Sources of evidence for assessing the safety, efficacy and effectiveness of medicines

June 2017

The Academy of Medical Sciences

Acknowledgements and disclaimer

The Academy of Medical Sciences is most grateful to Sir Michael Rutter CBE FRS FBA and to the members of the working group for undertaking this study. We thank the Academy's Officers, Council members and staff, the external review group, study observers, as well as our Fellows and all those who have contributed through the call for evidence, stakeholder meetings, or by providing oral evidence. We thank the study secretariat led by Dr Claire Cope.

The Academy is grateful to Arthritis Research UK, the British Heart Foundation, the British Pharmacological Society, the British Society for Immunology, the Medical Research Council, the Naji Foundation, and the National Institute for Health Research Health Technology Assessment Programme for their financial contribution to the workstream as a whole. Funding from a core grant from the Department for Business, Energy and Industrial Strategy to the Academy was also used to support this project.

This report is published by the Academy of Medical Sciences and has been endorsed by its Officers and Council. Contributions by the working group were made purely in an advisory capacity. The members of the working group participated in an individual capacity and not as representatives of, or on behalf of, their affiliated hospitals, universities, organisations or associations. Their participation should not be taken as endorsement by these bodies.

All web references were accessed in February 2017.

Sources of evidence for assessing the safety, efficacy and effectiveness of medicines

Contents

Executive summary

Recent high-profile media debates over the use of statins to prevent heart disease, anti-virals to treat influenza, and the human papilloma virus (HPV) vaccine to prevent cervical cancer, have brought to the fore discussions around the validity of the methods of collecting and analysing evidence, and the way that it is presented. To provide some clarity to the debate, the Academy of Medical Sciences convened a working group to explore the strengths and limitations of different methods of assessing evidence and to describe evolving approaches that are being, or have been, developed to address some of the major limitations of current methodologies. This working group project is part of a wider workstream considering how we can all best use evidence to judge the potential benefits and harms of medicines.

Different sources of evidence have individual strengths and limitations

All approaches to the evaluation of evidence have strengths and limitations, which we explore in depth in this report. The type of evidence, and the methods needed to analyse that evidence, will depend on the research question being asked. Randomised controlled trials (RCTs) are usually the best way of generating robust evidence about the effects of treatments. Their main strengths are their ability to minimise the effect of biases and confounding owing to their use of control groups, randomisation to different treatment groups, and often their use of blinding techniques (that, for example, prevent participants, and frequently those involved in administering the trial, from knowing who is receiving which treatment). Well-designed RCTs are ordinarily the only method that can detect, in an unbiased manner, moderate but clinically important effects that are directly caused by the treatment under investigation. RCTs by their nature involve interventions being tested on participants, who in practice are often people with no more than one disease. It is not uncommon for RCTs to have relatively strict participant eligibility criteria meaning that it may be challenging to generalise results from RCTs to wider patient populations and to know the potential impact on routine clinical practice. Individual RCTs also generally have limited ability to detect harms that are rare or have a long-latency (i.e. side effects that become apparent after the trial has ended). It should also be noted that RCTs may not be necessary when the treatment effect is very large.

In these circumstances, observational studies (where researchers observe and collect information about participants without actively intervening in a way that would affect the participant) can also play a useful role in gathering evidence about medicines. The interpretation of the results of observational studies is however limited by the lack of control for bias and confounding that RCTs provide (via randomisation), and they often do not use blinding. However, they may provide important information on the safety, and sometimes the effectiveness, of medicines, and on the generalisability of results to different groups and the wider population. Observational studies can be particularly good at detecting large effects on rare outcomes, which do not occur frequently enough to allow their reliable assessment in RCTs. More effort is needed to identify methods and best practice for combining information from RCTs and observational studies to maximise the amount of evidence available to decision-makers.

Alternative trial methods or analyses have been developed to address some of the limitations of conventional trial

designs. However, they typically still all use randomisation as their basis for minimising bias and confounding. Evolving approaches for analysing observational data, including propensity scores, natural experiments and the use of genetic variants in Mendelian randomisation, can also be used to strengthen research findings from these studies. We encourage researchers and regulatory authorities to continue to work together to resolve the limitations of different approaches to generating and evaluating evidence.

When collecting evidence to determine the effects of a medicine, researchers need to consider and communicate the strengths and limitations of their study design or of the existing evidence that they are analysing. For example, when designing and completing an RCT to test the effect of a new treatment, they should consider how the results might be extrapolated to the population of individuals with that disease or condition. Equally, when reporting the results of an observational study, researchers should consider biases and/or confounding as a possible explanation of their results. Whenever possible, researchers should always report clinically- and patient-relevant outcome measure endpoints to ensure that studies are as meaningful as possible.

Hierarchies of evidence for causal effects of medicinal products have been developed to help assess the strength of evidence (i.e. robustness) from different study types. They do so by ranking the different study designs according to how well the properly conducted studies of that design produce credible results. These place RCTs, or meta-analyses of RCTs, at the top of the classification. Such hierarchies are useful 'rules of thumb' but should not be used prescriptively. Importantly, they should not be a substitute for good judgement in the critical appraisal of the evidence and the rigour of the study from which the evidence has been generated.

Enhancing the reporting, accuracy and reliability of clinical study results

Concerns have been raised about the lack of transparency, standards, accuracy, and reliability of clinical study results. We support the efforts that have been put in place to:

- Increase the transparency of clinical trials, including the compulsory registration of clinical trials and the publication of summary results in public registries.
- Enhance the reporting of clinical trials, epidemiological studies, meta-analyses and systematic reviews via endeavours such as the CONSORT, STROBE, and PRISMA guidelines.

These initiatives are an important step towards improving standards in the dissemination of clinical study results. However compliance should be better and more consistently incentivised, and enforced by research institutions, funding bodies and journals. We also support the push for a cultural shift within the scientific community that ensures that all results are published (or made publicly available), including null, 'negative' or inconclusive results.

There is growing concern about the impact of a perceived lack of reliability of research findings (implied by a failure to reproduce some key research findings), which can hinder the scientific process, delay translation into clinical applications, and waste valuable resources. We welcome the pre-registration of clinical trials that has helped to increase the reliability of research by ensuring that all relevant trials undertaken can be found and that researchers give the necessary thought to the methods of design and analysis before the study is started. This approach should be explored for more basic research involving human subjects and in epidemiological observational studies, where it is not currently common practice.

Questions have been raised about whether the funding source (e.g. industry or government) for the collection and evaluation of evidence within the academic sector affects what research is conducted, how it is undertaken, whether and how it is disseminated, and how data are analysed. These concerns extend to the study design and analysis, data holding and access, personal payments from funders to academic researchers, and trial registration. We believe that consideration of these factors might help address some of the concerns about the validity of clinical research, and recommend that the Academy looks further at the principles governing relationships between academia and industry.

Evaluation of evidence for clinical practice

The first step in assessing the validity and applicability of the results of clinical studies for patients is to assess the quality of the study and the data. This requires consideration of whether the size of the effect is large enough to have been reliably detected in the study design that was used, and whether the findings are plausible (based on sound biological principles) and reliably demonstrate a causal link between the treatment and the outcome. Meta-analysis of multiple RCTs can inform decisions about whether the results can be generalised across a broad population, providing the choice of RCTs included in the meta-analysis is unbiased. Judgements about whether the evidence is 'fit for purpose' in a given clinical scenario and how variables such as age can moderate the effects of a treatment are key in balancing the benefits and harms of a given medicine.

Future challenges and opportunities

We have explored a number of future challenges and opportunities linked to the evaluation of evidence for the benefits and harms of medicines. These include the need for new strategies to evaluate the safety, efficacy and effectiveness of novel treatments in rare diseases (where small numbers of geographically scattered patients do not allow for large RCTs, although the convening power of collaborative research networks and patient groups may help in mobilising patients globally) and in emergency situations (where the spread of the disease, geographical dispersion and mortality rates can pose practical, ethical and methodological problems). Future challenges also include the smaller number of participants in stratified medicine trials, new sources of data, and increasingly the applicability of study results to populations where more patients will have more than one illness (multimorbidities). For example, the increasing availability of clinically-relevant data collected outside the context of conventional RCTs (such as data from primary and secondary care and disease registries) presents a potential opportunity to explore the benefits and harms of medicines outside of the controlled conditions of RCTs. To take advantage of these data, their use should be explored by researchers and regulators, and in health technology assessments (HTAs). Attention should be paid to approaches such as those in Scandinavia that enable access and linkage of data while also protecting data subjects' personal data and their desire for confidentiality and anonymity.

In the future it will not be realistic to carry out head-to-head comparisons of all new and existing products for particular conditions or diseases, due to the sheer number of possible comparisons. Evolving approaches, such as network meta-analysis of randomised trials, will be increasingly important to compare the benefits and harms of medicines where direct comparisons are impossible.

Communicating the benefits and harms of medicines

Patients, supported by their healthcare professionals, should have access to the best available evidence about the benefits and harms of treatment options. It is the responsibility of healthcare professionals, regulatory authorities and both the general and scientific media to present information on the level of risk or size of effects clearly, accurately and objectively, without causing undue concern or unsubstantiated reassurance. We recommend the use of measures of differences in absolute risk and benefit (an estimate of the likelihood that an outcome will occur in a group of people under clearly specified treatments). A number of studies have found that these measures are much more informative than others such as relative risk and attributable risk.

Transparency throughout the decision-making process

The concept of 'perfect evidence' is illusory. A substantial amount of research is conducted before a medicine has been approved for widespread use, but knowledge about a particular medicine will continue to increase once it has been approved and is used in much larger and heterogeneous populations. As this evidence accumulates, it can provide further information about the circumstances under which a medicine is effective. Decisions have to be made by: regulatory authorities about whether to license a new medicine; healthcare professionals considering recommending different treatments; and patients about their treatment strategy. These decisions must balance the need for sufficient high-

quality evidence to make an informed decision on the use of a medicine with a societal desire for faster access to medicines. Transparency around the decision-making processes and information on which decisions are made is needed at all levels, be it in research, regulation or clinical practice. This will allow wider society to judge whether decisions are made based on robust enough evidence.

Recommendations

Recommendation 1 (Chapter 2)

We recommend that absolute risk or absolute risk difference is always presented alongside any measure of relative risk or attributable risk so that the level of risk or size of intervention effects can be properly understood. This applies to the general and scientific media, regulatory agencies, and scientists.

Recommendation 2 (Chapter 3)

We recommend that funding bodies ensure appropriate support for research in the areas of: how to deal with the difficulties created by premature termination of RCTs (in particular estimation of treatment effect); and the extent to which under-representation of certain groups in these studies really affects the generalisability of the study results. Appropriate support should also be provided for trials that are sufficient in scale and duration to achieve the pre-specified outcomes.

Recommendation 3 (Chapters 3 and 6)

We recommend that all those evaluating evidence should pay particular attention to factors that are likely to affect the validity and applicability of the results, including:

- **Biological plausibility** – are the findings based on sound biological principles?
- **Generalisability** – do the results extend to the treatment populations of interest?
- **Effect size** – is the size of the treatment effect large enough to be reliably detected in the study design that was undertaken and/or is the sample size large enough to detect a clinically important treatment effect if it exists?
- **Causality** – do the results reliably demonstrate a causal link between the treatment and the observed effect or do they merely suggest a correlation or association?

Decision-makers should use their judgement as part of the critical appraisal of the evidence, to ascertain whether the evidence they are presented with is 'fit for purpose'. This judgement is central to the assessment of the benefits and harms of medicines, as well as for the evaluation of research findings when considering their application in clinical practice.

Researchers should be aware that these factors will be influenced by determinants such as bias, confounding, moderating variables, choice of comparator and endpoints, participant attrition and adherence to treatments, and the 'placebo' and 'nocebo' effects, which should

therefore be carefully considered in the study design. Alternative trial methodologies and analytical approaches (including Bayesian thinking) should be given due consideration, as should the investigation of outcomes that are of particular importance to patients.

Recommendation 4 (Chapter 4)

Electronic health records, research databanks and disease registries are valuable sources of so-called 'real world' data and we recommend that their use should be explored by researchers and regulators, and in Health Technology Association (HTA) assessments. In developing an approach for access to and linkage of data in the UK, attention should be paid to approaches such as those in Scandinavia, where the use of unique personal identifiers, supportive infrastructures and appropriate governance have enabled the straightforward linkage of data, and anonymity is protected by making data available to researchers in an irreversible encrypted fashion.

Recommendation 5 (Chapter 5)

We recommend that guidelines such as CONSORT, STROBE and PRISMA, among others, be comprehensively adopted and that funding bodies require their grant awardees to adhere to these reporting guidelines. We encourage research institutions and journals to provide incentives for their use and to better enforce their adoption.

Recommendation 6 (Chapter 5)

We recommend that higher education institutions and research institutes make training on research integrity, which should include elements relating to unconscious bias in biomedical research, mandatory and that they assess the effectiveness of such training programmes.

Recommendation 7 (Chapter 5)

We support the registration of RCTs and mandatory publication of their protocols by researchers as a means to help ensure that so-called 'negative', null or inconclusive trial results become publicly available, and to enhance research reproducibility. We recommend that similar methods be adopted for observational epidemiological studies exploring the effects of treatments, where currently such an approach is lacking. We encourage journals to make registration of such observational epidemiological studies and publication of their protocols a condition of publication, as is the case for clinical trials.

Recommendation 8 (Chapter 5)

We recommend that the Academy, through the oversight group, looks further into the principles governing relationships between academia and industry to address concerns about the validity of academic research funded by industry.

Glossary of abbreviations and acronyms

ABPI: Association of the British Pharmaceutical Industry

ASTERIX: Advances in Small Trials dEsign for Regulatory Innovation and eXcellence

BBSRC: Biotechnology and Biological Sciences Research Council

BMI: body mass index

BPS: British Pharmacological Society

CBT: cognitive behavioural therapy

CF: cystic fibrosis

CHD: coronary heart disease

CONSORT: CONsolidated Standards Of Reporting Trials

COX: cyclooxygenase

CPR: Central Population Registration

CPRD: Clinical Practice Research Datalink

CVD: cardiovascular disease

EACPT: European Association for Clinical Pharmacology & Therapeutics

EHR: electronic health record

EMA: European Medicines Agency

ESH: European Society of Hypertension

EU: European Union

FDA: Food and Drug Administration

GP: general practitioner

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HDL: high-density lipoprotein

HPV: human papilloma virus

HRT: hormonal replacement therapy

HTA: Health Technology Assessment

IDeAL: Integrated DEsign and AnaLysis of small population group trials

IMI: Innovative Medicines Initiative

InSPiRe: INnovative methodology for Small PopulatIons REsearch

IPD: individual patient data

IPTW: inverse probability of treatment weighting

ISD: Information Services Division

IUPHAR: International Union of Basic and Clinical Pharmacology

LDL: low-density lipoprotein

MHRA: Medicines and Healthcare products Regulatory Agency

MMR: measles, mumps, rubella

MRC: Medical Research Council

NIBSC: National Institute for Biological Standards and Control

NICE: National Institute for Health and Care Excellence

NIHR: National Institute for Health Research

NNT: numbers needed to treat

NNTB: number needed to treat to benefit

NNTH: number needed to treat to harm

NTD: neural tube defects

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROTECT: Pharmacoepidemiological Research on Outcomes of Therapeutics

RA: rheumatoid arthritis

RADR: Recognising Adverse Drug Reactions

RCT: randomised controlled trial

RDD: regression discontinuity design

SSRI: selective serotonin reuptake inhibitor

STROBE: Strengthening the Reporting of Observational Studies in Epidemiology

TTP: trusted third party

1. Introduction

Evidence that is generated, analysed, interpreted and appraised using the most robust and reliable methods appropriate for the research question under investigation is critical for the advancement of medical science, increased knowledge of disease processes, and the development of safe and effective treatments (see Box 1 for an overview of why, how and by whom evidence is used in the biomedical sciences). Further, effective communication of evidence is central to informed decision-making. It is therefore crucial that evidence is appropriately communicated to the relevant stakeholders, including the general public and patients who stand to benefit from it. Misrepresentation of the evidence can lead to questions about its validity, patients discontinuing prescribed medication or taking medication they may not choose to take if fully informed, and, in extreme cases, to the undermining of trust in the bodies that are associated with it.

The Academy has become increasingly concerned by high-profile media debates about the evidence underlying decisions about treatment options (such as ongoing controversies around statins, Tamiflu and the human papilloma virus (HPV) vaccine), and the debates around the validity of the different ways of collecting and analysing evidence. These concerns have been echoed to us from across the research community. The Chief Medical Officer for England also wrote to the Academy asking that we consider these issues in further detail. Accordingly, the Academy launched a new workstream in 2015, led by an oversight group chaired by Professor Sir John Tooke FMedSci, to explore how we can all best use evidence to judge the potential benefits and harms of medicines (Academy of Medical Sciences 2015a). As part of this workstream, the Academy brought together a working group, chaired by Sir Michael Rutter CBE FRS FBA, to specifically consider the

methods of assessing evidence. The objective of the working group study was to explore how evidence that originates from different sources is examined to inform decisions about the benefits and harms of medicines.

Although many of these concepts are well established, in light of recent debates about the evidence underlying the use of medicines, this report aims to collate all of the relevant concepts in order to provide some clarity. Although the principles are often well understood, they are not necessarily easy to apply in practice. The recommendations apply primarily to those who undertake research or use research findings in their clinical work. There are several issues that will be addressed by the oversight group. These are identified as the report proceeds.

Box 1. Why, how and by whom evidence is used in the biomedical sciences

Evidence that informs the development and use of medicinal products is gathered during the process of conducting biomedical research. Such evidence is generated for a variety of reasons. Firstly, it helps to further understand the biological processes that underpin our health and how these go wrong in disease. Secondly, it is used to identify novel therapeutic targets or treatment modalities, and to support the development of safe and effective new treatments. Lastly, it is used to inform clinical practice – either directly, or indirectly by feeding into the production of clinical guidelines – and to facilitate the adoption and use of new interventions in healthcare systems. Evidence is also more broadly used to inform government policies and the wider population on the harms and benefits of medicines.

Evidence is used by an array of stakeholders, including the following:

- Researchers, to inform their research projects and advance scientific knowledge.
- Industry, including pharmaceutical, diagnostic and biotechnology companies, to develop safe and effective new products that improve patient health.
- Healthcare professionals, to guide their treatment options and bedside practice.
- Regulatory authorities, such as the Medicines and Healthcare products Regulatory Agency (MHRA), to decide whether a new treatment is safe and effective, and can be licensed for use in humans.
- Health Technology Assessment (HTA) agencies, such as the National Institute for Health and Care Excellence (NICE), to inform the development of clinical guidelines and to input into their measures of cost-effectiveness.
- Policymakers and government, to ensure that health-related policies have the best outcomes for wider society.
- Patients and the public, to inform their healthcare choices.

1.1 Scope and terms of reference

The Academy's study on the sources of evidence was launched in June 2015. The study's terms of reference were to:

1. Explore the strengths and limitations of results and conclusions that originate from different study types or data sources to answer a range of research questions as well as to evaluate the risks and benefits of medicinal products.[1,2]
2. Start to consider the implications of the working group's findings for the communication of evidence, including the availability of the evidence around the risks and benefits of medicinal products. Initial concepts were to be further explored within the other elements of the overall workstream.

To this end, the working group study aimed to develop a list of broadly applicable principles relating to the presentation, interpretation and weighting of evidence to enable a range of stakeholders (including patients, the public, healthcare professionals and the media) to better consider the risks and benefits of medicinal products. It also aimed to draw on case studies and examples to illustrate these principles; for instance, the use of statins (the cholesterol-lowering drugs), hormone replacement therapy, and vaccination. Case studies that do not relate to medicinal products (such as surgical interventions, medical devices, screening procedures, and so on) were considered to be beyond the remit of this study.

The group did not seek to address all areas of contention, nor to replicate the work performed by the MHRA and NICE. Further, it did not consider how evidence is used to inform the cost-effectiveness of medicinal products.

Through these terms of reference, the working group sought to address the following questions:

- What are the research issues that all study types should aim to address?
- How do these affect traditional approaches to judging the benefits and harms of medicines?
- How do these issues impact on the evaluation of research findings when considering their application in clinical practice?
- What are the evolving approaches that seek to address some of the outstanding limitations of traditional approaches?
- What are the wider issues that should also be considered when evaluating evidence?

1.2 Conduct of the study

The study was conducted by a working group, chaired by Sir Michael Rutter CBE FRS FBA, which included expertise in methodology, epidemiology, regulation of medicines, general and clinical practice, social sciences, and communication. A lay patient representative was also present on the group to ensure that the work was accessible and gave due consideration to the patient perspective. Dr Melanie Lee CBE FMedSci kindly provided an industry perspective on the content of the report. Observers from the MHRA and NICE also joined the discussions but did not have sight or input into the conclusions and recommendations (see **Annex I** for a list of working group members and observers).

The Academy issued an open call for written evidence in June 2015 to inform the workstream as a whole. Submissions were received from a wide range of organisations and individuals. The responses to the call for written evidence, for which permission to publish was received, are available on the Academy's website (Academy of Medical Sciences 2016a). The evidence submitted was analysed and assimilated alongside many published papers. Additional consultation was achieved through:

- A meeting with Professor Sir Rory Collins FRS FMedSci (British Heart Foundation Professor of Medicine and Epidemiology and Head of the Nuffield Department of Population Health, University of Oxford) and Sir Michael Rawlins FMedSci (Chair, MHRA and author of the 2008 Harveian Oration) on 26 May 2015.

---

[1] By medicinal product we mean '*any substance or combination of substances presented as having properties for treating or preventing disease in human beings*' or '*any substance or combination of substances which may be used in, or administered to, human beings, either with a view to restoring, correcting or modifying physiological functions by exerting a pharmacological, immunological or metabolic action, or to making a medical diagnosis*'. Please see 'A guide to what is a medicinal product: MHRA Guidance Note 8' for further information: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/398998/A_guide_to_what_is_a_medicinal_product.pdf

[2] Animal models may be helpful in identifying the physiology and other biological effects of medicines. However, the issues involved are quite complex. Accordingly, it was decided that their consideration was outside the remit of this report, which is concerned with the evaluation of medicines in human participants.

- A meeting with the Academy of Medical Royal Colleges, the British Heart Foundation, the British Medical Association, the Medical Research Council (MRC), the National Institute for Health Research (NIHR) HTA Programme, the Royal College of General Practitioners and the Royal College of Physicians on 17 June 2015.
- A meeting with NICE and the MHRA on 17 June 2015.
- A meeting with the Association of the British Pharmaceutical Industry (ABPI)'s Innovation Board on 21 September 2015.
- An evidence-gathering workshop held jointly with the Wellcome Trust on 'Evaluating evidence in health' on 21 October 2015. A report of this workshop has been published separately (Academy of Medical Sciences 2016b).
- Oral evidence sessions with Claire Murray (Joint Head of Operations, Patient Information Forum) and Tracey Brown (Director, Sense about Science), who provided evidence to the working group on 5 November 2015. Notes of these sessions are available on the Academy's website (Academy of Medical Sciences 2016a).

