTs are:

- Was the randomisation procedure and concealment of allocation adequately described?
- Was the intervention and/or outcome assessment blinded, if appropriate?
- Was the outcome assessment done using validated measures?
- Was the sample size pre-calculated, and was the loss to follow-up adequately described?
- Was the analysis done on all patients who started the intervention (intention-to-treat analysis as preferred), or only on those who completed the follow-up (per-protocol analysis)?

For every study that is included in the review, the score on the 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. For instance, as described in the review of Mugunthan *et al*:

> 'Two review authors independently reviewed and selected trials from searches. Two authors then assessed the trials, rated the study quality, and extracted relevant data. Disagreements were resolved through discussion with the third author. Trial authors were contacted to request missing data or to clarify methods where needed.'

It is also important that authors consider other risks of bias, such as publication bias. Publication bias is caused by the fact that negative outcomes have a lower chance of being reported and published, and thus the retrieved results in the literature are not complete. This emphasises the importance of trying to localise unpublished reports. There are statistical procedures (producing illustrative figures, 'funnel plots') to give more quantitative information about the possible presence of publication bias.

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 retrieved, 25 relevant abstracts were selected for detailed evaluation and 15 potential studies were full text evaluated, resulting in only three studies (with 615 patients) which 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% CI = 1.5 to 2.8, *P*<0.001; cessation: RR = 2.31, 95% CI = 1.3 to 4.2, *P* = 0.008). 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

In the next step, the results of the included studies that 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 considered all symptoms, so the pooled results may be based on different patient numbers. The statistical significance of both the results of the individual studies and of the pooled result is expressed in the 95% confidence level. The pooling process is often visualised graphically in a forest plot, which demonstrates the pooled result at a glance. The forest plot in the review of Pritchett *et al*, for example, shows the standard mean difference as found in the separate studies, as well as the pooled standard mean difference with its 95% CI. Besides, it gives information about the tests used to pool (random effects model or fixed effects model) and about the amount of statistical heterogeneity and the statistical measures involved, such as I^2 (with heterogeneity often referred to if I^2 <40% as small, if 30–60% moderate, if 50–90% substantial, and if >75% large).

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, in intervention, or in outcome assessment, they may be too clinically heterogeneous to pool the result. Beware of pooling under these circumstances. Another reason may be that the studies used different types of measurement to assess the outcome, such as dichotomous (yes/no), ordinal (for example, 5-point scale), or continuous (for example, 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.

Another example of a therapeutic meta-analysis (Pritchett *et al*)