Annex II provides further details of the contributors to the study.

This report was iterated with the workstream's oversight group to allow the wider work package to inform, and be informed by, the working group's study. It was reviewed by a group appointed by the Academy's Council (see Annex I) and has been approved by the Academy's Council.

We thank all those who contributed to this study. We are grateful to Arthritis Research UK, the British Heart Foundation, the British Pharmacological Society, the British Society for Immunology, MRC, the Naji Foundation, and the NIHR HTA Programme for their financial contribution to the workstream as a whole.

1.3 Overview and audience

In this working group report we describe the principal research issues that all study types should seek to address (Chapter 2) and explore how traditional approaches to the evaluation of evidence are affected by these research issues (Chapter 3). We review the range of new or modified research methods that may be employed to study drug safety, efficacy and effectiveness, explore the study of drugs used in rare diseases and emergency situations, and briefly consider future challenges, namely in stratified medicine, so-called 'real world' evidence and patients with multiple illnesses (Chapter 4). In Chapter 5, we discuss issues apposite to the reproducibility and reliability of research, publication bias, and working with industry. We also briefly address concerns that have been expressed about possible over/underuse of medicines, including that in the context of vaccination. A book published by Nicolas Rasmussen illustrates some of the issues that we explore in this chapter (see Box 2). In Chapter 6, we consider the evaluation of research findings in the context of their application to clinical practice. Finally, we seek to draw conclusions and make a number of recommendations about future actions (Chapter 7).

The report is primarily aimed at a biomedical audience involved in the generation, analysis, interpretation, evaluation, and communication of evidence, including:

- Researchers in academia and industry
- Healthcare professionals
- Regulatory agencies
- Funding bodies
- Scientific journals

The report will also be of interest to the general media, as key communicators of research and evidence to the public, and to patients and the general public, who are directly affected by decisions on the use of medicines based on evidence from research.

Box 2. Nicolas Rasmussen's book, 'On Speed: The many lives of amphetamine'

Nicolas Rasmussen's book entitled 'On Speed: The many lives of amphetamine' describes several aspects that are explored in this report. For example, regarding the alleged medicalisation of social problems, the book describes the remarkable story of the rise, fall and surprising resurgence of amphetamines since they were first introduced as antidepressants in the 1930s. Although the focus of the book is on amphetamine, it also deals with the parallel story of other drugs used to treat depression, such as selective serotonin reuptake inhibitors (SSRIs) and the use of amphetamine derivatives to treat hyperactivity/attention deficit disorder as well as drugs used to aid dieting. One key point relevant to this report is that amphetamine psychosis, the most serious adverse effect, was discovered by an astute clinician who noted the phenomenon quite separately from any type of drug trial. The study of phocomelia caused by thalidomide showed much the same, with the potential link between them arising from clinical observations of the phenomenon rather than from any form of trial (see Academy of Medical Sciences 2007 for an account of the phenomenon).

2. The research issues that all study types should address

Research that involves human participants to investigate the safety, efficacy and effectiveness of medicinal products can be broadly divided into two study types: experimental research and observational research.[3] As the names suggest, in experimental studies interventions are actively assigned and tested in participants by researchers, whereas in observational studies the effects of interventions are observed but researchers do not actively assign subjects to treatment groups. Both of these approaches have respective strengths and limitations that are further discussed in Chapter 3. However, there are core research issues that all study types should seek to address in order to provide the highest quality and most reliable evidence possible.

This chapter outlines the strengths that should be aimed for and the limitations that should be minimised in all study types – be they experimental or observational in nature. For the purpose of this report, these have been grouped into the following topics that we think should be considered in the design, analysis, interpretation, evaluation and communication of studies by researchers and other stakeholders:

- Bias
- Confounding

[3] Efficacy refers to the performance of an intervention in 'ideal', controlled circumstances such as an RCT, whereas effectiveness describes an intervention's performance under 'real world' clinical conditions (Singal, Higgins & Waljee 2014).

- Blinding
- Internal/external validity and relevance
- Moderating variables
- Absolute risk, relative risk, attributable risk and number needed to treat
- Causality
- Choice of comparator
- Participant attrition and adherence to treatments
- The 'placebo' and 'nocebo' effects
- Surrogate endpoints

Many of the issues explored in this chapter are dealt with comprehensively in other publications; in particular, the Academy of Medical Sciences' 2007 report on 'Identifying the environmental causes of disease: how should we decide what to believe and when to take action?' and Rawlins' Harveian Oration of 2008 'De Testimonio: On the evidence for decisions about the use of therapeutic interventions'. In this chapter, we provide a summary of the key elements that should be considered when designing, conducting, analysing and, importantly, appraising clinical studies.

2.1 Bias

Bias is described as a systematic distortion of the estimated outcome away from the 'truth' and can be caused by problems in the design, conduct, or analysis of a study (Altman *et al*. 2001). Biases may be introduced at multiple stages throughout the conduct of a study and may be unconscious. Where biases in studies go unrecognised and systems are not in place to mitigate them, they can lead to erroneous results by over or underestimating the effects of the intervention.

Sources of bias in the design and conduct of research studies include the following (Higgins & Green 2011a; Higgins *et al*. 2011; Jadad & Enkin 2007):

- **Selection bias:** systematic differences between the baseline characteristics of the groups that are compared.
- **Performance bias:** systematic differences in how groups are treated in the study, or how they are exposed to other factors apart from the intervention of interest.
- **Detection bias:** systematic differences between groups in how outcomes are determined.
- **Attrition bias:** systematic differences between groups in withdrawals from a study, which can lead to incomplete outcome data. Attrition bias can lead to selection bias as described above.
- **Reporting bias:** systematic differences between reported and unreported findings.
- **Experimenter bias:** subjective bias towards a result expected by the experimenter.
- **Confirmation bias:** the tendency to search for, interpret, or favour information in a way that confirms one's preconceptions or hypotheses.
- **Ascertainment bias:** systematic distortion of the results due to knowledge of which intervention each participant received.

Bias can arise from a variety of sources including financial (e.g. source of funding) and non-financial factors (e.g. commitment to a scientific belief or career progression). Conflicts of interest are a widely recognised source of bias.

Studies should be designed to minimise important sources of bias in the conduct and analysis of the research as far as it is practicable and 'cost effective' (some biases will be prohibitively expensive to eliminate in certain circumstances). Researchers should be encouraged to consider potential sources of selection, performance, detection, and attrition bias at an early design stage and make considered decisions regarding strategies to mitigate these. They should also consider the potential impact of bias on the validity and relevance of the results. Sources of bias should be systematically considered when the results are published, alongside the measures put in place to minimise their effect, so that they can be appropriately taken into consideration in the critical appraisal and interpretation of the study results. It is also important that the source of funding, as a potential cause of bias, be disclosed for all studies. Where appropriate, studies should be designed and conducted by multidisciplinary teams (including patients) as this can further reduce the likelihood of errors in study design that increase the risk of bias, or that individuals' personal prejudices influence and bias a study.

2.2 Confounding

In observational settings, when the effect of an intervention is studied in two or more groups, two effects come into force: (1) the treatment effect, which is specifically related to the effect of the administered intervention on the outcome; and (2) the effect of other characteristics that differ between the groups being compared and that also affect the outcome (Academy of Medical Sciences 2007). This second effect is called confounding, because it literally 'confounds' (confuses) the estimate of the treatment effect. Where both the treatment and confounding effect exist in a study, what is assumed to be the treatment effect is in reality a combination of both the treatment and confounding effects unless methods are used to minimise the confounding effect.

Confounding arises from characteristics that influence both the proposed treatment and the outcome it is thought to affect, and so can explain (some or all of) the association between the two (Academy of Medical Sciences 2007). For a factor to be a confounder, it must satisfy three conditions: first, it must differ between the comparison groups; second, it should influence the outcome of interest; and third, it must not be part of the treatment or its causal effect (Altman *et al*. 2001). Confounding factors affect a study's findings by introducing variables (often unmeasured or unknown) that result in comparison groups that differ by more than the intervention under investigation (e.g. baseline characteristics, prognostic factors, or concomitant interventions). This makes it difficult to determine whether the apparent effect of the intervention is actually due to the intervention or to another characteristic, which is associated with both the intervention and the outcome. For example, if a group receiving a treatment is in general healthier than the corresponding control group, then it will be difficult to determine whether improvements in health are due to the treatment under investigation or due to the healthier lifestyles of the treatment group.

Therefore, in order to limit the distorting effects of confounding, potential confounding factors should be identified and minimised as far as is practically possible in the study design. As for sources of bias, the nature of known confounders and the measures used to minimise them should be detailed in publications to allow judgements to be made as to their possible effect on the results. Approaches exist to assess and adjust for measured confounders, and are discussed in **Chapter 4** (Mamdani *et al*. 2005; Normand *et al*. 2005). The challenge is adjusting for unknown confounders.

Properly randomised allocation of patients to comparison groups (e.g. to receive a drug or placebo) and blinding (see below) are the key ways to control selection bias and confounding, and should be implemented in the study design wherever possible.

2.3 Blinding

Blinding is the practice of keeping trial participants, care providers, data collectors, and/or those analysing the data unaware of which intervention is being (or has been) administered to which participant, to control for their possible confounding effects (Altman *et al*. 2001; Academy of Medical Sciences 2007). A common application is 'double-blinding', whereby both the participants and assessors are blinded to the intervention. As many members of a research team (e.g. those logging the study data, those analysing the data, or those undertaking the relevant laboratory tests) can and should (in many circumstances) be blinded, researchers should explicitly detail who was blinded to the allocation of patients to comparison groups, and by what means. Blinding is a particularly effective way of minimising potential biases – including selection, performance, detection, experimenter, confirmation and ascertainment bias – as it limits the extent to which individuals' inherent (and perhaps unconscious) biases can come into play. As we highlight in the next chapter, blinding is more common in RCTs than in observational studies (see **Chapter 3**).

Blinding is an important methodological feature that should be implemented wherever possible in all study types, during data analysis as well as data generation. There are a number of individuals who it might be important to blind from which comparison group the participants are in, including the participant, the individuals collecting and recording the outcome data, and those analysing the data. The integrity of the blinding process should also be monitored as it can be unintentionally compromised. There are clearly situations where blinding (of at least participants) is not pragmatically or ethically possible (e.g. if a drug is compared to a psychological intervention such as cognitive behavioural therapy (CBT), or if a drug is compared to an instrumental or operative procedure). To assess the quality of the blinding process, transparency around who has been blinded to knowledge about which comparison group people are in and how this blinding was achieved is required, as well as the explicit recognition of any limitations that have arisen due to a lack of blinding.

2.4 Internal/external validity and relevance

Internal validity describes the extent to which a study produces valid findings for the participants investigated and the population to which inference is being made (Academy of Medical Sciences 2007). It depends crucially on the control of bias and other sources of confounding. For example, findings that are investigated in a trial of Asian women aged 20–30 years old are likely to be applicable to women of the same age and ethnicity outside the context of the trial.

Conversely, *external* validity (also known as generalisability) refers to the extent to which study results are applicable to patients beyond the population that the participants who took part in the research came from. The generalisability (exportability) of results is dependent on the degree to which the participants involved in a study are similar to those for whom the treatments are intended to be used (Cartwright 2007; Rothwell *et al.* 2005; Woolcock 2013). For example, studies conducted only in participants within a specific age range might not provide results that are generalisable to people outside that age group. Explicit inclusion and exclusion criteria provide further information on who the results are likely to generalise to. Broad inclusion criteria might help reassure clinicians, policymakers and the public that the results apply to most people who might use the treatment, but they also often mean that more participants are needed in the study, which might increase the variation in the results; therefore it is common to undertake studies, particularly RCTs, in more select groups first.

The applicability of study results to a wider group of patients and their relevance to clinical practice should be considered when evaluating evidence. When designing studies, researchers should also take this issue into consideration by exploring whether the external validity of their study could be increased, or at least recognising the limitations of their research and outlining areas for further investigation in additional groups.

2.5 Moderating variables

Moderating variables – that is, variables that can influence the effect of treatment and thereby the generalisability (exportability) of study results – include age, sex, gender, ethnicity, comorbidities, prognostic factors, and other drugs, among others (Rawlins 2008). Although there is evidence of striking differences among participants used in research and those that will ultimately receive the treatment, the effects of moderating variables have rarely been systematically evaluated. **There is a need for further research into the effects of moderating variables on study results.**

Clinicians often have to make treatment decisions for the likelihood of benefit for individual patients when the results are only available from group comparisons (Rothwell *et al.* 2005; Thompson & Higgins 2005). Sometimes there is an attempt to deal with the possible influence of moderator variables through multivariate analyses, which statistically model possible effect modifications to get an average effect for all groups in the study, but do not explore whether effects differ between groups. However, when there is a known pre-specified group of special interest (such as there has been with respect to the use of statins in low-risk groups) a focus on robustly establishing effects in that group is important. With respect to statin use in low-risk groups, such analysis has been undertaken (for further details see Cholesterol Treatment Trialists' Collaboration 2012a; Ray *et al.* 2010; Taylor *et al.* 2013).

Pocock *et al.* (2002) have drawn attention to the importance of clearly pre-specifying the subgroup analyses that will be undertaken in an RCT to avoid spurious findings from so-called 'data-dredging' (see **Chapter 5**). This potential problem is nicely illustrated by the ISIS-2 large RCT, which showed that treatment with both streptokinase and low-dose aspirin was effective in preventing premature death in patients who had experienced a myocardial infarction (ISIS-2 Collaborative Group 1988). Following reviewer comments the editors of The Lancet asked the authors to undertake 40 subgroup analyses, none of which had been planned at the start of the trials. The authors were reluctant but compromised by undertaking the requested subgroup analyses and any additional ones they wished. They examined effects by astrological star sign and found that aspirin appeared not to have been effective in those who were Libra or Gemini. This demonstrates that spurious findings are likely to be obtained if ad hoc subgroup analyses are performed. Individuals can have more confidence in subgroup findings if they are pre-planned.

In addition, Pocock *et al.* (2002) highlighted that it is not valid to declare that effects differ between subgroups because in one subgroup there is a small p-value and in another there is not. Rather, a statistical test for interaction (difference in effect between subgroups) is required alongside examining the magnitudes of effect in each group (with their confidence intervals as an initial exploration) and ensuring adequate statistical power for detecting differences from the start of the

study. Further, within any one trial, tests of effect modification (differences between subgroups) will be different depending on whether the absolute risk or the relative risk of the treatment effect is tested (see below).

2.6 Absolute risk, relative risk, attributable risk and number needed to treat

Three measures are commonly used to quantify an intervention's effects (Academy of Medical Sciences 2007):

- **Absolute risk**, which is the absolute probability that a given outcome will occur in an individual exposed to a treatment.
- **Relative risk**, which describes the fraction by which the risk after exposure to an intervention is greater or lesser than that amongst those not exposed.
- **Attributable risk**, which is a measure of the overall effect of a new treatment on the incidence of the disease in the general population.

In these definitions, 'risk' is used as a statistical measure of probability and can represent either benefit or harm. In addition, absolute benefit is described as the change in absolute risk by using the new treatment.

The distinction between these terms is not necessarily understood widely outside of the field of epidemiology and, if taken out of context, the reporting of these measures can be misleading.

The differences between these ways of presenting results can be illustrated by thinking of a hypothetical example of patients being randomised between a new treatment and an existing treatment in an RCT. At the end of this study, if the percentage of people who are cured in the new treatment group is 6% and in the existing treatment it is 3%, then the absolute risk difference is 6% minus 3% = 3% (or 3 per 100). That is, if 100 patients were given the new treatment on average 6 would get better, which is on average 3 more than would have got better if all 100 had been given the existing treatment. The relative risk is 6 divided by 3 = 2 or a relative increase of 200%, meaning that twice as many people are cured with the new treatment compared with the existing treatment. The relative risk is usually a larger number than the absolute risk difference, which may be why researchers, funders and the media often prefer to emphasise it.

It is crucial that these measures of risk are put into appropriate context when reported in both the scientific and general media to avoid any misleading representation of the strength of an intervention's effect.

Recommendation 1

We recommend that absolute risk or absolute risk difference is always presented alongside any measure of relative risk or attributable risk so that the level of risk or size of intervention effects can be properly understood. This applies to the general and scientific media, regulatory agencies, and scientists.

Following an influential article by Laupacis *et al*. (1988), numbers needed to treat (NNT) was highlighted as the best method for comparing different treatment approaches. Two such figures exist:

- **Number needed to treat to benefit (NNTB)**: Describes the average number of patients who will need to be treated to get an additional positive outcome (National Institute of Health and Care Excellence 2016). For instance, if the NNTB is 20, then 20 patients on average would have to be treated to ensure an additional positive outcome in one patient. The lower the NNTB, the more effective the treatment, as fewer individuals will need to be treated to obtain an additional positive outcome in a patient.
- **Number needed to treat to harm (NNTH)**: Can be calculated for adverse outcomes, with the meaning being similar to that for NNTB. For example, a NNTH of 200 would suggest that 200 people would need to be given the treatment for one of them to experience a harmful outcome. For the NNTH, the higher the number the lower the risk of harmful outcomes.

Mathematically, NNTs are the reciprocal of the absolute risk reduction (for NNTB) or absolute risk increase (for NNTH). Many clinicians came to accept the NNT as the best available measure. Recently, however, it has been suggested that there needs to be some reconsideration of the problems involved with NNTs and the ways in which those might be dealt with (Roose *et al*. 2016). First, the most easily interpretable NNTs concern categorical outcomes (e.g. response vs. no response) but most deal with the treatment of disorders that function dimensionally (e.g. different disease severity). This would be true, for example, for the treatment of depression. The second difficulty is that there are particular problems with respect to using the NNT when there is a marked placebo effect. Also, pooling across RCTs requires that the placebo response is measured in the same way across all trials, but that is rarely the case. Further, NNT figures derived from clinical trials might not be directly relevant to clinical decision-making because the supposed control conditions used in the RCTs may not actually exist in standard practice. Roose *et al*. (2016) suggested that clinical utility of NNTs is likely to require effectiveness studies that include treatment conditions resembling actual clinical practice. This remains a challenge to be met and, in the meantime, we urge caution in the use of NNTs whilst not recommending that they necessarily be abandoned.

2.7 Causality

Many clinical studies aim to determine whether a specified intervention has an effect on an outcome of interest – in other words, whether the intervention *causes* an increase or decrease in the observed outcome. In his seminal paper, 'The Environment and Disease: Association or Causation?', Bradford Hill highlighted key considerations to help decide whether a causal link between an intervention and a potential outcome can be inferred – these are commonly referred to as the Bradford Hill considerations (see **Box 3**) (Hill 1965). We focus here on three features that are particularly important in the testing of causality. First, data are needed to determine the temporal order – the cause has to precede the effect. These can be prospective or retrospective data, but must allow the researcher to establish the sequence of events. Second, causal relations between the intervention and the outcome measured are best established when the control and treatment groups differ only by the intervention. To this effect, randomisation of participants – assigning them to treatment or control groups on the basis of chance alone – in conjunction with blinding is the only effective mechanism as it minimises selection bias and increases the likelihood that confounding factors (both known and unknown) are distributed in a balanced manner between the groups. Third, evidence on biological plausibility can support a causal inference. With respect to drug effects, evidence on biological mediation is particularly useful.

It is also helpful to determine the *mediating* variables by which an intervention is likely to have an effect, as this can assist with causal inference in both RCTs and observational studies. For instance, the effects of statin treatment on lowering low-density lipoprotein (LDL) cholesterol levels have been shown to mediate the reduction in the risk for cardiovascular disease (CVD) (Cholesterol Treatment Trialists' Collaboration 2005). Mediation is viewed as a matter of major interest because it can be used to determine which elements accounted for a risk or protection effect. The usual starting point is a strong direct effect (best measured by a path coefficient, MacKinnon & Fairchild 2009) of the hypothesised causal variable. Mediation analysis (again using path coefficients) tests the effects of the supposed causal variable on the postulated mediator, the effects of the mediator on the outcome and the overall indirect path going through the mediator. Finally, the analyses test whether the residual direct effect after mediation dropped to a non-significant level (Rutter & Pickles 2016). There is no doubt that the aims of mediation analyses are sound but it has become clear that the statistical demands are high (Rutter & Pickles 2016).

Causal Bayesian Networks, mixed treatment, and network meta-analysis are approaches that provide a way of bringing together RCTs in meta-analysis that allows inferences about interventions that may not have been evaluated directly against each other (Li *et al*. 2011; Mills *et al*. 2013; Spirtes 2010; Zhang 2008). Such causal networks are a highly attractive meta-analysis innovation but they present methodological challenges that require further methodological research. It is clear that in the future the sheer number of comparisons needed will make it unrealistic to carry out head-to-head comparisons. Given the importance of the issue, further study of such meta-analysis innovations should be a high priority.

Box 3. The Bradford Hill considerations

In 1965, Bradford Hill set out the following nine considerations when establishing a causal inference between an intervention and a potential effect (Hill 1965):

1. Strength: The larger the association, the more likely it is causal. For example, the causal link between tar and mineral oils and scrotal cancer was reasonably inferred as chimney sweep mortality from scrotal cancer was found to be 200 times that of other workers who did not work with such substances. However, a small association does not necessarily mean that there is no causal effect.
2. Consistency: A causal inference will be strengthened if it has been reproduced by different individuals, in different places, circumstances and times. Consistent findings when an experiment has been reproduced exactly as previously do not invariably strengthen the association, as similar results might be obtained due to replication of the same error in methodology.
3. Specificity: Causality is more likely if the association is limited to specific individuals, or sites or types of disease. The more specific the association, the more likely there is a causal link between an intervention and a proposed effect.
4. Temporality: The effect must occur after the cause if there is a causal relationship between the two. For example, if a diet causes a particular condition, the condition would appear following the change in diet; whereas if a particular condition affected dietary habits, then the condition would precede the change in diet.
5. Biological gradient: A relationship between the strength of the intervention and the strength of the effect (a dose-response) should be carefully considered. In certain instances, the greater the exposure, the greater the effect; in others, the inverse might occur – the greater the exposure, the smaller the effect. In other instances, there might not be any such proportionality, and the simple exposure to the intervention, regardless of its strength, causes the said effect.
6. Plausibility: The existence of a plausible biological mechanism to explain the association between an intervention and an effect can aid the causal inference. The extent of biological knowledge at a given time can, however, limit the determination of a biologically plausible mechanism.
7. Coherence: The causal inference should be coherent (in other words, not completely incompatible) with the well-established facts of the natural history and biology of the disease. Evidence from the laboratory or animal experiments can support findings in humans, but the lack of such evidence cannot nullify epidemiological observations in man.
8. Experiment: Experimental, or semi-experimental, evidence can help to support a causal inference. For example, does reducing or halting the intervention, or exchanging it for something else, impact on the outcome of interest?
9. Analogy: Where there is a precedent, the effect of similar factors can be considered in the process of examining causality. For example, the effects of thalidomide and rubella should be borne in mind when examining similar evidence of another drug or another viral disease in pregnancy.

2.8 Choice of comparator

2.8.1 Placebo and active comparators

Traditionally, experimental trials employ a placebo – a control intervention similar in every respect (e.g. size, colour, form, duration and administration of treatment) to the active treatment except that it contains no active ingredient. This is to ensure that study participants and those involved in administering treatments are blinded to which participants are receiving which treatments. Placebos might be the sole comparator (drug vs. placebo) or may be central to the comparator arm (e.g. drug A + placebo vs. drug A + drug B; or drug A [pill] + placebo [injection] vs. placebo [pill] + drug B [injection]). As highlighted previously, blinding is a particularly effective way of minimising potential biases and is a key methodological feature that should be implemented wherever possible.

When evaluating evidence, it is important to consider whether the choice of comparator is appropriate. For example, if there are already available treatments (e.g. standard of care), a study comparing a new treatment to already available treatments will be far more informative than a study comparing it to a placebo. Indeed, clinicians and patients will want to know whether the new treatment is better than the already available (or the best available) alternative treatments. This has led to the suggestion that three-arm trials may need to become the norm in late Phase II and Phase III trials (see **Box 5**) where alternative treatments are already available. That is, one arm deals with the placebo, a second arm deals with already available treatments and a third arm with the new experimental treatment (Khan & Brown 2015).

Further, in the situation where there are available alternative treatments of known efficacy, the use of a placebo control that involves stopping the use of effective treatments raises ethical issues, as there is a firm expectation that study participants should not be disadvantaged by taking part in a clinical trial. This concern does not arise if there are no alternative treatments with demonstrated efficacy. In both industry and academic research, there is always a consideration of ethical issues when designing clinical studies, and a placebo-controlled arm will never be adopted – or not without standard of care as integral to the placebo-controlled arm – if it deprives participants of an effective treatment. In **Chapter 3**, we detail alternative RCT designs that could be considered when the use of a placebo raises ethical issues.

2.8.2 Bayesian approaches

Bayesian approaches to the analysis of trial and other data should also be considered.[4] Briefly, Bayesian methods have been developed on the premise that all probabilities are conditional (Salsburg 2002), and can represent our uncertainty about facts (e.g. the average benefits of a treatment) as well as future events. Bayesian approaches interpret data from a study in the light of external evidence and judgement, taking into account prior plausibility of hypotheses (Spiegelhalter *et al.* 1999). This means they are ideal as a basis for combining data from multiple sources into a single analysis, possibly including some subjective judgements where data are thin. Such 'decision models' are already widely used in cost-effectiveness analyses for NICE, and allow, for example, an assessment of the probability that one treatment is superior to another (Rawlins 2008). However, especially in the absence of previous RCT data, there is an inevitable subjectivity regarding the prior probabilities. Thinking in Bayesian terms is often desirable, but translating it into a mathematical probability is complicated.

2.9 Participant attrition and adherence to treatments

2.9.1 Participant attrition in research studies

Attrition, due to participant withdrawal from a research study or loss to follow-up, can introduce bias and distort the estimates of the overall effect of treatment. In particular, bias can be introduced where attrition leads to an imbalance of factors that are prognostic for the outcome of interest between comparison groups. For example, if body mass index (BMI) is linked to the treatment effect, then a high drop-out of participants with a low BMI in one of the comparison groups will introduce bias into the study and affect the results. Further, adjusting for additional variables that were not specified in advance – such as those identified when exploring the potential causes of attrition – is poor statistical practice

[4] *Bayesian approaches* are ones involving subjective probabilities and emphasis on conditional probabilities; they are to be distinguished from *Causal Bayesian Networks* methods for causal inference, discussed earlier, which need not introduce subjective probabilities.

and can also introduce bias (Altman 2005; Dumville, Torgerson & Hewitt 2006). Pocock and Hughes (1989) have pointed out that premature termination of trials almost inevitably leads to an exaggerated (and, therefore, misleading) estimate of the size of the treatment effect (see also Bassler *et al*. 2010). It is also problematic that many studies do not have plans for when to terminate a trial (Pocock 1992). Regarding when this should be they note the clash between collective ethics (i.e. how to choose the best treatment) and individual ethics (focusing on the rather different question of what to do about the next patient to be randomised). The paper then shows how to apply Bayes' theorem to derive a statistical estimate of the probabilities if trials are, or are not, terminated. The approach sounds promising and deserves further exploration (see Wathen & Thall 2008). It is important to distinguish between attrition in terms of non-adherence to a treatment and loss to the primary outcome measure, which may result in significant bias.

The purpose of randomisation in RCTs is to ensure that the treatment under evaluation is the only systematic difference between the comparison groups. Any removals post-randomisation from the treatment group because of choice by the patient (drop-out or non-compliance) or by choice of the clinical researchers (such as a focus on completers) will reintroduce the selection bias the randomisation is designed to preclude. In order to deal with this problem it has come to be accepted that it is necessary to have an 'intention-to-treat' analysis (see Kraemer 2015). This means that every patient who is randomised must be included in the evaluation of the two treatments.

Attrition in randomised trials is often practically inevitable and does not always lead to an imbalance of the causally relevant characteristics of the participants remaining in the trial. Attrition is most commonly addressed in RCTs by conducting 'intention-to-treat' analyses, which assess clinical effectiveness by analysing all study participants based on the group they were initially randomly allocated to, regardless of whether or not they dropped out, fully adhered to the treatment, or switched to an alternative treatment (National Institute of Health and Care Excellence 2016). However, the assessment of the characteristics of those who no longer participate in the trial and those whose data are actually incorporated in other study analyses may be important – such detail would allow judgement as to whether the comparison groups are balanced or whether the study results might have been affected. These data are rarely published and it is therefore difficult to assess the effect of attrition on the overall study results. The CONSORT (Consolidated Standards of Reporting Trials) guidelines recommend that a table showing baseline demographic and clinical characteristics for each group is reported (The CONSORT Statement 2010). Additional tables detailing the baseline characteristics of participants and non-participants that are not included in the analysis might also be useful.

2.9.2 Adherence to medicines

Adherence to medicines on the part of participants is a key factor in clinical trials. Because 'intention-to-treat' analyses (described above) have become the standard approach in RCTs, the findings are more harmed by someone who drops out after enrolment than by somebody who does not enrol in the trial in the first place. It may be desirable to select participants who are more likely to adhere and this can be determined by means of a run-in phase prior to randomisation. However, although this approach is helpful to quantify the effect of treatment when taken as prescribed, pre-selection of individuals who are likely to adhere to a treatment regimen may make the results less generalisable to wider treatment populations where adherence might not be as high. A variety of techniques to maintain good adherence have been tried. Those that appear to be useful are frequent contacts and reminders, providing easy transportation and access to attractive facilities, providing continuity of care, providing special medication dispensers (such as calendar packs), and involving family members, particularly when the intervention involves the need for lifestyle change (Friedman & Schron 2015).

It is also important to monitor adherence during clinical trials. Where possible, measurement of patient adherence to medicines should become more routine practice to gain better insight into the effect of non-adherence on clinical outcomes and potential adherence problems that might arise when the medicine is available more widely. Various mechanisms exist to monitor adherence, including pill counts, patient diaries, questionnaires and electronic monitoring methods. Pill counts, patient diaries and questionnaires are susceptible to patients adjusting their dosing history; however, they are non-intrusive, widely accepted by patients and relatively cheap to implement. On the other hand, electronic monitoring methods may be more reliable and unbiased but can be expensive, intrusive and considered less acceptable by patients. More research into how best to monitor adherence in research studies is needed.[5]

[5] In December 2014, the Academy held a workshop on *Patient adherence to medicines* to identify the key challenges and opportunities of better adherence to medicines. Further information is available on our website at: http://www.acmedsci.ac.uk/policy/policy-projects/patient-adherence-to-medicines/

Information on adherence rates in clinical studies should be available to allow critical appraisal of the results in terms of the likely applicability and validity of results to different groups. An approach termed 'non-compliance and local average treatment effects' can be used to estimate a treatment effect in a subpopulation of an RCT treatment group that adheres to treatment, when it is known that not everyone in the wider treatment group has adhered to the treatment (Imbens & Angrist 1994; Angrist, Imbens & Rubin 1996).

2.10 The 'placebo' and 'nocebo' effects

A complicating factor with the use of a placebo comparator, and one that remains poorly understood, is the so-called 'placebo effect', where the health of patients seems to improve despite being given a control treatment with no active ingredient. For example, the placebo response has been studied in relation to drugs designed to treat depression, but these considerations have led in two somewhat different directions. First, imaging studies have shown changes in cerebral glucose metabolism as a consequence of placebo treatment and second, from Jerome Frank's initial book onwards, there has been an interest in common factors associated with supposedly specific psychological treatments (Frank 1961).

The converse is also observed: a phenomenon known as the 'nocebo effect', where patients report suffering from adverse events despite being treated with the placebo (Planès, Villier & Mallaret 2016). It is important that both the placebo and the nocebo effects are considered as they may amplify or undermine the apparent safety of a treatment under investigation. For example, for the patients in the recent studies discussed below, only a minority of the symptoms reported by patients taking statins were genuinely due to the treatment and the majority of reported symptoms occurred just as frequently on placebo (see **Box 4**) (Ebrahim & Davey Smith 2015; Finegold *et al*. 2014). **It is important to understand the nature and impact of the placebo and nocebo effects more comprehensively, so that their impact on study findings can be more appropriately assessed.**

The use of stopping a medication as a way of evaluating whether side effects are due to a drug rather than a placebo constitutes an example of a broader range of studies using randomised withdrawal as the way of testing efficacy of a particular therapeutic medication (Newcorn *et al*. 2016). Some have argued that this should be the standard design for documenting the maintenance of efficacy in a long-term disorder.

Box 4. Nocebo effect when taking statins

Many patients receiving statins are told that muscle pain is a possible side effect, since they are associated with a rare, but serious, risk of myopathy (muscle weakness due to dysfunction of the muscle fibres). It has been suggested that prior knowledge of side effects might adversely influence patients' experiences, as was recently highlighted in the ODYSSEY ALTERNATIVE RCT comparing three cholesterol-lowering treatments (Moriarty *et al*. 2014). In this study, patients who had previously reported muscle symptoms with statin therapy were randomised into three groups receiving either:

- A statin (Atorvastatin) plus placebo injections.
- A different cholesterol-lowering drug (Ezetimibe) plus placebo injections.
- Injections of an alternative cholesterol-lowering treatment (Alirocumab, a PCSK9 inhibitor) plus placebo tablets.

Patients were blinded to which treatments they were receiving. Around 22% of patients receiving a statin reported muscle pain; however, a similar proportion (20%) of patients taking the alternative treatment Alirocumab reported similar symptoms, as did about 16% of those receiving Ezetimibe. Further, the numbers reporting muscle pain dropped to less than 5% when they stopped taking the study tablets in all groups, irrespective of whether they were active statin tablets or matching placebo tablets. These results suggest that prior knowledge of potential side effects can negatively impact on patient experiences.

Recent n-of-1 trials in eight patients who experienced muscle pain (myalgia) when taking statins suggested that a similar approach might be a useful method for determining which patients' adverse effects can be specifically attributed to statin treatment (Ebrahim & Davey Smith 2015; Joy *et al*. 2014). In this study, patients with a history of statin-related myalgia switched randomly and blindly between placebo and statin over repeat three-week periods. The frequency and severity of symptoms were similar whether patients were taking statins or the placebo, suggesting that the symptoms were not directly caused by statin treatment. In fact, most patients resumed statin therapy after reviewing their results. To provide further clarification of the side effects that are likely to be caused by statin treatment, the Cholesterol Treatment Trialists' Collaboration is currently conducting a reanalysis of the adverse events reported in statin trials (Ebrahim & Davey Smith 2015).

2.11 Surrogate endpoints

Surrogate endpoints (e.g. tumour shrinkage or changes in cholesterol level) can be useful when determining whether a new intervention is biologically active and warrants further investigation of clinically meaningful outcomes. However, surrogate endpoints are frequently used in Phase III clinical trials in an attempt to reduce the cost and duration of these studies (Fleming & DeMets 1996). There are concerns that surrogate endpoints frequently fail to reliably predict the overall effect on the clinical outcome and thus often fall short of being an effective substitute for the clinical outcome. This could be due to a variety of reasons, such as the clinical outcome resulting from the intervention acting via multiple causal pathways (some of which may not even be recognised or anticipated) which are not reflected in the measurement of the surrogate endpoint, or moderating variables (known or unknown) also affecting the surrogate endpoint. This has led to claims (Fleming & DeMets 1996), which we endorse, that unless the validity of the surrogate endpoint has already been rigorously established, the true clinical outcome should be used as the primary endpoint in Phase III trials. Of course, data on surrogate endpoints that demonstrate biological activity can be collected in addition to clinical outcome measures during these trials, as they are likely to provide further clarity as to whether the intervention is active and targeting the pathways of interest, but surrogate endpoints should not be used as a substitute for true clinical outcomes.

2.11.1 Patient-relevant endpoints

A second issue concerning the choice of endpoints in clinical studies is that they may not always measure the outcomes that matter most to patients that stand to benefit from the treatment. This is particularly true when surrogate endpoints are used. For example, in an attempt to standardise disease measurements in rheumatoid arthritis (RA), health professionals and methodologists devised a core set of eight outcomes to be used as an international standard in RA clinical trials (Nicklin 2010; Boers *et al.* 1998; Felson *et al*. 1993). However, this initial set of outcomes failed to incorporate several outcomes of importance to patients. One of these was fatigue, which is reported by almost every patient. Fatigue was successfully added to the core set following a decade of research into this symptom (Nicklin 2010; OMERACT 2015). **Due consideration should be given to outcomes that are of particular importance to patients when designing, analysing and evaluating clinical studies.**

In **Chapter 3**, we outline how different study types are affected by the research issues described here, after briefly summarising their strengths and limitations. We focus on research undertaken since 2008, so as not to repeat comprehensive methodological reviews that have been published.

3. Addressing the challenges of research designs

In this chapter we aim to set out how different study types can, and have, addressed the research issues outlined in Chapter 2, and where there might still be challenges to address. The strengths and limitations of evidence from different study types have already been comprehensively reviewed elsewhere (Rawlins 2008; Academy of Medical Sciences 2007). To avoid duplication and repetition of previous work, we will summarise the key points from these accounts and focus predominantly on research undertaken since these documents were published. In doing so, we will highlight whether the points made by Rawlins in his 2008 Harveian Oration still apply today.

3.1 Randomised controlled trials

3.1.1 The traditional randomised controlled trial design

In experimental studies, interventions are tested in participants according to a pre-established protocol. The randomised controlled trial (RCT) is the most common form of experimental study and involves the comparison of two or more treatments that, from the outset of the trial, have been randomly allocated to groups of patients that are treated contemporaneously. During the process of drug development, the safety and efficacy of new treatments are typically evaluated through a series of phased clinical trials, most commonly RCTs. This controlled, staged approach enables the safety and efficacy of the novel therapies to be tested in humans as safely as possible, by ensuring the experimental

medicine is safe at low doses in a small number of volunteers before it is tested at higher doses and with larger numbers of participants (see **Box 5** for further details).

RCTs have a number of well-recognised strengths (Rawlins 2008). First, randomisation and blinding (when it is used) minimise different forms of bias by respectively increasing the likelihood that confounding factors (known or unknown) are distributed in a balanced manner between the groups, and reducing the likelihood of selection and ascertainment bias. Second, when properly conducted and analysed, RCTs provide confidence in the internal validity of the results (that is, they produce findings that are valid within the population that the sample investigated has been taken from), particularly if they are performed on a large trial population and replicated by subsequent studies. Third, by minimising confounding and bias, RCTs can provide a measurement of the causal effect of an intervention (i.e. its efficacy) that is likely to be correct. This makes RCTs particularly good at detecting causal effects on common outcomes, even when those effects are moderate in size (Collins 2016). Importantly, when the size of the effect of a medicinal product is modest, fully randomised and well-blinded large-population RCTs are the only reliable design for detecting that effect, with few substantive assumptions required.

For these reasons, RCTs have been considered the 'gold standard' for clinical studies. Their strengths at being able to establish a causal link between the intervention and the outcome in an unbiased manner also mean that they are required by regulatory authorities in marketing authorisation dossiers for new treatments.

However, RCTs can suffer from a number of limitations (Rawlins 2008):

- **Generalisability:** RCTs are often carried out in highly selected patient populations for relatively brief periods of time, whereas in practice interventions are likely to be used in more heterogeneous populations (potentially with comorbidities) for periods of time that may extend beyond that of the clinical study. As such, RCTs are sometimes considered to have limited external validity, with questions raised as to how far the results can be generalised or extrapolated to patients beyond those included in the study. For example, the findings of studies exploring the effects of acute treatment in low-risk young women are unlikely to be directly applicable to chronic treatment in high-risk older men.
- **Relevance to routine practice and patients:** RCTs may not necessarily reliably quantify effectiveness (i.e. benefit to patients in routine practice) and may have limited applicability to routine practice, particularly if they detect a modest but statistically 'significant' effect that is not clinically relevant. For example, a treatment shown to significantly reduce swelling of the knee in RCTs may not translate into improved mobility in routine practice as experienced by patients.
- **Assessment of harms:** RCTs are usually planned on the basis of the sample size needed to detect benefits and rarely is there a suitable assessment of the sample size needed to examine harms. It is important that there is adequate funding of sufficiently large samples to detect both benefits and common harms. However, it would be impractical for RCTs to be expected to detect rare side effects, which are better examined using observational studies, schemes for reporting adverse events such as the UK's Yellow Card Scheme (Yellow Card 2016), and registries such as the UK's Clinical Practice Research Datalink (2016).
- **Ethical implications and practicality:** There are instances where it would arguably be inappropriate to carry out RCTs, for example in certain research circumstances in rare diseases or emergency situations (see **Chapter 4** for more detailed discussion). In other instances, it might be impossible in practice to carry out a blinded RCT, for example when comparing a drug to behavioural therapy.
- **Resources:** RCTs can be very expensive and lengthy, due in part to increasing regulatory requirements that are aimed primarily at protecting patient safety and ensuring greater transparency. Such requirements are clearly important but have made it more difficult to carry out RCTs in practice.

Rawlins (2008) highlighted two areas which have not been systematically evaluated in RCTs: how to deal with the difficulties created by premature termination of RCTs (in particular estimation of treatment effect); and whether under-representation of certain groups in these studies really affects the generalisability of the study results (Rawlins 2008). Certain subgroups, particularly some ethnic groups, can be under-represented in trials in terms of the absolute numbers of participants included (that is even if the percentage of participants included is representative of the percentage of the eligible population). The balance between representation and interpretation of each subgroup, and the generalisability of the overall study population should therefore be considered. These research gaps are still relevant today.

The analysis of RCTs based on a null hypothesis, which presumes there is no difference between the treatments evaluated, as the best analysis approach has been questioned (Rawlins 2008). An alternative would be wider use of Bayesian approaches (see **Chapter 2**; Ashby 2006; Salsburg 2002; Spiegelhalter 2000), for which there is support in the analysis and interpretation of RCT data (Rawlins 2008). Even though Bayesian approaches might not be systematically

employed, thinking about prior expectations and their impact on the data might be useful. In practice, a prior probability might be determined based on Phase II clinical studies; unfortunately, such data may not always be available so the use of Bayesian approaches can become more of a subjective judgement.

Recommendation 2

We recommend that funding bodies ensure appropriate support for research in the areas of: how to deal with the difficulties created by premature termination of RCTs (in particular estimation of treatment effect); and the extent to which under-representation of certain groups in these studies really affects the generalisability of the study results. Appropriate support should also be provided for trials that are sufficient in scale and duration to achieve the pre-specified outcomes.

Box 5. The phases of clinical trials

During drug development, phased clinical trials enable the safety and efficacy of medicinal products to be assessed in a safe and effective manner. They are subject to strict rules and regulations, including the EU Clinical Trials Regulation (European Commission 2014). Clinical trials typically follow four phases (NHS Choices 2015):

- Phase I: Phase I trials are designed to test the safety of a new medicine. They aim to identify any potential side effects, as well as the most effective dose to use in further studies and subsequent treatment. Phase I trials involve only a small number of participants, including healthy volunteers.
- Phase II: Phase II trials are bigger studies involving larger numbers of patient volunteers with the target disease. These trials aim to gather further information on the most effective dose to use and the potential side effects.
- Phase III: Phase III trials compare the new medicine under investigation against a placebo or an existing treatment. The aim is to determine whether the new medicine is more effective than standard treatment and whether it has important side effects. These trials generally involve much larger groups of patient volunteers than Phase I or II trials, up to several thousands of patients.
- Phase IV: Phase IV trials occur once a medicine has been shown to be effective and has been granted a marketing authorisation to be prescribed. Phase IV trials provide further information on the effectiveness of a product when it is used more widely, side effects and safety profile, and the risks and benefits in the longer term.

3.1.2 Alternative RCT designs

Although there is currently a relatively wide range of robust effective methodologies for assessing the efficacy and side effects of medicines, a range of alternative methodologies has been developed in recent years to further address some of the limitations of traditional approaches. In the UK, the MHRA plays a major innovative role in adopting and advising on new trial designs. We focus here on multi-arm RCTs and pragmatic trials, and briefly describe further alternative designs in Annex IV.

3.1.2.1 Multi-arm RCTs

In Chapter 2, we noted that it had become increasingly recognised that clinicians, regulators and patients want to know

whether the specified treatment under investigation is better than already available alternative treatments. This has led to the suggestion that three-arm trials may need to become the norm in late Phase II and Phase III trials. We note here that the expectation that participants in any trial should not be disadvantaged by taking part cannot be met if the placebo requirement involves stopping taking alternative treatments of known efficacy. It has been argued that greater use should be made of multi-arm randomised trials because frequently one will need to be able to compare either the dose of the drug under investigation or alternatively many new drugs against a control arm (Parmar, Carpenter & Sydes 2014). This has been exemplified particularly in the field of oncology but it is likely to be applicable more widely.

3.1.2.2 Pragmatic trials

Although the results from traditional RCTs may be valid in the healthcare setting in which they were collected, they may have limited generalisability (external validity) to patient populations beyond those that were included in the study, owing to differences between the trial setting and the situations in which the treatment is to be more widely used (Zwarenstein *et al*. 2008). Pragmatic trials were therefore developed to bridge this gap. They typically evaluate the effectiveness of interventions in a broader, realistic, clinical setting, and aim to reconcile the constraints under which policymakers and service managers operate with the need for rigorous scientific evaluation (Schwartz & Lellouch 1967). For example, a pragmatic trial may explore the effect of a treatment for asthma in broader populations with or without other illnesses, such as high blood pressure or cancer. There is no sharp dichotomy between trials that are designed to test causal research hypotheses (commonly termed 'explanatory' trials, such as those that examine whether an intervention causes a particular biological change) and pragmatic trials that help to decide on care options. Rather, they are on a continuum of trials, where the safety and efficacy of treatments is (and needs to be) established, before their effectiveness is assessed in pragmatic trials. As for all trials, accurate reporting of findings is essential and, in the case of pragmatic trials, will help to determine whether the results are applicable or need to be extrapolated to a given situation, and whether a studied intervention might be feasible and acceptable. The CONSORT guidelines detail a checklist of 22 items that should be reported for pragmatic trials, including the need to blind the statisticians who are analysing the data (see Zwarenstein *et al*. 2008 for further details).