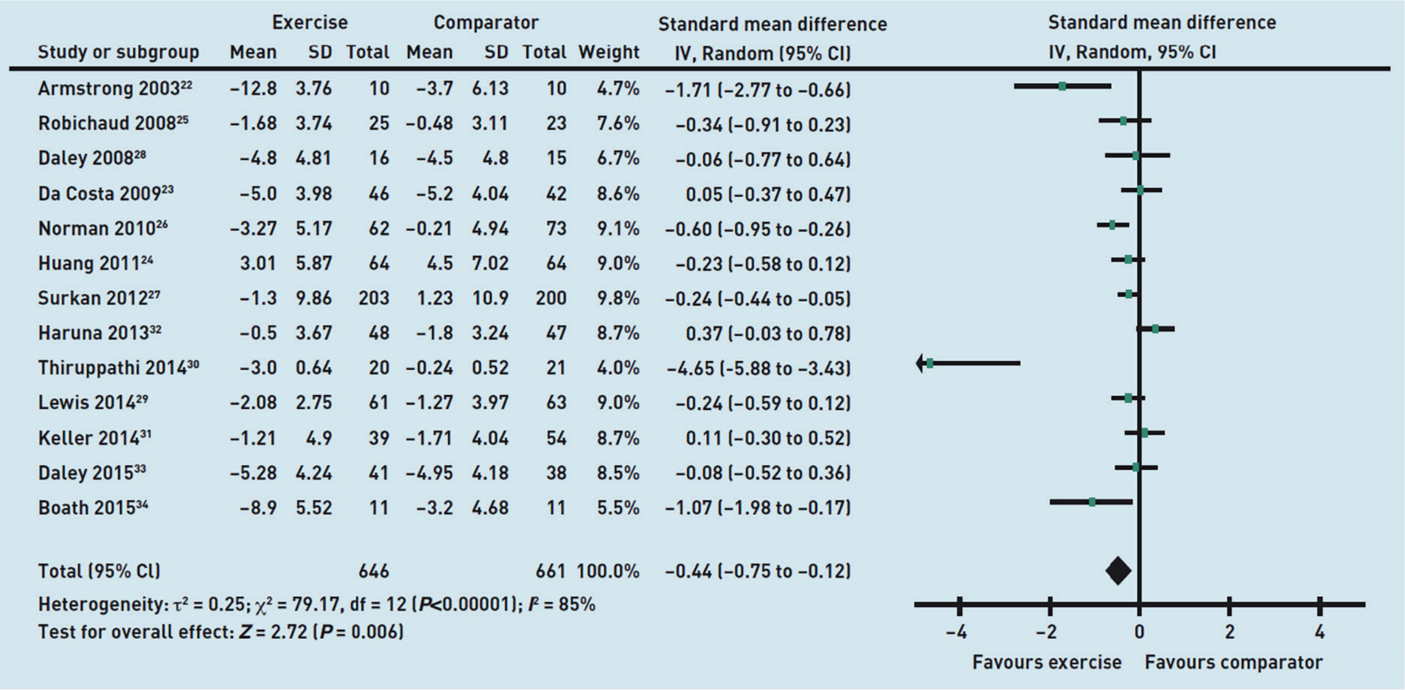
The authors aimed to systematically review the effectiveness of aerobic exercise on postpartum depressive symptoms. There was no restriction to study site or setting. The databases MEDLINE, EMBASE, Cochrane Library, PsycINFO, SportDiscus, Clinical Trials.gov, and the World Health Organization International Clinical Trials Registry Platform were searched. Titles and abstracts, then full text articles, were screened against inclusion criteria: randomised controlled trials (RCTs) measuring depressive symptoms in mothers ≤1 year postpartum, and interventions designed to increase aerobic exercise compared with usual care or other comparators. Included studies were assessed using the Cochrane Collaboration's risk of bias tool. Meta-analysis was conducted. Pre-planned subgroup analyses explored heterogeneity.

Thirteen RCTs were included, with 1734 eligible participants. Exercise significantly reduced depressive symptoms when all trials were combined (standardised mean difference –0.44, 95% CI = –0.75 to –0.12). Exploration of heterogeneity did not find significant differences in effect size between women with possible depression and in general postpartum populations, exercise only and exercise with co-interventions, and group exercise and exercise counselling.

The authors conclude that their systematic review provides support for the effectiveness of exercise in reducing postpartum depressive symptoms. Group exercise, participant-chosen exercise, and exercise with co-interventions may all be effective interventions. They state that these results should be interpreted with caution because of substantial heterogeneity and risk of bias.

Interpretation and clinical impact of the result

The final step is the interpretation of the effect size in relation to the clinical question under review. In a meta-analysis of randomised controlled trials of interventions 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, this statistical assessment of quantitative results 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. Thus, statistical significance does not always equal clinical significance.



	Exercise			Comparator				Standard mean difference
Study or subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random (95% CI)
Armstrong 2003[22]	−12.8	3.76	10	−3.7	6.13	10	4.7%	−1.71 (−2.77 to −0.66)
Robichaud 2008[25]	−1.68	3.74	25	−0.48	3.11	23	7.6%	−0.34 (−0.91 to 0.23)
Daley 2008[28]	−4.8	4.81	16	−4.5	4.8	15	6.7%	−0.06 (−0.77 to 0.64)
Da Costa 2009[23]	−5.0	3.98	46	−5.2	4.04	42	8.6%	0.05 (−0.37 to 0.47)
Norman 2010[26]	−3.27	5.17	62	−0.21	4.94	73	9.1%	−0.60 (−0.95 to −0.26)
Huang 2011[24]	3.01	5.87	64	4.5	7.02	64	9.0%	−0.23 (−0.58 to 0.12)
Surkan 2012[27]	−1.3	9.86	203	1.23	10.9	200	9.8%	−0.24 (−0.44 to −0.05)
Haruna 2013[32]	−0.5	3.67	48	−1.8	3.24	47	8.7%	0.37 (−0.03 to 0.78)
Thiruppathi 2014[30]	−3.0	0.64	20	−0.24	0.52	21	4.0%	−4.65 (−5.88 to −3.43)
Lewis 2014[29]	−2.08	2.75	61	−1.27	3.97	63	9.0%	−0.24 (−0.59 to 0.12)
Keller 2014[31]	−1.21	4.9	39	−1.71	4.04	54	8.7%	0.11 (−0.30 to 0.52)
Daley 2015[33]	−5.28	4.24	41	−4.95	4.18	38	8.5%	−0.08 (−0.52 to 0.36)
Boath 2015[34]	−8.9	5.52	11	−3.2	4.68	11	5.5%	−1.07 (−1.98 to −0.17)
Total (95% CI)			646			661	100.0%	−0.44 (−0.75 to −0.12)