Despite the usefulness of pragmatic trials, they also suffer from a set of limitations. Attrition rates can be higher in pragmatic trials than in other forms of RCTs and, although there may be substantial advantages in using patients attending an ordinary care setting for the RCT (e.g. a general practitioner (GP) surgery), there may be an accompanying danger that randomisation of participants to treatment groups may not occur because of the pressure to make the trial acceptable to participants. A further point is that, although pragmatic trials aim to increase the external validity of RCT findings, there is no single ordinary situation that covers all eventualities and, as such, the findings of a single pragmatic trial may not necessarily be generalisable to all situations (for a discussion of the strengths and limitations of pragmatic trials in a research context, please see Cape *et al*. 2016). This is of course true for any individual study, whether randomised or not.

3.2 Observational studies

In contrast to experimental studies, researchers undertaking observational studies *observe* participants and the interventions they receive, without actively assigning subjects to specific treatment groups. For example, an observational study might explore the effect of hormone replacement therapy (HRT) in women in Wales over 20 years. Researchers investigating this study will have no role in assigning subjects to treatment or control groups but will observe those that choose to take HRT and those that do not. Observational studies include cohort studies, historical controlled trials, case-control studies, and before-and-after designs, among others (see **Box 6** for further details) (Rawlins 2008).

Observational studies can have good external validity, as they typically evaluate the intervention effect in a heterogeneous section of the population. They can help to address research questions that cannot be addressed in RCTs due to ethical or pragmatic reasons (see **Chapter 4** for further details). They also provide a method for detecting rare and long-term harms. Observational studies are particularly good at detecting large effects on rare outcomes, which do not occur frequently enough to allow for their reliable assessment in RCTs (Collins 2016).

However, observational studies may also suffer from a number of major limitations:

- **Bias:** As subjects studied in observational studies are not randomly assigned to different comparison groups, nor are they (or the researcher) necessarily blinded to the intervention, biases such as selection bias, performance bias, experimenter bias or ascertainment bias (see **Chapter 2**) can be introduced.

- **Confounding:** Owing to the lack of randomisation to comparison groups, confounding factors (see **Chapter 2**) in observational studies are unlikely to be equally distributed between the groups. Therefore, the comparison groups might differ by factors other than the intervention under investigation (e.g. baseline characteristics, prognostic factors, or concomitant interventions). This makes it difficult to determine whether the apparent effect of the intervention is actually due to the intervention or to another characteristic, which is associated with both the intervention and the outcome. When possible causal effects of drug treatments are explored in observational studies, a specific form of confounding – confounding by indication – can occur. This is when treatment is given to patients in routine practice (rather than in a randomised trial) and the reason the treatment is given 'confounds' the association of the treatment with a particular outcome (an example is given below of the case of cimetidine and stomach cancer). Various approaches are available for assessing and adjusting the results to take account of measured confounding factors (Mamdani *et al*. 2005; Normand *et al*. 2005), but adjustment for unmeasured or unknown confounding variables, of which there are potentially many, is not possible (Davey Smith *et al*. 2008; Lawlor *et al*. 2004a; Lawlor *et al*. 2004b).
- **Causality:** As observational studies are likely to suffer from bias and confounding factors, it is difficult to conclusively establish a causal link between the intervention and outcome of interest.

3.2.1 Examples of erroneous assumptions of causality for treatments from observational studies

Erroneous inferences drawn from observational studies into the effects of HRT provide a good example of the major problem of selection bias and confounding in observational studies (see Academy of Medical Sciences 2007 for a discussion). HRT tended to be used by women whose lifestyle differed markedly from those not using HRT. The initial claims about the protective effects of HRT in relation to cardiovascular disease, which were not observed in the Women's Health Initiative RCT (Writing Group for the Women's Health Initiative Investigators 2002; Hsia *et al*. 2006; Manson *et al*. 2003), are likely to reflect the following: selection bias in relation to the women to whom doctors were happy to prescribe HRT; the socioeconomic and lifestyle differences between those who did or did not request HRT (i.e. confounding); and possible differences in the effect of HRT on CVD between when it was first started and after it had been taken for some time (Lawlor & Davey Smith 2006; Hernan *et al*. 2008).

The use of cimetidine, a treatment for peptic ulceration, is a good example of confounding by indication: its use was associated with the development of carcinoma of the stomach. However, it was found that cimetidine was often used to treat other diseases, which were most likely to be the cause of the increased cancer mortality rather than cimetidine itself (Academy of Medical Sciences 2007).

It is notable that concerns about the teratogenicity of thalidomide (and also amphetamine psychosis) arose from studies of particular clinical patterns that were not part of drugs trials of any kind (Mellin & Katzenstrin 1962; McBride 1961). Such examples provide an important reminder that other sources of information should not be overlooked. A critical element of regulatory processes is to consider the totality of the data from all types of data sources, including case reports and case series, which can act as important signals for unanticipated rare adverse events and avenues for further research.

Observational studies frequently come under criticism for producing data that are less reliable than those emanating from RCTs. However, **reasonably strong conclusions can be drawn on the effectiveness of interventions despite the lack of RCTs when the effects of the treatment have the following three characteristics:**

1. The treatment effects are large.
2. They occur in the treatment of a condition which almost invariably has a poor outcome (e.g. death).
3. They are biologically plausible (in other words, there is a strong scientific rationale explaining why the treatment results in the effect – see **Box 3**).

This was the case when the treatment of myxoedema by thyroxine and pernicious anaemia by vitamin B12 were discovered in the absence of RCTs (Glasziou *et al*. 2007). These situations are, however, more often the exception than the norm nowadays owing to significant advances in medicine and in the treatment of most common diseases. Further, evaluating the effectiveness of interventions in the absence of RCT data is unlikely to be possible in the case of remitting disorders. The role of observational studies in defining benefit when the effect size is modest is widely regarded as inappropriate (McKee *et al*. 1999; Vandenbroucke 2004).

Box 6. Observational study types

'Observational studies' is an umbrella term for a number of different study types, including the following (Rawlins 2008):

- Historical controlled trials: Trials in which a group of patients treated with an intervention are retrospectively compared with a group that had previously received a standard therapy.
- Case-control studies: Studies that compare the use of an intervention in groups with and without a particular disease or condition. Cases are identified on the basis of outcomes already achieved, and control groups are individuals selected to be similar but without the outcome. The differences in exposure are then compared. Such studies are widely used in epidemiology to identify risk factors for diseases, but they can also be used to investigate associations between interventions and effects, for example between medicines and adverse effects.
- Before-and-after designs: A design in which patients are studied before and after treatment to determine the effect of the intervention. The patients are effectively their own controls.
- Case series: Involves the use of routinely collected patient-level data, which detail both the interventions and patient outcomes. Such data are often collected in registries and can provide further information about the generalisability of RCT results, or further evidence about the safety of an intervention.
- Case reports: Many regulatory authorities have established schemes for reporting adverse events to marketed medicines. In the UK, this is the MHRA's Yellow Card Scheme (Yellow Card 2016). The data from such initiatives can provide important insights for monitoring the safety of medicines. Although manufacturers are legally obliged to report any suspected adverse events they are aware of to regulatory authorities, most reporting schemes are voluntary for healthcare professionals and patients. Therefore, these initiatives are susceptible to under-reporting and reporting bias. Case reports can be particularly valuable in highlighting areas for further research, perhaps requiring confirmation from other types of studies, and in guiding the research questions to be asked.
- Cohort studies: Studies in which groups of individuals who are and who are not exposed to an intervention are followed over time. Outcomes are compared across the groups that have received the intervention and those that have not. Cohort studies may be prospective (where the outcomes of interest are actively recorded as the study progresses), or retrospective (where both the intervention and outcome have already occurred at the time of the study) (Pearson 2016).

3.3 Attrition, adherence, choice of comparator and endpoints

Both RCTs and observational studies have to pay attention to issues associated with the attrition of research participants, lack of adherence to treatment regimens (although to some extent this may be easier to monitor and/or control where the intervention is administered by research staff, as for some RCTs), the choice of comparator, and the use of endpoints that are clinically meaningful and relevant to patient populations. When making judgements about evidence, these factors should be taken into consideration in the critical appraisal of the potential benefits and harms of medicines and can provide important insights into the generalisability of results.

3.4 Qualitative research

RCTs, and other trial designs, provide a framework within which researchers can collect both quantitative and qualitative data depending on the research question (this is commonly referred to as 'mixed methods research'). Qualitative research involves the collection, analysis and interpretation of non-numerical data. We provide only a brief discussion of qualitative evidence here but we recognise that it has the potential to provide valuable information about patient preferences (including relating to the importance of particular outcomes) and their attitude to risk, as well as their approach to trading risk for benefit (Rawlins 2008). In addition, it may also provide important insights into the social values expressed by society as a whole. Both qualitative and quantitative patient-reported outcomes can provide important insights into therapeutic treatments. The MRC has produced helpful guidance on developing and evaluating complex interventions, including those that use qualitative data (Medical Research Council 2008).

There is a need to develop criteria for assessing the quality of qualitative evidence. It has been suggested that there are seven key criteria, which equally apply to quantitative research (Harden, Weston & Oakley 1999):

1. An explicit account of the theoretical framework and/or inclusion of a literature review to provide the background to the research study and questions explored.
2. Clearly stated aims and objectives of the study undertaken.
3. A clear description of the content of the study.
4. A clear description of the sample of individuals employed.
5. A clear description of methodology including systematic data collection methods.
6. An analysis of the data by more than one researcher.
7. The inclusion of sufficient original data to mediate between data and interpretation.

It is possible that robust qualitative evidence could sometimes provide important insights into therapeutic interventions. The reliable use of qualitative evidence in the medical sciences should be considered more widely.

3.5 Meta-analyses and systematic reviews

Meta-analyses use statistical techniques to combine and analyse data from several studies in order to derive an overall estimate of a treatment's effect (Guyatt *et al*. 1995; Uman 2011). For example, a meta-analysis of RCTs might combine and analyse the data from all RCTs assessing a novel treatment for prostate cancer to get an overall estimate of its effect and determine whether it is better than an existing treatment. They constitute a standardised, statistical means of combining data across studies, giving greater weight to larger studies, and they assess the homogeneity of treatment effect sizes across studies (Shadish, Cook & Campbell 2002). The use of meta-analysis is a crucial way of combining quantitative evidence across studies. It has revolutionised the assimilation of evidence and has played a pivotal role in providing critical insights into the benefits and harms of treatments. While most meta-analyses synthesise aggregate (trial level) data, in individual patient data (IPD) meta-analyses the trialists contribute their raw data so that it can be analysed together as if it is one big trial. Providing that these are properly conducted (e.g. data is openly available) IPD meta-analyses can provide more robust evidence.

Inevitably, meta-analyses rely on what may be uncertain inferences regarding the comparability of studies (with respect to samples, measures, duration of follow-up and interventions). It is notable that when several meta-analyses of the same topic have been undertaken, they may differ in their conclusions because of differences in inclusion and exclusion criteria, drug dosage or regimen, trial outcomes, duration of follow-up, and so on (see Rutter & Pickles 2015). These differences can, however, provide an opportunity to explore heterogeneity and its influences on study results. Publication bias, that is the publication of so-called 'positive' results at the expense of null, 'negative' or inconclusive findings (see Chapter 5), is also a problem as important results may be omitted from these analyses. The 'Cochrane Handbook for Systematic Reviews of Interventions' (Higgins *et al*. 2008) describes steps that can be put in place to aid high-quality meta-analyses (further resources are also available on Cochrane's website: Cochrane Community Archive 2015).

Different RCTs that include large numbers of patients with different eligibility criteria can be combined in a meta-analysis to allow the findings obtained from narrowly-defined patient populations to be more widely generalisable. Such analyses have the potential to provide further information based on sufficiently large numbers of randomised individuals with different characteristics (e.g. from a spectrum of ages, gender and risk groups) while avoiding some of the inherent biases of non-randomised studies (Collins 2016). A good example of this approach (in this case IPD meta-analyses) is in the

assessment of the efficacy and safety of statin therapy in various populations (Cholesterol Treatment Trialists' Collaboration 2010, 2012a, 2012b, 2015).

Robust meta-analyses systematically assess the risk of bias in included studies. However, some meta-analyses do not always acknowledge systematic biases within the studies that are included, leading to the incorporation of biases in the overall results. Bias modelling allows the results of studies, including the meta-analyses, to be adjusted for the internal and external biases that they are judged to suffer from (Turner *et al*. 2009; Wilks *et al*. 2011). Ideally, this should be based on objective measures of internal and external bias but the evidence is rarely available. We note that a meta-analysis reports an average effect that does not necessarily represent the true value for any of the individual populations under investigation. However, meta-analyses of RCTs carried out in different populations enable the systematic assessment of treatment variation in these different groups to determine the extent to which effects are consistent.

Particular caution is required in interpreting meta-analysis results of observational studies, which can result in very precise but biased effect estimates (Egger, Scheider & Davey Smith 1998). This is because the quality of results from meta-analyses depends on the quality of the individual studies that contribute to them, and because bias and confounding in observational studies can limit their internal validity for testing treatment effects. Tests for assessing small study bias (often resulting from publication bias) were developed for use in meta-analyses of RCTs and may be unable to detect this source of bias in meta-analyses of observational studies (Egger, Scheider & Davey Smith 1998).

3.6 Hierarchies of evidence

In an attempt to provide a simple framework by which the 'strength' of the evidence can be intimated, so-called 'hierarchies of evidence' have been developed, which typically place RCTs at the top of their classification. Box 7 provides further information on these hierarchies and a brief overview of their development is included in Annex III. Some view hierarchies as an attempt to replace judgement with an over simplistic, pseudo-quantitative assessment of the quality of the available evidence (Rawlins 2008). We note that despite their serious limitations, they have provided useful indicators of the likely robustness of evidence from different types of study. The problem is that there have been a large number of different hierarchies proposed, and there is poor agreement among them (Boaz & Ashby 2003). Most crucially, hierarchies cannot accommodate evidence that relies on combining results from RCTs and observational studies and that is something that would ordinarily seem desirable.

It is important that hierarchies of evidence are not used too prescriptively nor as a substitute for judgement as part of the critical appraisal of evidence. The 'strength' of evidence will be dependent on the research question under investigation and the data utilised should be 'fit for purpose'.

Box 7. Hierarchies of evidence

To guide decision-making about the appropriate use of therapeutic interventions, 'hierarchies of evidence' have emerged which aim to provide an idea of the strength of the underlying evidence. Details of a subset of such hierarchies are provided in Annex III.

Hierarchies of evidence typically place RCTs at the top (or systematic reviews or meta-analyses of RCTs) as providing the most robust form of evidence, followed by cohort studies, case-control studies, case series, studies with no controls, and expert opinion. However, the use of such hierarchies is much debated. It has been argued that, while they are pragmatic, placing evidence sources on such hierarchies is inappropriate and that decision-making should be informed by all sources of evidence on a particular topic, with due regard given to the strengths and limitations of each study type to allow reasonable and reliable conclusions to be drawn (Rawlins 2008). Further, the term 'hierarchy' may not be helpful as it implies that certain forms of evidence are superior to others in all circumstances.

There are multiple examples where forms of evidence traditionally considered less robust in these hierarchies have provided important clinical information, further supporting the idea that the strict use of these hierarchies is not helpful in all situations. Examples include the use of thyroxine in myxoedema; vitamin B12 for the treatment of pernicious anaemia; and sulfonamides for the treatment of puerperal sepsis (Rawlins 2008). A commonality among all these examples is that the prognosis prior to treatment was poor and the effect of treatment was particularly marked. This meant that RCTs, which would usually be desirable to demonstrate the efficacy of treatments, were not necessary, especially given the contributing information on the biology of the disease, which provided insight into a plausible biological mechanism of action.

The decision as to what type of study should be undertaken is complex. It requires a thoughtful approach in which the wide range of research issues outlined in **Chapter 2** (e.g. the risk of bias and confounding, the importance of external validity, and so on) are carefully considered. It is unlikely that in most instances all potential study limitations can be addressed when designing a study; however, due consideration should be given to these different issues, and judgement as to which elements are of most concern should be made. For example, if it is most important that the findings of a research study are highly generalisable, then investigators need to consider whether they are prepared to tolerate the likely increase in certain biases, and vice versa. It is also important that researchers consider which biases they would be prepared to tolerate and the size of the effect – if the effect size is expected to be moderate then an RCT is most likely to be the study design that provides the most reliable evidence. The researchers' judgement on these considerations will dictate the most appropriate study type for the research question to be investigated.

Bodies who make decisions on behalf of the public (for example NICE and the MHRA, among others) should systematically consider all the research issues outlined in **Chapter 2** when assessing the evidence they are presented with, to ensure that their decisions are based on the most robust evidence possible. It is also important to remember that there are different data needs for different purposes. For example, the requirements to judge whether to invest in a Phase III programme, based on Phase II data, are different to the requirements for evidence for routine clinical use; or the requirements to warn about a possible side effect may be very different to requirements to evaluate the whole range of benefits and harms of a medicine; or the data needed for potential interaction between drugs, which in many cases can be simply obtained in small pharmacokinetic interaction studies (studies that examine the absorption, distribution, metabolism, and elimination of drugs by the body), whereas those needed for the study of pharmacodynamic interactions (the study of the effects of medicines on the body), might be more complex.

In **Chapter 6** we explore some of the issues outlined in **Chapters 2** and **3** in the context of evidence use in clinical practice.

Recommendation 3

We recommend that all those evaluating evidence should pay particular attention to factors that are likely to affect the validity and applicability of the results, including:

- **Biological plausibility** – are the findings based on sound biological principles?
- **Generalisability** – do the results extend to the treatment populations of interest?
- **Effect size** – is the size of the treatment effect large enough to be reliably detected in the study design that was undertaken and/or is the sample size large enough to detect a clinically important treatment effect if it exists?
- **Causality** – do the results reliably demonstrate a causal link between the treatment and the observed effect or do they merely suggest a correlation or association?

Decision-makers should use their judgement as part of the critical appraisal of the evidence, to ascertain whether the evidence they are presented with is 'fit for purpose'. This

judgement is central to the assessment of the benefits and harms of medicines, as well as for the evaluation of research findings when considering their application in clinical practice.

Researchers should be aware that these factors will be influenced by determinants such as bias, confounding, moderating variables, choice of comparator and endpoints, participant attrition and adherence to treatments, and the 'placebo' and 'nocebo' effects, which should therefore be carefully considered in the study design. Alternative trial methodologies and analytical approaches (including Bayesian thinking) should be given due consideration, as should the investigation of outcomes that are of particular importance to patients.

In this report, while we make recommendations aimed at further improving the way in which evidence is generated, analysed and evaluated, we are encouraged by the considerable progress that has been made in this area in recent decades. This is evident for example in the conduct and reporting of trials, in meta-analyses, and in the implementation of standards.

In the next chapter, we describe evolving approaches that aim to address some of the research limitations outlined in Chapters 2 and 3. We also outline areas where new research strategies are required, and the future challenges in measuring the benefits and harms of medicines.

4. Generating evidence – evolving approaches and special cases

In Chapter 3, we discussed the strengths and limitations of traditional approaches to assessing the benefits and harms of medicines. We also briefly outlined alternative trial designs that are variations on RCTs and that have been developed to address some of the challenges with the traditional RCT. In this chapter, we explore designs that have evolved largely as a result of studies outside of any kind of pharmaceutical trial, and that could help to further address some of the outstanding limitations of traditional approaches of assessing the benefits and particularly the harms of medicines. These include:

- Propensity scores, as a mechanism to minimise the effect of selection bias in observational data.
- Natural experiments, as a design that can aid in establishing causal links.
- Mendelian randomisation, which uses instrumental variable analysis to explore the causal effect of treatments or risk factors on outcomes.

These methods have not been widely used for testing drug treatment effects and we anticipate wider use in the future. The Academy's report 'Identifying the environmental causes of disease' further discusses some of the statistical aspects of these designs (Academy of Medical Sciences 2007).

In this chapter, we also explore research areas where specific points for consideration arise. These include research in rare diseases and in emergency situations, where alternative strategies may be needed to investigate the harms and benefits of medicines. Finally we outline a set of future challenges, where further research into best practice is needed. This includes: methods for assessing stratified medicines; approaches to using so-called 'real world' evidence; and consideration of patients with multiple illnesses.

4.1 Propensity scores as a way of dealing with social selection

Propensity scores refer to the conditional probability of an exposure to a particular experience (Rosenbaum & Rubin 1983). The propensity score enables the creation of comparison groups that are relatively balanced in their measured baseline characteristics, for instance demographic characteristics, signs and symptoms, risk factors, comorbid conditions, relevant disease history, laboratory tests, relevant prescriptions, and so on. Comparison groups that are matched in their measured baseline characteristics as far as is practically possible can then be compared and causality between the treatment under investigation and the outcome of interest may be more reliably inferred, as steps have been taken to minimise the effect of selection bias. The propensity may involve either exposure to risk factors for a disease, or the use of therapeutic medications to treat or prevent such a disease.

There are two main uses of propensity scores. First, as shown in the Sure Start example (see **Box 8**), they may identify the lack of overlap between groups and hence the need to exclude non-overlapping data from the analysis. Second, propensity scores may be used to match groups using an inverse probability of treatment weighting (IPTW) as shown by the example of studying the effect of marriage in protecting against criminal behaviour (see **Box 9**). Propensity scores have been developed mainly outside of the domain of drug studies and hence we discuss their practicalities and principles in relation to Sure Start and antisocial behaviour before moving on to examples where they have been applied to medical research (see **Box 10**).

Conventional adjustment for covariates, by contrast, focuses not on the probability of exposure to treatment but rather on the variables concerning subject characteristics as obtained prior to treatment that are associated with outcome in the disease or disorder being studied. The statistical appendix to the Academy's 2007 'Identifying the environmental causes of disease' report argued that there were several advantages of propensity scoring over the more usual covariate adjustment. In particular, the fact that propensity scores should be undertaken without knowledge of the outcomes means that the same initial analysis can be used for a range of response variables. Accordingly, we regard propensity scores as a potentially useful analytic tool. Nevertheless, we caution that they work best when there are large samples and when they are constructed from as many predictors of group membership and outcome as is contextually feasible (Garbe *et al.* 2013). However, even in these circumstances it has been argued that there is still the inevitable limitation that propensity scores can only be as good as the variables used to construct them, and that missing variables may well constitute an important bias (Shadish, Cook & Campbell 2002). The Heckman correction, devised in Nobel Prize-winning research by the economist James Heckman, aims to address this issue of unmeasured or unknown variables. It models selection bias by using a two-step statistical approach, which essentially treats the selection problem as an omitted variable. It has been adopted to a degree for the analysis of clinical studies, particularly those impaired by missing data (Sales *et al.* 2004).

Box 8. The use of propensity scores to identify the lack of overlap between groups – the Sure Start programme example

Propensity scores can be used to determine the lack of overlap between groups that are being compared. A good example of this is the Sure Start programme, which was set up in 1998 with the aim of giving 'children the best possible start in life' by improving education,

childcare, health and family support. Due to government opposition to running an RCT to study the effects of the programme, these were instead examined by a comparison between areas where Sure Start was available and areas where it was not. Propensity scores showed that there was scarcely any overlap between the two groups with respect to the characteristics associated with the use of services. However, exclusion of the extremes from both groups allowed the reanalysis of the data with the diminished sample to examine the effectiveness of the programme (Department for Education 2010).

Box 9. The effect of marriage on crime – the use of propensity scores to match groups

Sampson, Laub and Wimer (2006) used propensity scores to examine whether crime rates varied according to whether the individuals were in a marital relationship. Some 20 variables were found to be associated with whether or not someone got married. The combination of propensity scores and age curve variations in crime rate were brought together with the statistical technique of IPTW. The findings suggested that marriage did indeed exert a protective effect with respect to crime and that this was so in both short and long time spans.

Box 10. The use of propensity scores in medical research

In recent years, the use of propensity scores has been adopted in medical research. The following studies are examples of where they have been used, and demonstrate the range of research fields in which they can be employed:

- One recent cohort study in HIV research that compared the virological responses to lamivudine and emtricitabine as part of combination antiretroviral therapy used propensity scoring to reduce confounding from treatment allocation bias (Rokx *et al.* 2016).
- Inverse propensity score weighting was used to balance treatment groups and reduce bias from cluster randomisation in a study evaluating the impact of consultations on patients with osteoarthritis of the knee (Ravaud *et al.* 2009).
- Adjustment for propensity scores was used to reduce potential confounding from selection bias in a study determining the effectiveness of non-professional mentor support in reducing intimate partner violence and depression in mothers (Taft *et al.* 2011).
- In a study of primary care management of major depression in patients of over 55 years of age, propensity scores were used to match GPs who performed the intervention (van Marwijk *et al.* 2008).
- Adjustment for propensity scores, based on a very large number of variables in an administrative database, was used in a study of the effect of cyclooxygenase (COX)-2 inhibitors on gastric toxicity (Schneeweiss *et al.* 2009).

4.2 Natural experiments

Bradford Hill (1965) argued that a key need in testing causal effects was to conceptualise and consider possible alternative explanations for an observed statistical association. Natural experiments provide an approximation to experimental conditions by pulling apart variables that ordinarily go together (Rutter 2007; Rutter & Thapar 2015). There are two kinds of natural experiments that are applicable and potentially useful in assessing the benefits and harms of medicines. However, it should be noted that instances where these sorts of natural experiments might arise are rare, limiting the extent to which they can be used in practice.

First, there is the situation of universal exposure whereby the choice of whether or not to use a pharmaceutical agent is decided by governments rather than by the individual. The example that follows constituted a natural experiment through the combination of universal coverage and the use of a vaccine, and international variations in its use. The measles, mumps, rubella (MMR) vaccine had been alleged to have led to an epidemic in the rate of autism. The opportunity to test this claim arose because the vaccine had been withdrawn in Japan at a time when its use continued in other parts of the world (Honda, Simizu & Rutter 2005). These before-and-after studies combined with cross-country studies suggested that there was no link between the MMR vaccine and autism. There was a similar situation with the use of thiomersal (a mercury preservative used in many vaccines). This stopped being used in Scandinavia but it was continued to be used in the rest of the world (Atladóttir *et al*. 2007). The comparison between countries in this natural experiment suggested that thiomersal was not linked to a rise in autism. A main limitation of this design is that it is impossible to randomise participants in these situations, and it is possible that other differences between the groups (in these two examples, the countries) might have confounded the observed associations.

The second natural experiment design that might be applicable to the study of drug effects is the use of a regression discontinuity design. Briefly, in this design participants are assigned to groups (e.g. treatment or control groups, or different treatment groups) based on whether they sit above or below a cut-off score on an 'assignment variable'. The assignment variable is a measure taken pre-test (e.g. illness severity measurement) and determines which group the participant is assigned to. Post-test scores and regression lines are then plotted for each group on the same graph, allowing the displacement of the regression lines for each group to be compared. If there is no or little displacement of the regression lines, the treatment is unlikely to have an effect on the outcome (**Figure 1a**), whereas if there is a marked displacement of the regression lines, the treatment is likely to have an effect on the observed outcome (**Figure 1b**) (Shadish, Cook & Campbell 2002). As such, the approach focuses on the distance between regression lines rather than mean scores.

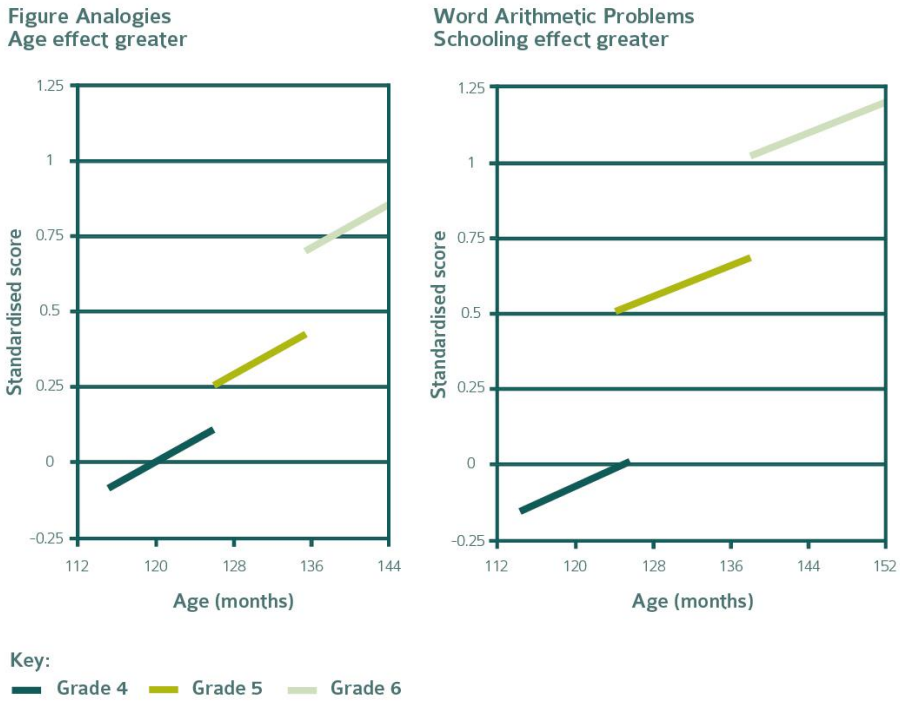
Although this approach has some similarities to an RCT, it is much more demanding in its requirements and makes major assumptions; therefore, it is not surprising that it has been rarely employed. This method was used as part of a study that looked at policies on the age at which children began schooling or changed from primary to secondary schools to determine whether duration of education or age was more influential for educational attainment (Cahan & Cohen 1989; **Figure 1**). It found that duration of education appeared most important for the majority of the educational outcomes. The approach could conceivably be used for: antidepressants when these are only given when some score on a depression scale is above a predetermined threshold; stimulants used in the treatment of attention deficit hyperactivity disorder (ADHD) when given only above the cut-off on some scale; or drugs to lower blood pressure that would only be given once a certain level had been exceeded. As far as we know, they have not been used in this way, but there may be some benefit to using this approach, possibly as a precursor to deciding whether there is value in proceeding to intervention development and RCT assessment.

Figure 1. The effects of children's age and schooling on cognitive performance

A diagrammatic representation of the regression discontinuity design dealing with duration of schooling as the treatment under consideration.

a. Increase in age

b. Duration of schooling



The notable displacement between the regression lines of the three groups in Figure 1b suggests that duration of schooling has a marked effect on attainment outcomes in word arithmetic problems. In contrast, the small displacement between the regression lines of the three groups in Figure 1a suggests that duration of schooling has less of an effect on attainment outcomes in figure analogies; in this instance, further results indicate that age has more pronounced effect. The study population consisted of all the fourth, fifth, and sixth graders attending Jerusalem's Hebrew-language, state-controlled elementary schools in 1987 (adapted from Cahan & Cohen 1989).

4.3 Mendelian randomisation as a way of inferring causality

For many years, instrumental analyses have been considered as an additional or alternative method to multivariable analyses in both RCT and observational study designs, as a way of minimising confounding and bias due to measurement error in the risk factor/treatment of interest (McClellan, McNeil & Newhouse 1994; Korn & Baumrind 1998; Newhouse & McClellan 1998). In these analyses, an instrumental variable is identified that, in essence, supplies a source of variation in treatment that is equivalent to randomisation. For example, an instrumental variable might be the distance to hospital or a specific genetic variant (see below) which is used to assign study participants to different treatment or control groups. There are three key assumptions for instrumental variable analyses – the instrumental variable must: (1) influence the treatment (or risk factor); (2) not be influenced by confounding factors for the treatment-outcome association; and (3) only affect the outcome via their effect on the treatment (i.e. there is no path independent of the treatment/risk factor from the instrumental variable to the outcome) (Glymour, Tchetgen & Robins 2012; Greenland 2000; Lawlor *et al*. 2008). This last assumption is known as the 'exclusion restriction criteria' and is the one that tends to cause most concern when used in RCTs or Mendelian randomisation studies (see below). Instrumental variables can also be used in RCTs to get a measure of local average treatment effect (i.e. the effect of the treatment in those who take it as prescribed) (Greenland 2000).

Mendelian randomisation is the specific form of instrumental variable analysis that uses genetic variants as instrumental variables. There is empirical evidence that genetic variants are not related to the large number of confounding factors that tend to affect observational associations using multivariable regression (Davey Smith *et al*. 2008). Supplementary material in the Wellcome Trust Case Control Consortium shows remarkable consistency of genome-wide data between two control groups, one of which (blood donors) was highly select and the other a non-select population cohort (1958 national birth cohort), suggesting selection bias is unlikely with genetic based studies (Wellcome Trust Case Control Consortium 2007). Further, since genetic variants are 'allocated' at conception and cannot be changed by subsequent disease or behaviours, 'reverse causality' is unlikely here. For these reasons Mendelian randomisation has been linked to 'nature's randomised controlled trial' (Hingorani & Humphries 2005). Recent developments to this method have focused on assessing and limiting the impact of direct pleiotropy, where a gene has effects on several different outcomes. Indeed, pleiotropy results in violation of the exclusion restriction criteria that states that the instrumental variable should only affect the outcome via its effect on the treatment.

Mendelian randomisation was used to study whether increased high-density lipoprotein (HDL) cholesterol (the 'good' type of cholesterol) would lower the risk of myocardial infarction (Voight *et al*. 2012). Consistent with recent RCT evidence of drugs that increased HDL cholesterol (Schwartz *et al*. 2012; Barter *et al*. 2007), this study suggested that there was no causal effect of HDL cholesterol on coronary heart disease (CHD). This study is useful for highlighting some of the strengths and limitations of Mendelian randomisation. In brief, a very large sample (over 12,000 cases of myocardial infarction and 41,000 controls) was used, which is often the case with Mendelian randomisation studies as they tend to have low statistical power. In addition to the main study exploring effects of HDL cholesterol on CHD, the authors undertook a positive control testing the effect of LDL cholesterol (the 'bad' type) on CHD, which supported previous Mendelian randomisation studies and RCTs of statins by providing evidence that suggested higher LDL cholesterol causes an increased risk of CHD. The consistency of these findings for both HDL cholesterol (no causal effect) and LDL cholesterol with RCT evidence strengthens the likelihood that the findings are correct, and highlights the value of integrating different study designs. However, the authors highlighted the difficulty for Mendelian randomisation (and RCTs) of truly differentiating effects of different lipids when they are biologically and statistically correlated with each other (Würtz *et al*. 2016).

4.4 Areas where new strategies are required

4.4.1 Research in rare diseases

Although rare diseases are individually rare, they are actually common when considered all together. Research into rare diseases clearly involves the same need for robust evidence on the efficacy of novel treatments. RCTs, which are able to generate causal evidence with minimal bias and confounding, would be the ideal method to generate such evidence but present particular difficulties. The first practical difficulty is that patients with such diseases are likely to be geographically scattered. Collaborative research networks may help in pooling patients, and patient groups might be able to help in

mobilising these (Goss *et al*. 2002; Rowe *et al*. 2016; Corbyn 2012). The Cystic Fibrosis Foundation took such an approach to aid the development of a new drug, ivacaftor (trade name: Kalydeco) (see **Box 11**).

A methodological review of innovative methods for evaluating treatment effects for rare disease found more methodological developments of RCTs than of observational studies (Gagne *et al*. 2014). Several methods have been developed for use with RCTs that reduce the total sample size required and maximise the number of patients who are randomised to the new treatment, whilst maintaining the strengths of RCTs in general. In observational studies, propensity scores and the use of 'self-controls' (i.e. before-and-after studies) were noted to be frequently used for exploring treatment effects in patients with rare diseases. Given the nature of rare diseases, where appropriate, continuously measured outcomes, which have greater statistical efficiency, may be preferable to binary or other categorical outcomes in both RCT and observational studies. Whilst earlier we cautioned against the overuse of surrogate outcomes, in the case of rare diseases their use, particularly where these are continuously measured (such as biomarkers), may be valuable as long as their validity has been established. Adaptive randomisation (randomisation changes during the enrolment period to increase the proportion randomised to the treatment that appears more effective) or sequential designs may maximise participation, as will crossover designs (see **Box 12**). Propensity scores may help by modelling exposure rather than outcomes.

Case studies may also be of particular value in studying rare diseases. Those involving a mix of both qualitative and quantitative methods have been used in legal cases, ethnographies, policy evaluation and post-hoc fault diagnoses to establish causality in single cases and might have value in health research. They can suggest cause-effect relations for individuals in a population, and could therefore be used as the basis for more robust methods aimed at estimating the magnitude of causal effects at a population level (Byrne & Ragin 2009).

None of the advances in study design and methods described above provides a totally satisfactory solution to the difficulties involved in treatments for rare diseases but, in combination, they may help (Gupta *et al*. 2011). In particular, we would endorse greater effort to fund large (international) collaborative networks to support research in rare diseases.

Box 11. The Cystic Fibrosis Foundation

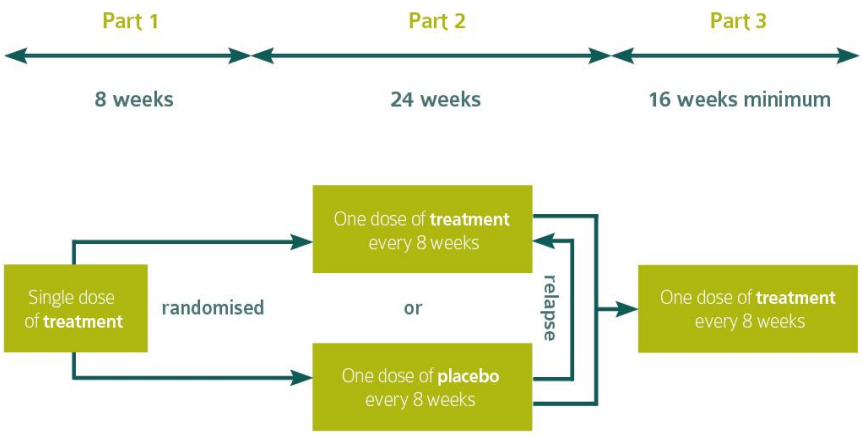
The Cystic Fibrosis Foundation was founded in 1955 in Philadelphia by concerned parents of children affected by this rare genetic disease. At the time, few children with cystic fibrosis (CF) lived to attend primary school. In 1998, the Foundation launched the Cystic Fibrosis Therapeutics Development Network, the largest CF clinical trials network in the world (Goss *et al*. 2002). They also began to pursue an innovative venture philanthropy model of drug development, providing early stage funding to biotechnology and pharmaceutical companies to develop breakthrough drugs for adults and children with cystic fibrosis.

This model proved successful in 2011 when Phase III clinical trials of a drug developed in collaboration with Vertex Pharmaceuticals showed significant results. The drug, ivacaftor (trade name: Kalydeco), is the first to treat the underlying causes of CF rather than the symptoms, and is highly effective in individuals with specific genetic mutations (Corbyn 2012). Nowadays, people with CF can be expected to live well into adulthood.

Box 12. Example of a crossover design that was used in the evaluation of a rare disease treatment

A variation on the crossover design was used in a trial of the monoclonal antibody canakinumab as a treatment for cryopyrin-associated periodic syndromes, a spectrum of rare, inherited inflammatory disorders (Lachmann *et al*. 2009). The drug was initially given to all participants before a randomised withdrawal period. At the end of this period, or at the time of relapse, patients were put back on the drug (see **Figure 2**). This method allowed the safety and efficacy of canakinumab to be explored without denying patients access to a promising new therapy.

Figure 2. A schematic illustration of the Lachmann *et al*. (2009) crossover trial



4.4.2 Research in emergency situations

Research in emergency situations provides a further set of challenges, as the recent Ebola virus disease outbreak in West Africa has illustrated. At the time of the recent outbreak, several experimental therapies had been developed or proposed, but none had been tested for efficacy in humans prior to the 2014 epidemic. In response to the outbreak, several clinical trials were launched, after much debate about appropriate trial designs. Different methodologies were adopted, some of them innovative, and only one was a RCT (albeit an unconventional RCT with a quasi-Bayesian adaptive design). This was because the use of a placebo or usual care (which was not therapeutic) was considered impractical and unethical, and blinding was also considered to be impractical (Adebamowo *et al*. 2014). The rapidity of the spread of the disease also created problems, as did the short duration of the disease, the geographical dispersion of the infection (in this case in resource-poor locations), and the high mortality rate for both patients and healthcare staff (and potentially research staff). Trial designs therefore needed to generate information quickly to inform treatment and the control of the disease, whilst ensuring these efforts did not put additional people in danger or worsen the situation. One viable approach in a unique situation like this would be to try different treatments in parallel at different sites. Although fully-randomised, well-blinded RCTs would have provided the most robust evidence on safety and efficacy, in this emergency situation the conditions for which reliable evidence on treatment effects can be obtained from observational studies were satisfied. Indeed, any promising treatment would have had a large effect (survival) on a poor outcome (death) and

the treatments were devised based on plausible biological mechanisms. Nevertheless, the US National Institutes of Health have announced that they will be undertaking an RCT in Liberia and the US to further evaluate the safety and efficacy of the investigational drug ZMapp as a treatment for Ebola virus disease (National Institutes of Health 2015).

The issues with respect to the testing of vaccines to *prevent* infection with Ebola are somewhat different to those of the emergency situation of *treating* the disease, where affected individuals are at high risk of dying within a short timeframe. The ethical implications therefore do not apply to the same extent. A double-blinded RCT of vaccine efficacy has been undertaken in Liberia (Kennedy *et al*. 2016).

4.4.3 Novel research methodologies in rare diseases and emergencies

There has been a drive to develop novel trial designs that address the constraints that arise when assessing the benefits and harms of medicines for rare diseases or in emergency situations. For example, there are currently a number of EU-funded initiatives attempting to further develop research methodologies for rare diseases, including Advances in Small Trials dEsign for Regulatory Innovation and eXcellence (ASTERIX) (ASTERIX project 2016), Integrated DEsign and AnaLysis of small population group trials (IDeAL) (Integrated DEsign and AnaLysis of small population group trials 2016), and INnovative methodology for Small PopulatIons REsearch (InSPiRe) (InSPiRe 2016). The recent Ebola crisis also generated much debate surrounding appropriate methodologies to adopt in emergency situations, and many institutions are now reflecting on the lessons learnt from the outbreak, including the UK House of Commons Science and Technology Committee and the Academy (www.parliament.co.uk 2015; Academy of Medical Sciences 2015b). We await the outputs from these initiatives to determine their impact on research methodologies.

4.5 Future challenges

4.5.1 Stratified medicine

Stratified medicine (also known as personalised or precision medicine) refers to the grouping of patients based on risk of disease, or response to therapy, using diagnostic tests or techniques (Academy of Medical Sciences 2013a). Stratified medicine has the potential to benefit: patients and healthcare providers, through the development of more targeted and effective treatments; the healthcare system, through improved efficiency and capitalising on healthcare gains; and industry, through more efficient therapeutic development and access to an expanded market for specialised treatments. As medicines become more specialised and the precision required moves towards the level of the individual patient, the stratification of disease has the potential to make almost every disease a rare disease with the corresponding challenges raised in the section above. Accordingly, assessing the safety and efficacy of such treatments may require smaller and more specialised trials, designed in a way that allows the safety and efficacy of new treatments to be evaluated in small subgroups of a tested population while ensuring they have sufficient statistical power. **We would encourage regulatory authorities and researchers to engage with each other to determine whether alternative methodologies (such as adaptive designs) and data sources could be informative when assessing the safety and efficacy of such specialised novel treatments.**

4.5.2 'Real world' evidence

There is a growing awareness of the potential for routine health and related records to provide data that, alone or integrated with observational or trial study data, could be of immense value in assessing both the benefits and harms associated with therapeutic drugs (see **Box 13** for a recent example). The availability of electronic health records (EHRs) can aid this activity and the UK's Clinical Practice Research Datalink (CPRD), which at present covers only a small minority of the population, could be used much more widely. The well-established Yellow Card Scheme for reporting adverse outcomes in particular could be combined with other aspects of EHRs to support the rapid identification of important side effects of new treatments. Without such linkage, individual case reports (e.g. based on the Yellow Card reporting system) can rarely be used to estimate the frequency or impact of adverse events because they lack a denominator in terms of the relevant at-risk population. With appropriate use of linked EHRs this issue could be ameliorated, though under-reporting of such events is likely to remain a problem.

Box 13. The use of GP records to examine the link between ethnic differences in depression rates and neighbourhood ethnic density

A recent study used data from GP records to examine whether the diagnosis of depression and antidepressant prescribing varied by location and ethnic density (Schofield *et al*. 2016). This had been investigated previously but the results had been inconclusive. By accessing primary care data, the authors were able to cover over one million patients of Indian, Pakistani, Bangladeshi, Black Caribbean and Black African origin in four London boroughs. Such a large dataset enabled them to conclusively determine that new depression diagnosis and antidepressant use was less likely in areas where there was a higher density of people with the same ethnic background for some, but not all, ethnic groups. It was noted that certain antidepressants can be prescribed for other indications (such as pain control), in addition to a mental health indication. A potential limitation of the study was that the GP records did not indicate whether the antidepressant therapies were prescribed for a mental health indication or not.

The UK has a number of disease registries of various kinds that have, by themselves or in combination with other data, sometimes provided useful insights into treatment effects (Toschke *et al*. 2011). However, to date a systematic approach to linking them all and making the data readily accessible, in the way that registers in Scandinavia and some other countries operate, has been lacking. In Scandinavia, a unique personal identifier known as the Central Population Registration (CPR) number is used to easily link data from healthcare records, disease registries, and national databases (Furu *et al*. 2010; Lone 2003). This can for example enable the tracking of individual diseases in prescription data over time (Haerskjold *et al*. 2015) or improve postmarketing surveillance of drugs (McNeil *et al*. 2010). People have been concerned about privacy, but Scandinavia has led the way in dealing successfully with these issues by making data available to researchers in an irreversible, encrypted version, thereby carrying the general population with them (Haerskjold *et al*. 2015).