Heterogeneity: $\tau^2 = 0.25$; $\chi^2 = 79.17$, df = 12 ($P<0.00001$); $I^2 = 85\%$

Test for overall effect: $Z = 2.72$ ($P = 0.006$)

Meta-analysis of the effect of exercise on depressive symptoms (standardised mean difference).

CI = confidence interval. SD = standard deviation.

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 Methodol* 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
Director Research Department of Medical Education, Deputy Director, University College London Medical School

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: https://doi.org/10.3399/bjgp11X601389
- Mitchell C, Dwyer R, Hagan T, *Mathers N.* Impact of the QOF and the NICE guideline in the diagnosis and management of depression: a qualitative study. *Br J Gen Pract* 2011; DOI: https://doi.org/10.3399/bjgp11X572472
- Verbakel N, de Bont A, Verheij T, et al. Improving patient safety culture in general practice: an interview study. *Br J Gen Pract* 2015; DOI: https://doi.org/10.3399/bjgp15X687865
- Banks J, Farr M, Salisbury C, et al. Use of an electronic consultation system in primary care: a qualitative interview study. *Br J Gen Pract* 2018; DOI: https://doi.org/10.3399/bjgp17X693509

What is qualitative research?

Qualitative research is interpretative, and as a research paradigm it encompasses a broad range of approaches. However, all methodologies share one common aim: to explore complex, contextual, and subjective phenomena examining in depth the features that help us develop a deeper, nuanced understanding of the object of our enquiry. Qualitative research examines 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. It can be used to explore a field, provide a theoretical understanding, or evaluate practice.[1] 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 researchers would contest the very notion.[2,3] However, published checklists exist in order to help authors, reviewers, and readers make a more informed decision about the quality of qualitative research.

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.[4] 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 research design that has methodological integrity. How to demonstrate this coherency and rigour will be discussed and illustrated by reference to four 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; and
- structures such as the workplace, its organisation, and its culture.

Typically, qualitative research uses first-person accounts (for example, interviews and focus groups), texts (for example, policy and free text on questionnaires), and observations (including video) as its main supply of data. This section will present 12 tips on how qualitative research can be shown to be systematic and robust, and yet remain 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 examine an issue in its fullest sense. A qualitative researcher will want to hear, read, or see people's everyday experiences and, through interpretative work, generate meaning. Research in this paradigm should:

- aim to reveal specific features of a phenomenon that until now have been hidden or tacit;
- be conscious of, and explicit about, the context in which the empirical work is carried out; and
- go beyond merely reporting, to also explore, reinterpret, assess, provide explanations, theorise, 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 concepts and theory?

Health sciences research is sometimes criticised for its insufficient use of theory but, increasingly, researchers are recognising its importance.[5] 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, as well as 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. This is a conceptual, ontological position that, broadly speaking, means that people's understanding of the world is gained through their active engagement in it. Social interactions are regarded as the catalyst to generate meaning or knowledge. The meanings people make are therefore unique, multiple, and dynamic. Importantly, the construction of meaning also applies to the research process and the researcher. The Blakeman *et al* paper rightly acknowledges that meaning is actively constructed by the researcher through their engagement with the data.

Verbakel *et al*'s paper used 'communities of practice'. This is an example of situated learning theory, a sociocultural perspective, which suggests that an important source of learning is through social encounters.

For further explanation about concepts and theory see Seale 2018.[6,7]

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,[8] 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 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?