In Scotland, National Records of Scotland acts as a trusted third party (TTP) that enables de-identified health and non-health data to be linked. After appropriate approvals, access to de-identified data is provided to researchers in data safe havens (such as the Farr Institute Scotland), secure environments where data can be accessed and used for research while upholding data subjects' confidentiality and rights to privacy (for further information, please see Information Services Division (ISD) Scotland 2010; The Scottish Government 2012 and 2014).[6] A similar approach for linking health data using a TTP and providing data access to approved researchers in data safe havens is utilised in Wales (SAIL Databank 2016). We welcome the advances in developing data linkage in Scotland and Wales. However, because an equivalent to the Scandinavian CPR number is not available for linkage, a complex system using a TTP and research safe havens is needed to allow linkage whilst ensuring confidentiality. For example in Scotland, the Community Health Index (CHI) number, a unique number allocated to all individuals registered with a GP (ISD Scotland 2016), cannot be used for linkage without the involvement of a TTP because it links to personal information contained in the CHI register. Looking to the future, there are cultural, ethical and technological issues that need to be addressed to enable linkage of all relevant data across the entirety of the UK in order to provide useful data for evaluating evidence.[7]

For reasons given in McNeil *et al.* (2010), registries rarely have the precision of a randomised controlled trial, although

[6] In March 2014, the Academy held a workshop in partnership with the Medical Research Council and the Wellcome Trust on *Data in safe havens* to discuss the meaning and adequacy of data safe havens, and the next steps in their development. Further information is available on our website: http://www.acmedsci.ac.uk/policy/policy-projects/data-in-safe-havens/

[7] In September 2015, the Academy, in partnership with the ABPI, held a workshop on *Real world evidence* to explore aspirations and opportunities for future use of real world evidence in a regulatory context. Further information is available on our website: http://www.acmedsci.ac.uk/policy/policy-projects/real-world-data/.

they can provide useful information.

The term 'real world' has become accepted but many people find it misleading because it implies that RCTs, for example, do not deal with the real world when obviously they do. For example, pragmatic trials, by focusing on clinical effectiveness, most obviously do so.

In addition to the initiatives mentioned above, various others have considered, or are currently exploring, the use of new data sources. Examples are given in **Box 14**. We welcome these initiatives that will help clarify how so-called 'real world' evidence can be integrated into the medicines development process and how it can best be used to assess the safety, efficacy and effectiveness of medicines. However, such analyses are prone to bias and confounding, and therefore the limitations to such approaches should be considered when evaluating the evidence generated as part of these 'real world' studies.

Box 14. Examples of initiatives exploring the use of new data sources

- The UK Farr Institute of Health Informatics Research, which specialises in research linking electronic health data with other forms of research and routinely collected data (The Farr Institute 2016).
- The EU Innovative Medicines Initiative (IMI) GetReal project is currently examining how so-called 'real world' evidence can more effectively, and at an earlier stage, inform pharmaceutical research and development and the healthcare decision making process (Innovative Medicines Initiative 2013).
- The EU IMI WEB-RADR (Recognising Adverse Drug Reactions) project is currently researching the utility of publicly available web and social media content to detect new drug side effects, and the right methodologies for harnessing these data (Innovative Medicines Initiative 2014).
- The EU IMI PROTECT (Pharmacoepidemiological Research on Outcomes of Therapeutics) project looked at ways of using data from different sources to strengthen the monitoring of the benefit-risk of medicines and the early detection and assessment of adverse drug reactions (IMI PROTECT 2009).
- The US Food and Drug Administration (FDA) is also exploring how so-called 'real world' evidence can be better utilised and has established the Sentinel Initiative to identify and investigate safety issues in near real-time (US Food and Drug Administration 2015).

Recommendation 4

Electronic health records, research databanks and disease registries are valuable sources of so-called 'real world' data and we recommend that their use should be explored by researchers and regulators, and in HTA assessments. In developing an approach for access to and linkage of data in the UK, attention should be paid to approaches such as those in Scandinavia, where the use of unique personal identifiers, supportive infrastructures and appropriate governance have enabled the straightforward linkage of data, and anonymity is protected by making data available to researchers in an irreversible encrypted fashion.

4.5.3 Patients with multiple illnesses

A separate need concerns the challenges associated with an increasingly aged population with multiple illnesses (known as comorbidities). Many such patients may be excluded from RCTs, although they represent a significant group that would potentially stand to benefit from novel treatments. Accordingly, **more research is needed into moderating variables that affect the response to treatments**. **Modelling and simulation may be explored in the future as a means of providing supporting evidence in subpopulations for which limited experimental data are available**, including older people and children. It could also be used to ascertain whether a trial is necessary in particular target populations, and if so, could help to refine or optimise the protocol design (Saeed, Vlasakakis & Della Pasqua 2015). Further, it could help generate information on drug-drug or drug-disease interactions, which are particularly important for patients with multiple illnesses. A key issue here is how this type of evidence can be appraised by regulatory authorities. The European Medicines Agency (EMA) has developed guidance on techniques for trials in small populations and extrapolation, which might be useful in this context (for example in paediatrics, see European Medicines Agency 2016).

5. Further issues for consideration when assessing research findings

In previous chapters, we have explored the research issues that should be considered when designing studies, and analysing, interpreting or appraising clinical evidence, described how these research issues affect traditional study types, and provided an overview of approaches that have evolved with the aim of addressing some of the remaining limitations. There is, however, a set of wider issues that are associated with the evaluation of evidence and which warrant further consideration. In this chapter, we briefly explore concerns around research reproducibility and reliability, publication bias, working with industry, and concerns about over/underuse of medicines.

5.1 Research reproducibility and reliability

Reproducibility is a core principle of scientific progress (Nosek *et al*. 2015).[8] Yet, it has been pointed out that it is likely that most published findings are chance findings in that they do not replicate when the same methods are applied in a different sample (Ioannidis 2005). In recent times, there have been systematic attempts to determine how far this charge

[8] Results are regarded as reproducible when an independent researcher conducts an experiment under similar conditions to a previous study and achieves commensurate results.

is, or is not, correct. For example, the Open Science Collaboration examined the extent to which 100 experimental and correlational study results published in three psychology journals could be replicated, and reported that very few of these could (Open Science Collaboration 2015). The Open Science Collaboration is also attempting to independently replicate selected results from 50 papers in cancer biology (Reproducibility Project: Cancer Biology 2013). Similar problems with relatively limited replication have been found within the pharmaceutical industry, which has joined other research sectors in calling for reproducibility to be improved (Prinz *et al*. 2011; Begley & Ellis 2012; Ioannidis *et al*. 2014). **Efforts to enhance the reproducibility and reliability of research will require co-operation across the research landscape with research funders, publishers, research institutions, professional bodies and individual researchers working together globally.** The Academy, Biotechnology and Biological Sciences Research Council (BBSRC), MRC and Wellcome Trust recently held a symposium to explore how issues of reproducibility can be tackled in the UK and have committed to developing and implementing changes, while working alongside international partners to facilitate action on a global scale (Academy of Medical Sciences 2015c).

For those evaluating evidence, independent replication of results increases the confidence that the findings are likely to be true. It should be noted, however, that for findings to be deemed reproducible, the results obtained and the methods employed in replication studies do not need to be exactly *identical* to the original study. Indeed, results can be viewed as reproducible if independent researchers conduct experiments under *similar* (but not necessarily identical) conditions and achieve commensurate results. If the finding is robust, minor changes in the details of the methods and results should not matter, and can provide evidence that the findings can generalise beyond the specific conditions of the original study (Robins 1978). In addition, it needs to be recognised that there are valid reasons why there might not be a high level of replication between studies and the lack of replication might be informative in itself (see Lindsay 2015 for further details).

A key consideration when designing clinical trials is to ensure that they have adequate statistical power – that is, they are of adequate size to detect an effect, should that effect exist (Friedman & Schron 2015). Many trials are underpowered and this has a negative effect on the likelihood of replication. A recent study into the reproducibility of findings in psychiatry found that among 83 highly cited studies claiming effective psychiatric treatments from 2000–2002 and recommending effective interventions, 40 had not been subject to any attempt at replication, 16 were contradicted, 11 were found to have substantially smaller effects and only 16 were fully replicated (Tajika *et al*. 2015). There must be caution exercised when reading a report based on a small sample that reports a very large effect, because it is highly likely to be false. The Open Science Collaboration examination of psychological studies mentioned above echoes these findings: the replication effects were found to be only half the magnitude of the original effects, with replication being better with higher quality studies using large samples (Open Science Collaboration 2015; Patil, Peng & Leek 2016). **We recognise the importance of funding for large-scale trials and commend the schemes from major funding bodies that fund such research. Requiring applicants to submit details of sample size calculation (with justification of choice of effect size and outcome) and underpinning justification for their research in their funding applications is clearly essential, and we encourage funding bodies to continue to request this.**

A different yet related concern has been the distorting effect of measurement error. It has been shown through both simulations and actual experiments that undisclosed flexibility in data collection and analysis (i.e. unreported divergence from the original research aims and protocol) leads to a high likelihood of false positive findings because of unconscious bias in both analyses and reporting (Simmons, Nelson & Simonsohn 2011). There is abundant evidence of the pervasiveness of unconscious bias (in all individuals) and it is clear that it could influence expectations of outcome. What remains unknown is the extent to which training succeeds in improving the situation. Concerns such as these led to the CONSORT (Consolidated Standards of Reporting Trials), STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) and PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines, which have helped in providing a means of standardising structured reporting of randomised trials, observational studies in epidemiology, and systematic reviews and meta-analyses respectively (The CONSORT Statement 2010; STROBE Statement 2009; PRISMA 2015). Despite the usefulness of these guidelines, there are concerns that they have not always been adequately adopted (Rutter & Pickles 2015).

Recommendation 5

We recommend that guidelines such as CONSORT, STROBE and PRISMA, among others, be comprehensively adopted and that funding bodies require their grant awardees to adhere to these reporting guidelines. We encourage research institutions and journals to provide incentives for their use and to better enforce their adoption.

Recommendation 6

We recommend that higher education institutions and research institutes make training on research integrity, which should include elements relating to unconscious bias in biomedical research, mandatory and that they assess the effectiveness of such training programmes.

Rather than assume all biases are equally damaging, it is desirable to identify and model all internal and external biases. A strategy for identifying and modelling internal and external biases has been devised; however, at the present time, determining the presence of bias has to be based on a considered judgement rather than on some methodological algorithm (Turner *et al*. 2009).

New models have recently emerged to try and address some of the issues associated with false positives that occur as a result of data dredging – that is, the repeated analysis of a dataset until a statistically 'significant' result is found (also known as 'p-hacking'). These involve pre-registration of protocols, which set out in advance how the results will be collected and analysed, with a journal or a data repository before the experiments are carried out (Registered Reports 2014; Open Science Framework 2011). Such prior specification of methods aims, among other things, to guard against multiple analyses once the results are obtained. Protocol registration is relatively common practice for RCTs and has played a key role in driving up standards. Registration of protocols and pre-specification of methods should become routine for more basic research involving human subjects and in epidemiological observational studies, where this is not currently common. We note that the Health Research Authority, which aims to promote and protect the interests of patients in health research and to streamline the regulation of research, has statutory duties to promote research transparency. As such it expects all clinical trials to be registered, and all researchers, research sponsors and others to meet this best practice standard (Health Research Authority 2015).

Although it is in principle highly desirable to specify methods in advance, it is also important to appreciate that initial analyses may engender further analyses (Rutter & Pickles 2015). In this instance, **it is critical that an explicit explanation is provided as to why additional analyses are required and to specify how these are to be undertaken in order to avoid data dredging and the consequent false positives.** Exploratory analyses are clearly important and should be conducted wherever necessary; however, they should be explicitly presented as such, and not misleadingly as testing prior hypotheses.

Improving the quality and integrity of research clearly has implications for the training of researchers and mentorship programmes. For example, researchers at all levels should receive training in methodology and statistics to ensure that the evidence they produce is robust and of high quality. More widely, a change in research culture is needed to incentivise the production of reliable findings (for an in-depth discussion see The Lancet series on 'Research: increasing value, reducing waste' 2014).

5.2 Publication bias

Publication bias (or the failure to publish all research results, often favouring the reporting of so-called 'positive' results at the expense of null, inconclusive or 'negative' results) was widely raised as an ongoing concern in the responses to our call for evidence. Failure to publish research results is particularly problematic for the assessment of treatment effects, as analysis of the published evidence provides a skewed evidence base if null, 'negative' or inconclusive results are not available. There have been concerted efforts from funding bodies and journals to incentivise the publication of these results, which may be viewed as less exciting and thereby less worthy of publication. Nevertheless, dramatic changes in approaches for dealing with null, 'negative' or inconclusive findings are still required (Rutter & Pickles 2015). **We support the push for a cultural shift within the scientific community that ensures that all results are published (or made publicly available), including null, 'negative' or inconclusive results.** We do not intend to provide a detailed account of all of the issues here, and instead refer readers to comprehensive published reports for further information on 'fashions' in publishing, the problems associated with selective publication of findings, and the pressures created by both employers and funding agencies to claim that some finding has much more importance than it warrants (Hampton 2015; Rutter & Pickles 2015).

With regards to publication of clinical trials, the compulsory registration of trials (which is a condition of publication for many journals) and publication of summary results on public registries has helped to increase transparency about the trials that are undertaken and their initial findings (ClinicalTrials.gov 2012; EU Clinical Trials Register 2012; De Angelis *et al*. 2004 Rawal B & Deane BR 2014; Deane BR & Sivarajah J 2017). However, studies exploring compliance with the requirements of trial registration have shown that there needs to be continued focus on ensuring that these are enforced (Ross *et al*. 2012; Prayle, Hurley & Smyth 2012; Wieseler *et al.* 2012).

The Academy believes that the existence, methods and results of clinical and health research involving patients – whether positive or negative – should be made swiftly available for patient, social and scientific benefit (Academy of Medical Sciences 2013b). Due consideration should also be given to access to patient-level data from both academic and industry-sponsored trials, and this should be provided in a manner that protects patient confidentiality and ensures that the data are intelligible, assessable, reliable and usable. The Academy supports the principles underpinning the AllTrials campaign, which calls for all past and present clinical trials to be registered and their full methods and summary results reported (Academy of Medical Science 2013b; AllTrials 2014).

Related issues include drug companies using ghost writers (individuals who produce a manuscript for publication, specified to order by the pharmaceutical company, but who are not included in the author byline or the acknowledgements – see below) and suppressing results that they find inconvenient, as recently suggested for the underreporting of harms associated with antidepressant treatment (Sharma *et al*. 2016). These findings show major problems in study design and discrepancies in reporting. While this is concerning, we welcome the European Medicines Agency decision that all newly submitted reports be made publically available (European Medicines Agency 2014). We await evidence that this has resulted in action.

5.3 Working with industry

The Academy believes that a vibrant innovation ecosystem and strong links between academia, industry, the NHS and the regulatory sector play a key role in addressing health and scientific challenges, by capitalising on strengths across the healthcare sector and shared expertise, skills and resources (Academy of Medical Sciences 2013c). Working with the pharmaceutical industry has particular advantages for drug development. Indeed, industry can provide the funding necessary to take forward costly RCTs that are of appropriate size for generating meaningful results, and collaboration with industry allows external experts to input into the development of treatments that could have a real clinical benefit, particularly in disease areas that might not be an obvious priority for industry. These issues will be considered more fully by the oversight group. However, these partnerships should be conducted in a manner that minimises biases and undue influence on the results to ensure that the findings are robust and reliable.

Funding for clinical research can come from a variety of sources (see **Box 15** for further details). In the wider workstream of which this working group is part, we have been exploring how sources of funding (or other potential conflicts of interest) might impact on the generation or interpretation of medical evidence, and how conflicts can be effectively

managed.[9] One aspect of this is whether commercial pressures may influence what research is carried out, how it is carried out, whether and how it is disseminated, and the analysis of evidence in decision-making. With closer links between academia and industry, there are concerns that these pressures may also influence those working within the academic sector. A number of general issues have arisen during the course of this working group:

- **Study design**: Consideration should be given as to how involved industry should be in the study design and whether research studies should be developed by the academic researcher alone or in conjunction with industry. We note that the design can be enhanced by the peer-review process (e.g. pre-registration of protocols with journals).
- **Autonomy in data analysis**: Whether the data analysis is conducted entirely independently by the academic researcher, and whether it should be free of any constraints imposed by the funder, should also be considered. We note that regulators expect industry funders to be able to robustly defend every aspect of the study – from design through analysis and reporting – and they will challenge and inspect all elements of the evidence submitted to them. This might make it harder for the analysis to be conducted entirely independently from the funder, who will ultimately need to know and understand all the features of the design and analysis of the research. Further, such independent analysis might be impractical in situations where investigators are participating in a multi-centre study.
- **Data holding**: Thought should be given as to whom should hold the data set – the academic researcher, industry, or both.
- **Data access**: Access to data should be given due consideration, both in terms of publication of the results regardless of the outcome and timeliness of publication. We note that there may be tension between the priorities of academic researchers (who might want to retain exclusive access to the data for further research) and industry (who are increasingly required to be transparent).
- **Personal payment**: The impact of academic researchers receiving *personal* payment for their work from industry funders should be considered, particularly in terms of how it might affect the perception of bias and trustworthiness of the study. Alternative models where funding is placed into a general fund for the academic's research group should be considered.
- **Trial registration**: Registration of both academic and industry trials, and publication of summary results on public registries should be explored.

Consideration of these issues when academic research is supported by industry might help to address some of the concerns about the validity of clinical research. We anticipate that with the advent of new development pathways (such as adaptive pathways), industry will increasingly be conducting studies throughout the entire lifecycle of a medicine, including post-approval safety and efficacy studies. This will offer further opportunities for academia-industry collaboration beyond drug development once a medicine has been approved. Such future opportunities further emphasise the need to address concerns about the validity of academic research when supported by industry.

Recommendation 7

We support the registration of RCTs and mandatory publication of their protocols by researchers as a means to help ensure that so-called 'negative', null or inconclusive trial results become publicly available, and to enhance research reproducibility. We recommend that similar methods be adopted for observational epidemiological studies exploring the effects of treatments, where currently such an approach is lacking. We encourage journals to make registration of such observational epidemiological studies and publication of their protocols a condition of publication, as is the case for clinical trials.

[9] Further information on this work strand is available on our website: http://www.acmedsci.ac.uk/policy/policy-projects/conflicts-of-interest-workshop/

Box 15. Funding for clinical trials

Clinical trials can be expensive endeavours. Research costs that have to be covered include: the treatments included in the study; research staff to conduct the trial, and collect and analyse the data; administrative costs; and any tests or hospital stays.

Different types of funding are available for clinical trials. Funding sources include:

- Government agencies, such as the MRC and NICE. (Medical Research Council 2016; National Institute for Health Research 2015).
- Medical research charities, such as Arthritis Research UK (2016), the British Heart Foundation (2016), Cancer Research UK (2016), and the Wellcome Trust (2016), among many others.
- The pharmaceutical industry.

Pharmaceutical companies will sometimes choose to conduct their own trials of treatments they have developed, or to outsource such research to a contract research organisation that will run the trial on their behalf. If companies would like a trial to be carried out independently from their organisation, they might choose to collaborate with academia and provide money or supply the treatment free of charge to an academic organisation that will carry out the research in an independent fashion (Pfizer 2016; Roche 2015).

Recommendation 8

We recommend that the Academy, through the oversight group, looks further into the principles governing relationships between academia and industry to address concerns about the validity of academic research funded by industry.

5.4 Concerns about over or underuse of medicines

The aim of this report is to examine how evidence can best be evaluated to determine the benefits and harms of treatments. As both under and over medication can result in harm, relevant methodological considerations are an appropriate part of our deliberations. Overmedication in particular has recently been widely debated in the general and specialist press in relation to the use of statin therapy. We recognise that many elements – such as communication, real or perceived interests, and public attitudes to medicines – are important factors to consider when addressing this issue. These aspects will be explored in detail by the oversight group and discussed in its report. We focus here on the methodological issues that are relevant to the over or underuse of medicines, using statins, vaccination, medicines to treat ADHD, depression and bipolar affective disorders, and antimicrobials as examples.

5.4.1 Use of statin therapy

In 2014, NICE recommended that the threshold for offering statin treatment for the primary prevention of CVD be lowered from a 20% or greater 10-year risk of developing CVD to a 10% or greater risk (National Institute for Health and Care Excellence 2014a). This decision was based mainly on two meta-analyses, which both suggested that the guidelines for primary prevention be reconsidered:

- A Cholesterol Treatment Trialists' Collaborators meta-analysis of 27 clinical trials, which showed that statin therapy significantly reduced the risk of major vascular events even in individuals with a five-year risk of major vascular events to lower than 10% (Cholesterol Treatment Trialists' Collaboration 2012a).
- A Cochrane Review, which reported that statins were associated with a reduction in all-cause mortality, major vascular events, and revascularisation in patients without established CVD, and were not associated with adverse events (Taylor *et al*. 2013).

Despite these studies, the resulting change to the NICE guidelines received strong objections, with concerns over: the generalisability of results to patients beyond those included in the trials; the (over)medicalisation of healthy individuals; and whether the potential benefits of reducing CVD in low-risk patients outweighed the risks of side effects associated with a treatment that would need to be taken lifelong, amongst others (National Institute for Health and Care Excellence 2014b).

The use of statins has been studied in a wide range of different treatment populations. With respect to the relative benefits, evidence from RCTs (or meta-analysis thereof) suggests that the use of statins is beneficial in reducing CVD risks, largely irrespective of an individual's background risk, gender or age. Indeed, a recent meta-analysis of 27 trials concluded that statin use elicits similar effects on lipid concentrations and major coronary events in men and women (Cholesterol Treatment Trialists' Collaboration 2015), and several meta-analyses have provided evidence to support the safe and effective use of statins in older people (Cholesterol Treatment Trialists' Collaboration 2012a; Cholesterol Treatment Trialists' Collaboration 2010; Taylor *et al*. 2013; Savarese *et al*. 2013).

Three types of adverse events have reliably been attributed to the use of statins, namely:
- Myopathy, also termed myositis (muscle pain accompanied by a ≥10-fold rise in normal levels of creatine kinase – about one case per 10,000 patient-years of treatment).
- New-onset type 2 diabetes mellitus (about one to two cases per 1,000 patient-years).
- Haemorrhagic stroke (about one case per 10,000 patient-years).

The evidence shows that these are typically rare and marginal events, which are generally accepted to be outweighed by the beneficial effect of statins on CVD risk (Collins 2016). Claims that statin therapy can have other side effects (including cancer, Parkinson's disease, rheumatoid arthritis, and dementia, among others) or beneficial effects on non-cardiovascular events (such as respiratory conditions, cognitive impairment and cancer) have largely been refuted (Cholesterol Treatment Trialists' Collaboration 2012b; Ebrahim & Taylor 2014; Smeeth *et al*. 2009; Collins 2016). As discussed previously (see **Box 4**), there is also evidence to suggest that many of the side effects commonly reported in routine clinical practice, including the most frequently reported myalgia (generalised muscle pain without raised creatine kinase levels), are not a direct consequence of statin therapy (Taylor *et al*. 2013; Finegold *et al*. 2014; Moriarty *et al*. 2014; Moriarty *et al*. 2015).

The fact that statins are used as a preventative treatment in individuals who may not have any clinical signs or symptoms does not affect the underlying evidence that supports their use. The studies that informed the change in NICE guidance provided high-quality evidence of the worthwhile benefits (including in low-risk patients) and the low rate of adverse side effects.

Although there is no convincing evidence of greater harms associated with statins in low-risk patients (Cholesterol Treatment Trialists' Collaboration 2012a), there may be a legitimate concern that there is a reduced need for preventing CVD if the base rate of CVD risk is low. Indeed, by lowering the threshold for treatment, the proportion of patients who could receive treatment and will not directly benefit from it increases. This needs to be balanced against the significant numbers of low-risk patients who will benefit from treatment.

Whereas overestimation of treatment benefits (see also the example of HRT in **Chapter 3**) may result in unnecessary use of drugs, inaccurate claims about side effects may result in treatments not being offered to patients, or patients not taking them due to unwarranted safety concerns (i.e. under-medication). In the case of statins, this is of particular concern for individuals at high-risk of CVD, who may be deterred from using a potentially life-saving treatment. Recent evidence from Australia has shown that lipid-lowering therapies were taken in 2011–2012 by less than half of patients at a high five-year risk of CVD, less than two-thirds of patients aged 45–74 years of age who had prior CVD, and less than half of patients aged above 75 years who had prior CVD (Banks *et al*. 2016).

5.4.2 Vaccination

The use of vaccines in healthy individuals to protect against future infections has some similarities with the use of statins to protect against heart disease in so much as it is a preventative treatment, the use of which is still disputed by some groups. As for statins, the use and availability of vaccines should be determined in light of evidence on efficacy and on side effects. There is a long history of successes in the use of vaccines, which has for example seen the eradication of smallpox and poliomyelitis. The World Health Organization has provided stringent guidelines on the testing of vaccines so that there can be a high level of confidence in relation to both efficacy and safety (see Academy of Medical Sciences 2005 report 'Safer Medicines'). There are however issues that are specific to the use of vaccines, as described in **Box 16** (for a discussion of the ethical implications of vaccination, see the Nuffield Council on Bioethics report 'Public health: ethical issues' 2007).

5.4.3 Medicines to treat ADHD, depression and bipolar affective disorders

There is more empirical evidence on the efficacy of stimulant medication in treating ADHD than almost any other disorder in child psychiatry (see Barkley 2000; Banaschewski *et al*. 2006; National Institute for Health and Care Excellence 2008 and 2013). There can be no doubt that stimulant medication is effective because it has been systematically compared with non-pharmacological interventions (Jensen & the MTA Group 2002) and its use has been studied in a wide range of ages spanning childhood to adult life. Nevertheless, there have been concerns that the use of stimulants to treat ADHD is medicalising a social problem (Visser & Jehan 2009). In addition, for many years concerns have been expressed in the United States about the heavy reliance on stimulants, with particular concerns about their use in preschool children (Diller 1998).[10] It is clear that there are biological changes associated with ADHD and that it is not an exclusively social problem. As far as the UK is concerned, there would be general agreement that for severe cases of ADHD, stimulants are well-warranted and have a substantial empirical research base. However, as recommended by NICE, particularly with milder cases, it is usually best to start with behavioural approaches. Recently, questions have been posed about ADHD with an onset in adult life (Moffitt *et al*. 2015). Unlike ADHD that begins in childhood, it is much less likely to have associated neuropsychological abnormalities, and twin analyses have indicated that it seems to be genetically separate from ADHD with an onset in childhood (Agnew-Blais *et al*. 2016). It remains to be determined what implication, if any, this has for the use of medication.

Somewhat similar concerns could be expressed about the use of antidepressants to treat depression (Khan & Brown 2015). Once more, there is a wealth of evidence on efficacy and side effects. As far as children are concerned, it is striking that tricyclic antidepressants are not effective in children with depression despite the fact that they are in adults (Hazell *et al*. 2002). In contrast, SSRIs are effective. On the other hand, it is notable that there is a high placebo response (Khan & Brown 2015). There is also a paucity of data on the differences among SSRIs with respect to efficacy. About half of the studies of approved antidepressants have failed to show any significant difference between medication and placebo (Zimmerman 2016). A particular concern is that strict inclusion and exclusion criteria in drug trials of antidepressants have meant that there are major problems in generalisability of findings. That was largely overcome in the Star*D trial (Trivedi *et al*. 2006) where there was a specific focus on the sample that was more representative of patients treated in routine clinical practice. A limitation of this, and other effectiveness studies, is that conclusions about efficacy are not possible because there was no placebo control group (Zimmerman 2016). Because the antidepressant drugs involve a substantial placebo effect, it could be argued that drugs are being used because they are less time consuming than CBT, which may be equally effective.

Much less is known about the use of medication in the treatment of bipolar affective disorders in childhood (see Leibenluft & Dickstein 2015). There is also concern about the serious metabolic side effects of second generation antipsychotic medications (SGAs). Less is known about the use of lithium to prevent mania in children as compared with the more substantial evidence in adult life (Correll *et al*. 2009).

5.4.4 The use of antimicrobials

The one class of drugs where there is no doubt whatsoever that there has been damaging overuse is the use of antimicrobials to treat mild viral infections, which do not respond in any case to antibiotics. Such misuse of antimicrobials

[10] In this report we do not deal further with medicalisation (the process by which human conditions and problems come to be defined and treated as medical conditions and thus become the subject of medical study, diagnosis, prevention, or treatment) or concerns about over-medication.

is one of the factors contributing to the emergence of antimicrobial resistance, an area the Academy continues to monitor and work on with partner organisations to drive a coordinated international approach to address this issue (Academy of Medical Sciences 2013d). The recent O'Neill review of antimicrobial resistance highlights the scale of the problem and outlines a list of necessary actions to tackle this rising threat (Review on Antimicrobial Resistance 2016).

As mentioned above, there is a range of other important issues pertaining to the over or underuse of medicines, which will be dealt with in more detail by the oversight group report.

Box 16. Specific considerations for the use of vaccines

Vaccination involves the administration of a pathogen (an organism that causes an infectious disease) that has been modified so that it does not cause the disease itself. In response to the vaccine, the immune system produces antibodies as though the body has been infected with the disease. Should the vaccinated individual subsequently come into contact with the infectious disease, their immune system will rapidly recognise it and produce the antibodies necessary to fight it. When a large proportion of a population has been vaccinated against a disease, herd immunity emerges. Herd immunity provides indirect protection against the disease to those who have not been immunised by preventing the spread of the infection (as most individuals will be immune to it).

Vaccination is a preventative strategy that has special considerations. First, there is a tension between the need for herd immunity and the risks to individuals. Indeed, it could be argued that, at an individual level, a person could be better off not being vaccinated because of the rare risk of side effects; however, vaccination programmes are only effective if the majority of the population is immunised and therefore it is in the wider public good to ensure herd immunity (in-depth analysis of the ethical implications of vaccination is beyond the scope of this study; for further discussion of these issues, see the Nuffield Council on Bioethics report 'Public health: ethical issues' 2007). Second, there have been alleged risks from combining vaccines – as with the MMR vaccine and more recently administering flu and shingles vaccinations concomitantly. However, as far as we are aware, there is no evidence that these proposed negative interactions actually operate. Third, many vaccines are based on attenuated live viruses or bacteria. The securing of safety requires an understanding of the mechanism of attenuation and its genetic stability (Academy of Medical Sciences 2005). Fourth, pandemic infections may involve genetic mutations that are not linked to seasonal infections. This was noted in the study of neuraminidase inhibitors, which are not vaccines, but this issue also applies to vaccines (Academy of Medical Sciences & Wellcome Trust 2015). Finally, unlike statins, vaccines may vary in efficacy in different age groups.

6. Evaluation of research findings when considering their application in clinical practice

The majority of this report is concerned with the methods for assessing the safety, efficacy and effectiveness of medicines – the main task that we were given to deal with. However, it is equally important to consider factors that contribute to the value of research evidence in the clinical decision making process. Four main steps are involved.

6.1 Step 1: Quality of evidence on efficacy and harms

To assess the crucial issues of efficacy and harm of the medication being considered it is essential that the methodologies employed allow causality to be legitimately inferred. Both require methodologies that can provide a rigorous test of the *causal* inference. For the reasons discussed in some detail in **Chapter 3**, RCTs provide the main feasible means available for this purpose with respect to benefits, but pharmacovigilance may also be needed to detect uncommon harms. As discussed in the section dealing with natural experiments in **Chapter 4**, the regression discontinuity design (RDD) has been shown to be mathematically comparable (see Shadish, Cook & Campbell 2002). However, for this to work in the way intended the rules have to be strictly followed without any exception and this is rarely feasible. Moreover, the RDD has less power than the RCT – mainly because of colinearity between the assignment and treatment variables. Accordingly, with respect to the causal inference, the main need is to consider whether the requirements for a valid RCT are met: Has randomisation been fully implemented? Has there been effective blinding throughout, including during data analysis as

well as measurement of outcomes? Have precautions been taken that treatment and control groups have not been subjected to any systematic differences after randomisation that affect outcomes? Have outcomes been accurately measured and recorded?

Pragmatic trials in ordinary clinical settings, despite their huge advantages in studying applications in such settings, may occasionally involve researchers being under pressure to include particular patients in a non-randomised fashion (see Chapter 3). Also, it has been found that drop-out rates may be high in pragmatic trials. Nevertheless, all RCTs are likely to involve some drop-out (attrition). If the reasons for drop-out differ between the experimental group and the control (or comparison) group, bias will be introduced. An appropriate way of dealing with this issue is to analyse by 'intention to treat' (see Chapter 2). This effectively deals with avoidance of bias, but only at the cost of being limited to an overall average causal effect. Angrist *et al.* (1996) and Imbens and Angrist (1994) have shown that the use of an instrumental variable (one associated with the outcome but operating only through the treatment) may enable realistic estimates of 'local' causal effects to be assessed. That is, it measures the effect of the medication (or other interventions) in the participants who actually receive it, provided that certain assumptions (which are often difficult to test) are made.

On the other hand, it is well known that a substantial proportion of patients fail to take the medicines that have been prescribed (see Chapter 2) and it is, therefore, necessary to ask about the effects of the treatment in those individuals who actually receive it. Problems here include the uncertainty of the methods available to monitor the adherence to treatment, and the possibility that adherence to a medication may differ in clinical practice from that observed in research studies. In Nelson, Fox and Zeanah's (2014) RCT of the possible benefits of foster care in the treatment of children who had been in very deprived institutions in Romania, they give a very clear account of why an 'intention to treat' analysis was necessary to avoid bias and why it was also informative to look at the effects of current living conditions at age eight years, long after the RCT had been completed. In these circumstances, this might have still allowed bias to creep in but what was special about their design was the fact that whether children did or did not remain in foster care or in an institutional environment was decided by the authorities and there was no choice by the children themselves (or their families). The value of this is that it avoids the downside of an 'intention to treat' approach, which would not take into account whether children did or did not continue in their placement. Their findings showed that there was even a stronger beneficial effect of foster care on IQ than that evident in the original 'intention to treat' analysis. It would be rare for this to be possible when examining therapeutic medication but we mention it because of the potential value of overcoming the methodological challenges.

Throughout the whole of medical sciences there is added strength if multiple research strategies can be brought together in order to determine whether they support the same hypothesis. In connection with the evaluation of RCT findings for application in clinical practice, use can be made of a randomised *withdrawal* design (Newcorn *et al*. 2016). The question being tackled is whether the withdrawal of the medication results in a progressive loss of benefits. The design is particularly useful when a medication is being used to treat a long-term disorder.

In Chapter 3 we concluded that RCTs constituted the only feasible means of testing for both efficacy and common harmful effects (observational studies were particularly useful for picking up rare or delayed harms). Here, we emphasise that for applications in medical practice we need to ask, not only *which* method was used, but also how *rigorous* were the detailed steps taken to deal with possible biases. In other words, the first step has to be determining the robustness of the findings on efficacy and harms. This will need to include attention to biological plausibility, effect size, and evidence of causality (Hill 1965; Rawlins 2008). As per Recommendation 3 discussed at the end of Chapter 3 we advise that all those evaluating evidence as used in medical practice should pay particular attention to factors that are likely to affect the validity and applicability of the results, including:

- Biological plausibility – are the findings based on sound biological principles?
- Generalisability – do the results extend to the treatment populations of interest?
- Effect size – is the size of the treatment effect large enough to be reliably detected in the study design that was undertaken and/or is the sample size large enough to detect a clinically important treatment effect if it exists?
- Causality – do the results reliably demonstrate a causal link between the treatment and the observed effect or do they merely suggest a correlation or association?

Decision-makers should use their judgement as part of the critical appraisal of the evidence, to ascertain whether the evidence they are presented with is 'fit for purpose'. This judgement is central to the evaluation of the benefits and harms of medicines.

Researchers and others should be aware that these factors will be influenced by determinants such as bias, confounding, moderating variables, choice of comparator and endpoints, participant attrition and adherence to treatments, and the

'placebo' and 'nocebo' effects, which should therefore be carefully considered in the study design. A range of methodologies and analytical approaches (including Bayesian thinking) should be given due consideration, as should the investigation of outcomes that are of particular importance to patients.

6.2 Step 2: Combining RCT data from multiple studies

The second step requires putting together multiple RCTs by means of a meta-analysis which can provide a more precise estimate of the *size* of a treatment effect and the *size* of the harms as they apply across the broad population constituted by the various subpopulations from the studies included (Cumming 2014). It may be desirable when considering multiple alternative medications or multiple dosages, to combine everything as in multi-arm RCTs (see **Chapter 3**). A conventional meta-analysis cannot do that, but there are now statistical techniques available to bring together multiple treatments in a single meta-analysis (Leucht *et al*. 2013). In **Chapter 2**, we also noted the value of using Bayesian causal networks and mixed treatment comparison meta-analyses for this purpose. As with all other studies, application to healthcare practice needs to consider the *quality* of the evidence, and not just the method used. As Uher and McGuffin (2010) noted, the meta-analysis reported by Risch *et al*. (2009) was heavily skewed by the inclusions of very large studies using weak measures. Murray (2014) commenting on the Matheson, Shepherd and Carr (2014) study, also noted that some meta-analyses (through their choice of which studies to include) were clearly biased by the researcher's own views. Once more, the basic point is that, in considering the application of findings to clinical practice, it is essential to examine the *quality* of the evidence.

6.3 Step 3: Balancing benefits and harms

The third step requires a balancing of the benefits and the harms. This is necessary with respect to all studies but it is particularly critical in the situation in which the medications, shown through RCTs to bring the greatest benefits in terms of symptom reduction, also carry a particularly high risk of serious harms. This can be illustrated by considering the use of haloperidol in treating aggression in children with autism (Cambell, Cohen & Small 1982; Campbell *et al*. 1983). Of all the drugs tested, it showed the clearest effects with respect to symptom reduction. However, the drug is particularly likely to lead to tardive dyskinesia (abnormal movements). During the 1980s, this was seen to constitute an adverse risk-benefit ratio and most clinicians ceased to prescribe the drug. A telling example from non-psychiatric medications is provided by what Rawlins (2008) describes as the 'sorry tale' of rofecoxib (an anti-inflammatory drug used to treat arthritis) in which the high level of harms led to withdrawal of the drug. The 'sorry tale' descriptor refers to both the failure to consider biological plausibility and the fact that there had been a lack of disclosure of demonstrated toxicity. A second example is provided by cimetidine, a drug used to treat peptic ulceration (Colin-Jones *et al*. 1985) for which non-randomised studies showed an increased rate of malignancies. It became clear that this did not truly reflect the balance of risks and benefits because it seems likely that selection bias was operating. Cuervo and Clarke (2003) noted that one of the key problems in balancing benefits and harms is that trialists usually know what benefits to assess but they may be unaware of what harms to look for. Thus, it took some time for the serious cardiotoxic effects of COX-2 inhibitors to be identified because they had not been systematically searched for in earlier trials.

The situation with respect to serious metabolic side effects of certain antipsychotic drugs such as olanzapine (Taylor, Paton & Kapur 2015; Arango *et al*. 2014; Fedorowicz & Fombonne 2005; Stigler *et al*. 2004) is very similar with respect to the need to balance harms and benefits but differs slightly in that despite the high rate of metabolic harms, clinicians and patients may consider that the risks of harm are worth taking in view of the risks of not treating the psychosis. However, there are now many different antipsychotic drugs (see Taylor, Paton & Kapur 2015) and it should usually be possible to find one with a level of metabolic risk that is more acceptable to both clinicians and patients.

The key point is the need to consider both harms and benefits side by side and appreciate that clinical decisions need to be based on sound research data but, at the same time, must recognise the need to take on board the preferences of the patient. This issue may be illustrated by the inclusion of low-risk groups in the use of statins. The evidence that statins were effective and that the level of harms was low was convincing, but some commentators wrongly inferred that the evidence meant that statins should *always* be used in low-risk groups. Rather, the proper procedure is that clinicians should discuss the evidence with patients, but respect the importance of allowing individual patients to make up their own minds as to whether or not statins were to be prescribed for them. In other words, in applying research findings to clinical practice, both a dispassionate account of findings should be provided *and* there should be freedom for the patient

to exercise their own choice.

6.4 Step 4: Consideration of moderating effects

The fourth step concerns the importance of moderators – namely, variables that can influence the effects of treatment. This topic was briefly discussed in Chapter 2 in relation to the topic of external validity (generalisability) but it is most important when considering how best to apply research findings to clinical practice. Patients need to know whether particular RCT findings (or others) apply to them. The need is to consider possible subgroup differences. In Chapter 2 we noted Pocock *et al*.'s (2002) warning that it is not valid to assert that effects differ between subgroups because in one there is a statistically significant effect and in another subgroup there is not. They also drew attention to the importance of pre-specifying subgroup analyses (see also Rutter & Pickles 2016 for a discussion of the importance of using pre-planned analysis guidelines such as provided by CONSORT). Simmons, Nelson and Simonsohn (2011) showed, through both simulations and actual experiments, that undisclosed flexibility in data collection and analysis led to a high likelihood of false positive findings. There are circumstances in which pre-planned analyses cannot be undertaken because new groups have to be compared but particular care is needed to avoid data dredging (by clearly specifying why new analyses are necessary and how they are to be undertaken).

An historical example illustrates how, at their best, subgroup analyses can be clinically informative. Rapoport *et al*. (1978 and 1980) showed that far from stimulant medication having a paradoxical effect in children with ADHD, there were the same beneficial effects in improving attention in all individuals with and without ADHD. On the other hand, the euphoria (and, therefore, the addictive) effects were found only in adults and not children. Curiously, however, we still do not understand the biological mechanisms that are involved. Age-related differences in effects are also evident in antidepressants. Children do not respond to tricyclic antidepressants but they do respond to SSRIs (Rutter & Pickles 2016). Again, we do not know why this is so.

Moderators have been most studied in the fields of psychology and psychiatry. However, a number of non-psychiatric medical studies have undertaken subgroup analysis to identify candidate moderator (sometimes labelled 'interactive') variables. For instance, in a well conducted study of intravitreal injections of ranibizumab as a treatment for neovascular age-related macular degeneration, it was found that the patients' baseline visual acuity score, the size of the choroidal neovascularisation area and age were the strongest predictors of a positive drug response (Kaiser *et al*. 2007). The authors were appropriately tentative in their conclusions about the meaning of these observed subgroup differences in view of the difficulty of taking account of all confounders, and the uncertainty about biological mediators.

The same reservations apply to Rubins *et al.*'s (2002) analysis of subgroups in their high-density lipoprotein intervention study. They found that, in elderly males, fasting plasma insulin levels in both those with and without diabetes were the best predictors of an effect of gemfibrozil in reducing the risk of major cardiovascular events.

6.5 Conclusions

The key point is that, in discussing research findings with patients, there should be attention to the question as to whether the RCTs that have been undertaken did or did not include individuals similar to the particular patient who is being seen in clinical practice. There should be no pre-assumptions either that any drug always has the same pattern of benefits and harms, or that subgroup differences are so pervasive that no extrapolation across subgroups is possible. Moreover, attention needs to be paid, not only to well recognised ethnic, gender or age differences, but also to less studied differences such as whether or not it matters, with respect to the disorder being studied, if there are other associated disorders beyond those being primarily targeted. In this connection it would be good to know the size of the variation in the individual effect size. The smaller it is, the more similar the subjects' responses would be. But, unlike the mean of the individual effect size, the variance cannot usually be calculated without substantial further assumptions.

Sometimes, there may be a wish to pit the importance of assessing the benefits and harms of particular therapeutic medications against the supposedly different issues of considering how the findings may be applied in ordinary clinical practice. In this chapter we have shown how *both* topics need to be based on the same research strategies. The difference, however, is that applications to ordinary clinical practice need to focus, not just on which methods have been used and on the *quality* of their use, but also on the balance of benefits versus harms and the key question of the extent

to which they apply to the individual patient being seen. Inevitably, judgement is needed as part of the critical appraisal of the evidence in order to determine whether or not it is fit for the purposes to which it is to be put.

7. Conclusions and recommendations

In this report we have focused primarily on research done since Rawlins' 2008 Harveian Oration and the Academy's report 'Identifying the environmental causes of disease' (2007). We have therefore concentrated on the methodological issues raised by the Chief Medical Officer for England in her letter to the Academy, the respondents to our call for written evidence, the stakeholders we interviewed, the ABPI, and the Academy's Council. The report concentrates on identifying areas where the way in which evidence is generated, analysed and most importantly evaluated, can be improved. Nevertheless, we welcome the considerable progress that has been made in this area in recent decades, for example in the conduct and reporting of trials, in meta-analyses, and in the implementation of standards.

7.1 The research design and conduct issues that studies should seek to address

A major concern throughout our discussions was the pervasiveness of opportunities for bias. While this is an issue that extends across the whole of medicine we concentrated particularly on the situations involved in studies of efficacy and side effects of drugs. The CONSORT, STROBE and PRISMA guidelines were explicitly devised to address these issues, including their emphasis on the protective value of pre-planned analyses or, if that was not possible, by detailing what

new analyses needed to be done and why. **We have therefore recommended that these guidelines be followed and we are concerned by the evidence that, despite their ready availability, they are not always applied in practice** (Rutter & Pickles 2016)**. We have also recommended that higher education institutions and research institutes make training on research integrity, which should include elements relating to unconscious bias in biomedical research, mandatory.** It is important that they also assess the effectiveness of such training programmes.

The key consideration with respect to the study of efficacy is the evidence on the internal validity of findings. This is ordinarily best provided by RCTs as a result of both randomisation of participants to treatment and control groups, and blinding procedures. Blinding should involve not only the trial participants but also those who are involved in administering treatments, measuring the outcomes and analysing the results. We recognise the need to deal with situations where patients cannot be blinded to the intervention they are receiving (for example when a pharmaceutical product is being compared with a psychological intervention). In these circumstances, every effort should be made to blind those rating or measuring the findings.