As discussed, Blakeman *et al* took a social constructivist perspective and therefore an interpretative, qualitative methodological approach aligns well. They used ethnographic (observational) and interview data to explore the interaction between computer-based disease management templates and any impact on self-management dialogue in primary care consultations. To examine the relationship, they used a constant comparative method, which is drawn from a methodological approach known as grounded theory. In this methodology, theory emerges from the data; an inductive approach.

Verbakel *et al* used the theory of communities of practice to guide their methodology. They deliberatively deployed the central tenets of the theory in shaping their data analysis, using a deductive approach. Therefore, these studies provide good examples of research designs that demonstrate methodological alignment.

5. What are the characteristics of the participants?

The characteristics of the participants are an important consideration in qualitative research. Samples are usually not 'representative' of the population as a whole in the way that they often aspire to be in quantitative research. Rather, participants are 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.[9]

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?

Banks *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?

Banks *et al* purposefully sampled. They selected those practices who had been involved in the e-consultation pilot. They also sampled to include various organisational demographics, as well as including various types of participants (GPs, receptionists, and practice managers), thereby also generating a maximum variation sample. Mitchell *et al* and Banks *et al* showed the demographics of their participating practices. This allows the reader to make their own decision about whether what is reported will be applicable to other settings.

How many participants, interviews, focus groups, or observations is reasonable? This is a difficult question. Small sample sizes and a lack of 'representativeness' are criticisms frequently levelled at qualitative work. However, the aim of qualitative research is not to produce generalisable results, but to acknowledge the importance of subjective, unique accounts. The in-depth nature of this work means that it can be legitimate to study very small numbers.

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 usually 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, though 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 also of 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 one 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, for example narrative research, are also possible for interviews and observations. 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, and so on), was it applied correctly?
- Was software used and, if so, how did it contribute to the process?

Grounded theory has been used by Blakeman *et al.* Grounded theory, which was devised by Glaser and Strauss,[10] is an approach typically used when very little is known about the empirical field and the research is exclusively exploratory. It is an inductive approach. The researcher invites informants to talk freely about the area of study and tries not to control the outcome, by avoiding 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 usually have some prior knowledge and wish to ask certain questions.

Verbakel *et al* illustrates a deductive approach, using theory to shape data gathering and analysis.

8. How should data analysis be presented?

Presenting qualitative work in a meaningful way, demonstrating the complexity of issues, and including consensus as well as outlying voices, is a major challenge, particularly given the restrictions of journals word counts. Tables of coding themes and/or models may help the reader to quickly get a better idea of the scope of the analysis. The increased flexibility inherent in online publishing is a great advance for qualitative researchers and publishers alike.

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* presented interactions, and those by Mitchell *et al* and Banks *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;
- presenting views that are confirming as well as disconfirming;
- clearly identifying limitations of the research design; and
- a logical research design that permeates the entire research process.

What approaches have these three papers taken?

See chapter 32 in *Researching society and culture* by Seale for further guidance.[11]

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, with their 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, and 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 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 undermining 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'. Verbakel *et al* applied a theory to data.

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 still many significant ethical issues inherent in a qualitative approach that 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 developed in a discussion section showing how they are related to the bigger picture, existing literature, and possibly other contexts or practitioners, and how they contribute to the existing empirical field. Research should contribute to building upon our existing knowledge base, as well as to developing our theoretical and methodological understanding. High quality work will demonstrate how it can impact upon practice, change healthcare outcomes, and benefit patient care.

- 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?

Three of the 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 requires a detailed description of each step in the research design, from theory to analysis through to empirical and theoretical conclusions. 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, and presentation of results, 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. Sidhu K, Jones R, Stevenson, F. Publishing qualitative research in medical journals. *Br J Gen Pract* 2017. DOI: https//doi.org/10.3399/bjgp17X690821.
2. Barbour RS. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *BMJ* 2001; **322(7294):** 1115–1117.
3. Mays N, Pope C. Qualitative research in health care. Assessing quality in qualitative research. *BMJ* 2000; **320(7226):** 50–52.
4. Park S, Griffin A, Gill D. Working with words: exploring textual analysis in medical education research. *Med Educ* 2012; **46(4):** 372–380.
5. Reeves S, Albert M, Kuper A, *Hodges BD*. Why use theories in qualitative research? *BMJ* 2008; **337:** a949.
6. Seale, C. Philosophy, politics and values. In: *Researching society and culture*. 4th edn. London: Sage, 2018.
7. Silverman D. Research and theory. In: Seale C, ed. *Researching society and culture*. 4th edn. London: Sage, 2018.
8. Clough P, Nutbrown C. *A student's guide to methodology: justifying enquiry*. London: Sage: 2002.
9. Robson C. *Real world research: a resource for social scientists and practitioner–researchers. 2nd edn.* Oxford: Blackwell Publishing, 2002.
10. Glaser B, Strauss A. *The discovery of grounded theory*. Chicago, IL: Aldine, 1967.
11. Seale C. Research quality. In: *Researching society and culture. 4th edn*. London: Sage, 2018.

Section 7

Critical evaluation of a health economic journal article

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

Relevant *BJGP* paper:

- Oppong R, Smith RD, Little P, *et al*. Cost effectiveness of amoxicillin for lower respiratory tract infections in primary care: an economic evaluation accounting for the cost of antimicrobial resistance. *Br J Gen Pract* 2016; DOI: https://doi.org/10.3399/bjgp16X686533

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); and
- 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, in relation 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. QALYs take into account the individual’s preference for health.

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 Care 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.

Authors and study question

The authors, Oppong *et al,* are well-known health economists with a wealth of knowledge. The title is clear, telling us what the paper is about and should gain the attention of the readers, since antimicrobial resistance is an important discussion point now in health care. Antibiotics are still used inappropriately, and antimicrobial resistance is on the rise. This study identifies and evaluates the cost-effectiveness of antibiotic use for acute cough/lower respiratory tract infection (LRTI), and the implications of the costs associated with antimicrobial resistance. The authors also refer to cost–utility analysis in the methods; the QALY can be used in both CEA and CUA, depending on whether you are showing effectiveness or utility. The use of both terms in this paper is confusing.

Introduction

The introduction to the paper clearly outlines the issues and the gaps in knowledge on this topic. We are made to understand that LRTI is a major source of morbidity and that antibiotics are not beneficial in the treatment of all patients, but it is difficult to identify which patients need treatment. Antibiotic use has also been found to increase the cost of treatment because of antimicrobial resistance. Therefore, including the cost of resistance would affect the cost-effectiveness of antibiotic treatment of LRTI.

Method

The patient data are collected from a parallel, randomised trial in which patients receive either amoxicillin or placebo. This large, multicentre international trial is presented in another paper.[1] The economic evaluation took the perspective of the health system; that is, there will be no personal costs to the patient or costs associated outside the healthcare system included in this cost analysis. This also excludes any societal costs. The authors collected data on resource use from the case report forms from the primary care physicians, and patient report outcomes from patient diaries. The paper clearly reports and defines all the costs and outcomes (EQ-5D-3L) that will be used in the analysis. Since the timeframe for the study was a 4-week period, no discounting was required. NICE's recommendation thresholds of between £20 000 and £30 000 per QALY were used to judge the cost-effectiveness of the intervention.

The paper states that there are no good or accurate estimates for the cost of resistance. This is unfortunate, as it would be beneficial to have a published and validated value for the cost of resistance to use in this study. The authors instead estimate the cost of resistance using an Excel 'what if' analysis. They also estimate the cost of resistance per prescription using literature values from the US, EU, and globally for the cost of resistance and the number of antibiotic prescriptions dispensed in each region examined (US, EU, and globally). This cost is at a societal perspective which is at odds with the health service perspective of the amoxicillin costs. This cost can be added to the costs incurred in the intervention arm of the study.

An all-important sensitivity analysis was also performed to determine whether amoxicillin is cost-effective in patients aged ≥60 years old. The variation of the cost of resistance could also be seen to be a sensitivity analysis.

Results

Table 1 gives the results of the cost collection for the intervention and placebo groups. The health outcome measures are also given. The amoxicillin group had a slightly higher cost than the placebo, which would account for the cost of the antibiotic in the intervention arm. The patient outcome measure, the EQ-5D-3L, shows that, at baseline, patients are similar in health quality. However, the placebo arm significantly reduces in health quality in the first and second weeks compared to the intervention arm, where patients are given amoxicillin. The health quality returns to no difference between the groups in weeks 3 and 4. This demonstrates that LRTI infections may get better with antibiotics, but the benefit is short lived compared to the risks of increased antimicrobial resistant infection. The message of a 'cold lasting a week and then you will feel better' may be shown here.

The paper goes on to calculate the cost-effectiveness with and without resistance costs. The results are reported as the incremental cost-effectiveness ratio (ICER). In this type of analysis, the authors are reporting the differences in cost and benefits between the two interventions based on the costs recorded, rather than on the total costs and outcomes in each arm. This type of analysis, also known as marginal analysis, is common and more straightforward to perform than a full economic analysis, as it focuses on the differences between the arms, rather than trying to count every cost and outcome.

The paper states that the cost difference between the amoxicillin and placebo group is €3.04 (£2.42). However, it is difficult to determine where this cost comes from. Table 1 shows the cost difference between the two arms as €2.43. The calculation of the former cost is not illustrated, so it is hard to determine what costs are used and what the difference is between the table and text costs, but it may be that the units in Table 1 should be £ Sterling. The difference in QALYs between the two groups is also reported in the text of the paper as 0.00037. However, no clear explanation of how this value was calculated is presented in the paper. These two values are used subsequently in the cost per QALY calculation before resistance costs are added.

Costs of resistance are given in the paper and, again, it is unclear where these costs come from or how they were calculated. The cost difference with resistance costs are given, greatly increased, but the reader is left unsure how these were done. The different cost of resistance per prescription calculated earlier from each region (US, EU, and globally) are used. Once more, it is unclear how and why we are getting the values we are seeing in this paper. The paper illustrates that with the cost of resistance included in the cost-effectiveness analysis, the costs greatly exceed the NICE cost-effectiveness threshold of £20 000 to £30 000, unless the EU data for cost of resistance are used.

The sensitivity analysis found that patients who were ≥60 years of age were more costly in the amoxicillin group, and that treatment was slightly less effective. Therefore, amoxicillin was less cost-effective in this age group. The paper could have also mentioned that antibiotic resistance increases with age. This is the age cohort effect, where the probability of antimicrobial resistant infections increases with age because of increased time to resistant environments and increased probability to exposure to antibiotics, compared to a younger cohort.[2] The use of three different costs for resistance could also be classed as a sensitivity analysis, as it shows that the cost of resistance used has a major effect on the cost-effectiveness of amoxicillin in LRTI.

Discussion

The Discussion section of the paper reiterated that the cost-effectiveness of amoxicillin for a LRTI is of marginal benefit, and probably not cost-effective in most cases. The strengths and limitations presented are logical and well explained.

Because the cost of resistance is not clearly or accurately established, the analysis in this study is difficult to reproduce. The cost of resistance that is used is not clearly shown, so that the reader is convinced that the added cost to the intervention arm is somewhat in the 'ball park' of what it should be. In the author's calculation, three different sources of the cost of resistance per prescription were estimated, but even then, because of study design, not all costs were collected in the literature values.

Key points in the critical appraisal of health economic studies[3]

- Ensure that the question is clear, and the correct analysis is used.
- State the perspective of the analysis — is it patient, NHS, or societal?
- Specify all costs, and be clear where they come from.
- Ensure outcomes are clearly defined.
- Undertake sensitivity analysis on any data that may have uncertainty.
- Ensure that the conclusion relates to the question asked.

Conclusion

Overall, this is a good demonstration of difficulties in health economic analysis of interventions because of assumptions that have to be made and costs that have to be added from various sources. It is helpful to be clear where the costs are calculated from, so that the reader is convinced of the authority and robustness of the entire calculation of final costs for comparison. This is particularly of importance when the cost of amoxicillin treatment changes between the tables and the text.

References

1. Little P, Stuart B, Moore M, *et al*. Amoxicillin for acute lower respiratory tract infection in primary care when pneumonia is not suspected: a 12-country, randomised, placebo-controlled trial. *Lancet Infect Dis* 2013; **13(2):** 123–129.
2. Steinke DT, Seaton RA, Phillips G, *et al*. Prior trimethoprim use and trimethoprim-resistant urinary tract infection: a nested case–control study with multivariate analysis for other risk factors. *J Antimicrob Chemother* 2001; **47(6):** 780–787.
3. Elliott R, Payne K. Essentials of economic evaluation in healthcare. London: *Pharmaceutical Press*, 2004.