External validity (generalisability) concerns the extent to which findings on efficacy can be extrapolated or generalised across different populations. This is particularly an issue when RCTs have involved highly selected groups that may differ markedly from the groups for whom the therapeutic treatment is intended. This issue has widely been considered in relation to variables such as age or gender, but there has been less attention paid to the issue of the co-occurrence with other disorders. This is an important issue, not only because such co-occurrences are common, but also because many trials do not specify the extent to which these have been included or excluded. **Greater attention should be paid to a wider range of variables impacting on external validity.**

In the past, relatively little attention has been paid to the ethical issues involved in using placebos in RCTs. We do not consider that there is an ethical issue if there is no alternative treatment available. However, if alternatives are available, using placebos runs counter to the universally accepted principle that participants in trials should not be disadvantaged as a result of taking part in the trial. From a practical, clinical, point of view, if alternative treatments of known value are available, the issue is whether the new treatment is better than these and not whether the new treatment is better than no treatment at all. RCTs involving an active comparator will therefore provide the greatest relevance and insight.

Many RCTs are statistically powered for the assessment of efficacy rather than the identification of common harms. While we acknowledge that it would be impractical for RCTs to be expected to detect rare side effects, **we encourage funding bodies to continue to fund RCTs of sufficient size to detect common harms. Funding should also be available for observational studies that aim to build on initial RCT findings and examine long latency and rare adverse events that might arise once the medicine is more widely available.** Linking patients in RCTs to electronic health record systems also provides an opportunity to obtain an unbiased assessment of the long-term effects of previous exposure to treatment.

7.2 Trial designs

RCTs are sometimes discussed as if they were a homogeneous group of research designs, but they are not. For example, multi-arm RCTs have been devised to deal with both the comparison of different pharmaceutical products, and differences in dosage. They constitute an important way of going forward. There is also a range of alternative RCT designs (such as basket and umbrella trials) which deal with specific requirements of the research question under investigation. We note the value of these relatively new varieties of RCT and encourage their wider use where applicable.

Particularly in recent years, there has been enthusiasm for the use of pragmatic trials examining effectiveness of interventions when the products are used in ordinary clinical settings (such as GP surgeries). We accept the value of pragmatic trials to better understand the effectiveness of medicines in realistic, clinical settings but we note that they should be preceded by evidence on efficacy. We note also that the results from pragmatic trials may not always generalise to other populations as there is no single situation that would enable any trial (or observational study) to cover all clinical eventualities. In addition, there is the limitation of possible pressure to include non-randomised participants and evidence that the attrition rate from pragmatic trials tends to be higher than in other RCT designs.

The working group considered the extent to which the principles in the report applied to the new situations of rare diseases and emergency situations. We considered that the same principles should apply but we also note the variety of modifications that may be necessary. With respect to rare diseases, although these are individually rare they are actually common when taken as a whole. We have noted the ability of rare disease charities to provide a network (and in some

cases funding) to reach out to patients across the globe, thereby enabling trials with an adequate sample size to be conducted (for example the Cystic Fibrosis Foundation). With respect to emergency situations (such as the Ebola outbreak) the need for RCTs should be balanced against the spread and morbidity of the disease, which often makes it impractical, or even impossible, to undertake RCTs at the time with the randomisation and blinding processes they entail. This does not apply to the study of vaccines to prevent Ebola, as such studies aim to *prevent* infection as opposed to *treat* a potentially fatal disease. We acknowledge and appreciate the creativity of researchers involved with both these situations. More broadly, we encourage researchers and regulatory authorities to continue to work together to resolve the limitations of different approaches to generating and evaluating evidence.

Meta-analyses have become the accepted way of combining results from multiple studies, usually many randomised clinical trials. Their use is undoubtedly highly valuable and has transformed the assimilation of evidence, providing vital insights into the benefits and harms of medicines. The quality of meta-analyses is of course crucial. Some poor quality meta-analyses have unhelpfully combined results from RCTs and observational studies, failed to systematically evaluate sources of bias in the studies contained in the analysis, or been selective in the choice of the studies included in the meta-analysis. Processes need to be in place to prevent this occurring. The 'Cochrane Handbook for Systematic Reviews of Interventions' (Higgins *et al*. 2008) describes some precautionary measures. It should also be recognised that a single measure of efficacy may not be meaningful when there is major heterogeneity in target populations. While advocating the greater use of meta-analyses, we acknowledge the difficulties and emphasise that meta-analyses must be of high quality. Because of the importance of heterogeneity, we also urge more research into the study of moderators of efficacy.

There have been numerous proposed hierarchies of evidence that attempt to provide an idea of the strength of the evidence based on the design of the research study, but they do not always agree on the order of the hierarchy. RCTs almost invariably come at the top, or near the top, but there are differences in terms of whether meta-analyses are placed above or below RCTs. Although hierarchies may provide some utility in guiding people about the relative usefulness of different sorts of evidence for the study of efficacy, they should not be used too prescriptively nor as a substitute for judgement as part of the critical appraisal of evidence. The 'strength' of evidence will be dependent on the research question under investigation and the data utilised should be 'fit for purpose'. We note that there are rare occasions in which an RCT may not be needed. These generally concern circumstances where there is a very big effect in relation to a condition with a poor, known prognosis and a plausible biological mediator.

7.3 Evolving approaches for the study of the benefits and harms of treatments

The working group was asked to consider whether new research strategies might be available to improve the study of the benefits and harms of medicines. There is already a wide range of robust research strategies that are employed in the medical sciences; however, there are new strategies arising from outside the study of medicinal products, such as in the social sciences, which could be applied to the study of the benefits and harms of medicines. These include, for example, the more sophisticated use of propensity scores, natural experiments and instrumental variables.

Propensity scores aim to reduce the impact of confounding introduced by social selection and differ from the more usual adjustment for covariates by focusing not on the variables associated with the disease outcome but rather on the variables associated with exposure to the intervention that is being studied. A key limitation of this strategy is that it cannot account for unknown, and therefore unmeasured, variables that might affect the findings.

Natural experiments (see Chapter 4 for some examples) concern situations in which attention to naturally occurring circumstances may serve to pull apart variables that ordinarily go together. However, instances when natural experiments might arise are rare and it is impossible to randomise patients to treatment groups when they do, limiting their applicability in practice.

The use of instrumental variables, including genetic variants in Mendelian randomisation, has value in that it also attempts to minimise confounding in observational studies by using the instrumental variable as a means of controlling confounding. However, they can be brought to bear on the study of medicinal products only rarely.

These three strategies have been shown to be useful, but to our knowledge they have rarely been applied to the study of medicinal products to date.

7.4 Evaluation of research findings for clinical practice

When evaluating evidence for clinical practice, the quality of the evidence – irrespective of the method of analysis used – is crucial. In Chapter 6 we highlighted the need to consider biological plausibility; generalisability; effect size and proof of causality when assessing the validity and applicability of evidence about medicines.

We outlined four key steps:

- Assessing quality of evidence on efficacy and harms.
- Combining RCTs from multiple studies.
- Balancing benefits and harms.
- Considering moderating effects.

Judgement will be required about whether the evidence is 'fit for purpose', and its applicability to a particular patient or sector of the population.

7.5 Associated issues

7.5.1 Overuse and underuse of medication

The debate about overmedication was one of the triggers for this report. While the overestimation of treatment benefits may result in unnecessary use of drugs, inaccurate claims about side effects may result in under-medication where treatments are not being offered to patients, or patients are not taking them due to unwarranted safety concerns. We considered the issues more broadly by looking at the use of: statins and vaccination as preventative treatments; antibiotics; antidepressants; and stimulants to treat ADHD.

As far as statins to treat heart disease are concerned, we focused on the extensive evidence on benefits and harms, and on the evidence that they were useful even in low-risk populations. Of course, that does not mean that statins should invariably be used by these groups but, rather, that the possibility of their use should be jointly considered by the clinician and the patient, taking into account what is known about benefits and harms but also what the patient preference is. We noted, however, that although there was evidence of the benefit of statins in low-risk groups, the balance between benefits and harms (or even need for the drug) changed if there was extension to ever lower risk groups.

In terms of preventative treatments more widely, we consider that the use of medicinal products for an individual patient should be jointly decided by the patient and their clinician on the basis of the empirical findings of benefits and harms, taking into account the patient's personal preferences and health beliefs. What is important is that patients are presented with all the information available on the benefits and harms, which should be considered with their clinician in order to make an informed decision. A particular consideration for the use of vaccines is the need for 'herd immunity' in the population, which only occurs if the majority of the population is immunised.

We noted the evidence of antibiotic overuse to treat mild viral infections and its contribution to the emergence of antimicrobial resistance. In relation to antidepressants, NICE guidelines already indicate the desirability of using psychological interventions before moving on to antidepressants. The situation with respect to stimulants for the treatment of ADHD is somewhat more complicated in that there have been instances of overuse (in the United States) and perhaps underuse (in the United Kingdom).

The broader issues concerned with the over and underuse of medication will be considered by the oversight group.

7.5.2 Working with industry

The Academy has long recognised that working with the pharmaceutical industry has particular advantages for drug development. There have, however, been instances in the past when pharmaceutical funding has led to biased findings or biased reporting. A recent example of a pharmaceutical company not making available findings on the side effects of antidepressants indicates that the concerns are still relevant today (Sharma *et al*. 2016). We welcome the action taken by

the pharmaceutical industry to change its practice. We appreciate that the situation is improving as a result of their actions to move towards increasing transparency and of efforts by the publishing sector to encourage best practice to deter ghost writing (The World Association of Medical Editors 2005; Jacobs & Wager 2005).

During the course of investigations by the working group, we heard of a number of issues relating to working with industry that had the potential to influence the research questions and methodologies, as well as the subsequent analysis and dissemination of results. In particular, questions were raised about: the involvement of commercial partners in the study design and data analysis; data holding and data access (in terms of publication and access to the raw data for further analysis); personal payment of academic researchers by research funders; and trial registration. We believe that careful consideration of these issues when academic research is supported by industry might help to address some of the concerns about the validity of clinical research. **We have recommended that the Academy, through the oversight group, looks further into the principles governing relationships between academia and industry to address concerns about the validity of academic research funded by industry**.

A key issue today is that the diversity of medical products is broader now than ever before. In addition, the healthcare delivery world into which new products are launched is increasingly complex as decision-makers emerge at all points on the route of the product to the market, each demanding their own special knowledge base about the product before embracing it onto the next step.

7.6 Recommendations from the report

The following recommendations have emerged from our deliberations:

Recommendation 1 (Chapter 2)

We recommend that absolute risk or absolute risk difference is always presented alongside any measure of relative risk or attributable risk so that the level of risk or size of intervention effects can be properly understood. This applies to the general and scientific media, regulatory agencies, and scientists.

Recommendation 2 (Chapter 3)

We recommend that funding bodies ensure appropriate support for research in the areas of: how to deal with the difficulties created by premature termination of RCTs (in particular estimation of treatment effect); and the extent to which under-representation of certain groups in these studies really affects the generalisability of the study results. Appropriate support should also be provided for trials that are sufficient in scale and duration to achieve the pre-specified outcomes.

Recommendation 3 (Chapters 3 and 6)

We recommend that all those evaluating evidence should pay particular attention to factors that are likely to affect the validity and applicability of the results, including:

- **Biological plausibility** – are the findings based on sound biological principles?
- **Generalisability** – do the results extend to the treatment populations of interest?
- **Effect size** – is the size of the treatment effect large enough to be reliably detected in the study design that was undertaken and/or is the sample size large enough to detect a clinically important treatment effect if it exists?
- **Causality** – do the results reliably demonstrate a causal link between the treatment and the observed effect or do they merely suggest a correlation or association?

Decision-makers should use their judgement as part of the critical appraisal of the evidence, to ascertain whether the evidence they are presented with is 'fit for purpose'. This judgement is central to the assessment of the benefits and harms of medicines, as well as for the evaluation of research findings when considering their application in clinical practice.

Researchers should be aware that these factors will be influenced by determinants such as bias, confounding, moderating variables, choice of comparator and endpoints, participant attrition and adherence to treatments, and the 'placebo' and 'nocebo' effects, which should therefore be carefully considered in the study design. Alternative trial methodologies and analytical approaches (including Bayesian thinking) should be given due consideration, as should the investigation of outcomes that are of particular importance to patients.

Recommendation 4 (Chapter 4)

Electronic health records, research databanks and disease registries are valuable sources of so-called 'real world' data and we recommend that their use should be explored by researchers and regulators, and in HTA assessments. In developing an approach for access to and linkage of data in the UK, attention should be paid to approaches such as those in Scandinavia, where the use of unique personal identifiers, supportive infrastructures and appropriate governance have enabled the straightforward linkage of data, and anonymity is protected by making data available to researchers in an irreversible encrypted fashion.

Recommendation 5 (Chapter 5)

We recommend that guidelines such as CONSORT, STROBE and PRISMA, among others, be comprehensively adopted and that funding bodies require their grant awardees to adhere to these reporting guidelines. We encourage research institutions and journals to provide incentives for their use and to better enforce their adoption.

Recommendation 6 (Chapter 5)

We recommend that higher education institutions and research institutes make training on research integrity, which should include elements relating to unconscious bias in biomedical research, mandatory and that they assess the effectiveness of such training programmes.

Recommendation 7 (Chapter 5)

We support the registration of RCTs and mandatory publication of their protocols by researchers as a means to help ensure that so-called 'negative', null or inconclusive trial results become publicly available, and to enhance research reproducibility. We recommend that similar methods be adopted for observational epidemiological studies exploring the effects of treatments, where currently such an approach is lacking. We encourage journals to make registration of such observational epidemiological studies and publication of their protocols a condition of publication, as is the case for clinical trials.

Recommendation 8 (Chapter 5)

We recommend that the Academy, through the oversight group, looks further into the principles governing relationships between academia and industry to address concerns about the validity of academic research funded by industry.

7.7 Guidelines for different stakeholder groups

In addition to the recommendations outlined above, we propose the following guidelines to assist researchers in academia and industry, healthcare professionals, regulators, scientific/medical journals and publishers, and funding bodies, in the evaluation of evidence. It is by no means a comprehensive list, as the proposed guidance focuses predominantly on issues relating to the methodology of medicine evaluation, providing high-level guidance of the various issues to be aware of.

7.7.1 Guidelines for researchers in academia and industry

When designing a research study and deciding what type of study should be undertaken, researchers should carefully consider the following wide range of research issues (see **Chapter 2** for further details on each of these considerations):

Bias	Causality
Confounding	Choice of comparator
Blinding	Participant attrition
Internal/external validity and relevance	Adherence to treatments
Moderating variables	The 'placebo' and 'nocebo' effects
Absolute risk, relative risk, attributable risk and number needed to treat	Surrogate endpoints

- After careful consideration of these issues, the researcher should make a judgement as to which elements are of most importance, reflecting on the implications their decision might have on the introduction of bias, internal/external validity, determination of causation, and so forth. The researcher's judgement on these considerations will dictate the most appropriate study type for the research question to be investigated. Alternative trial designs (such as basket or umbrella trials) and analysis approaches (for example, Bayesian thinking) should be given due consideration, as should the investigation of outcomes that are of particular importance to patients.
- Following on from this decision, measures should be put in place to minimise as far as is practically possible the limitations of the study design, conduct and analysis. These limitations and the measures to mitigate them should be clearly stated in research publications, alongside the strengths of the findings.
- Studies involving human participants should be registered and summary results made available in public registries. Study findings should be published in full and include details of the source of funding. Due consideration should also be given to access to patient-level data, which should be provided in a manner that protects patient confidentiality and ensures that the data are intelligible, assessable, reliable and usable.
- Researchers should comply with guidelines such as CONSORT, STROBE and PRISMA when publishing their research. In particular, they should present, as standard, absolute risk alongside any measure of relative risk or attributable risk, as well as complete baseline demographic and clinical characteristics of participants and non-participants, and details of attrition levels.
- Researchers at all levels should have the relevant training in methodology, statistics and research integrity (including unconscious bias) to ensure that their studies are as robust as possible.
- Academic researchers should participate in debates around the level of autonomy required from their funders in terms of the design and publication of their research to ensure that the validity of their research is not undermined.

7.7.2 Guidelines for healthcare professionals

- In assessing the applicability of evidence about medicines for clinical practice, healthcare professionals need to judge the quality of the evidence (which will require a consideration of whether the findings are biologically plausible, whether causality has been demonstrated, and whether the study was large enough to reliably detect the target effect size). Healthcare professionals should be aware that studies may (or may not) vary in how applicable (or generalisable) they are to patients that differ from those that participated in the research. Such differences, and the potential impact of other variables that can moderate the benefits or harms of a treatment, which may not be well understood, may need careful communication to patients to inform shared decision-making.
- Whenever possible, healthcare professionals should communicate absolute risk and/or absolute risk difference figures alongside any measure of relative risk or attributable risk to inform shared decision-making. We acknowledge that these figures are not always easy to find, although we are optimistic that this will become easier in the future.

7.7.3 Guidelines for regulators

- There are a number of evolving approaches which may prove beneficial in place of traditional methods for evidence generation in certain situations (see **Chapter 4**). Regulators should provide clarity on how alternative methodologies and new data sources could be most usefully used in regulatory submissions.
- Regulators should ensure that the true clinical outcome should be used as the primary endpoint in Phase III trials and not surrogate endpoints (unless the validity of the surrogate endpoint has already been rigorously established). Data on surrogate endpoints that demonstrates biological activity can be presented alongside the clinical outcome data to provide evidence that the intervention is active and targeting the pathways of interest.

- Regulators should encourage the study of outcomes that are of particular importance to patients in regulatory submissions. Robust qualitative evidence has the potential to provide important insights in this regard.

7.7.4 Guidelines for scientific and medical journals, and publishers

- Journals should provide incentives to researchers for the use of guidelines such as CONSORT, STROBE and PRISMA, and require their adoption. Scientific and medical journals should be familiar with these guidelines, and question studies where these have not been used.
- Publishers should encourage and support best practice in the reporting of complete baseline demographic and clinical characteristics of participants and non-participants.
- Publishers should globally co-operate with research funders, research institutions, individual researchers, and wider professional bodies to enhance the reproducibility and reliability of research.
- Journals should continue to incentivise the publication of null, inconclusive or 'negative' results to mitigate the problems associated with publication bias.
- Journals should continue to encourage good practice in order to deter ghost writing, and refuse to publish manuscripts where there is concern about the involvement of such ghost writers.

7.7.5 Guidelines for funding bodies

- Funding should be made available for:
    - Robust studies employing new methodologies.
    - RCTs of sufficient size to detect common harms.
    - Observational studies that aim to build on initial RCT findings and examine long latency and rare adverse events that might arise once the medicine is more widely available.
- Funding bodies should provide incentives to researchers for the use of guidelines such as CONSORT, STROBE and PRISMA, and better enforce their adoption.
- Funders should continue to request that applicants submit details of sample size calculation and underpinning justification for their research (e.g. of choice of effect size and outcome) in their funding applications.

Annex I. Report preparation

Working group membership

This report was prepared by a working group of the Academy of Medical Sciences. Members participated in a personal capacity, not as representatives of the organisations listed. A summary of the working group members' interests is summarised below.

Chair

Professor Sir Michael Rutter CBE FRS FBA (until November 2016) is Professor of Developmental Psychopathology at the Social Genetic and Developmental Psychiatry (SGDP) Research Centre at the Institute of Psychiatry, Psychology and Neuroscience, King's College London. His research interests include the use of natural experiments and animal models to test hypotheses about causation; the use of epidemiological longitudinal studies for the same purpose; gene-environment interplay; and studies of psychosocial risk. He founded the SGDP in 1994 and was its first honorary director. He retired from his administrative posts in 1998 but remains active in research and teaching. His textbook on child and adolescent psychiatry remains distinctive in attention to both conceptual and statistical issues, as well as the integration of science and clinical work. He is Governor of the Coram foundation and Chair of the Scientific Advisory Group for the Canadian Institute for Advanced Research programme on Child and Brain Development.

Members

Sir Alasdair Breckenridge CBE FRSE FMedSci was Chairman of the MHRA for 10 years. Previously, he was Chairman of the Committee on Safety of Medicines and Professor of Clinical Pharmacology at the University of Liverpool. He currently chairs the Advisory Scientific Committee of the Centre of Regulatory Excellence for the Government of Singapore and was a Non-Executive Director of University College London Hospitals from 2012 to 2015.

Professor Nancy Cartwright FBA is a methodologist and philosopher of the natural and social sciences specialising in issues of causal inference, objectivity and the nature of evidence in the natural and social sciences and for social policy. She currently holds joint appointments at Durham University and the University of California, San Diego and has worked previously at Stanford University and the London School of Economics and Political Science. She is a fellow of the British Academy, the American Academy of Arts and Sciences, the German National Academy of Science (Leopoldina) and the American Philosophical Society, and is a recipient of a MacArthur fellowship. She has also been President of the American Philosophical Association (Pacific Division) and of the Philosophy of Science Association. She has been awarded a number of grants for her work, including from: A. K. Knight; the Arts and Humanities Research Council; the British Academy; the European Research Council; the John Templeton Foundation; the Latsis Foundation; the London School of Economics and Political Sciences Seed Fund; the Durham University Seedcorn Award; the Spencer Foundation; and the US National Science Foundation.

Professor Dame Nicky Cullum DBE FMedSci is Professor of Nursing and Head of the School Division of Nursing, Midwifery and Social Work in the School of Health Sciences at the University of Manchester. Her research mainly focuses on the epidemiology and management of complex wounds such as leg, foot and pressure ulcers and non-healing surgical wounds. She was a founding member of the Cochrane Collaboration and has been Coordinating Editor of the Cochrane Wounds Group since 1995. A particular interest is how research evidence of relevance to clinical nursing decisions is produced and that evidence is translated into practice. She founded the Centre for Evidence-Based Nursing at the University of York in 1995. She is a member of the Royal College of Nursing and a Fellow of the American Academy of Nursing. She has a number of current grants, either as a principal or co-applicant, from the NIHR.

Mr Edward Green is a patient representative on the Great Ormond Street Hospital Members' Council (Patient and Carer Constituency), where he strives to improve the overall patient experience by listening to patients and families and holding the executive to account. He has helped drive the implementation of new systems for patients and clinicians to communicate and work collaboratively throughout the period of care across the NHS. He is also a Business Consultant at Block Solutions, a technology focused consultancy.

Professor Deborah Lawlor FMedSci is Professor of Epidemiology and Deputy Director of the MRC Integrative Epidemiology Unit at the University of Bristol. Her research is concerned with women's reproductive health, and how their pregnancy experiences and characteristics relate to their own, their offspring's, and their grandchildren's future cardiometabolic health. She has developed and applied a range of novel methods for obtaining best estimates of causal effects in population science. She is currently the Chair of the MRC Population Health Sciences Group and a member of the MRC Strategy Board, and has served on various boards and panels for the MRC, Wellcome Trust and British Heart Foundation. During the working group project, she also served on the Academy of Medical Sciences' Council. She has received a number of grants, either as a principal or co-applicant, from: Alcohol Research UK; Australian Government National Health and Medical Research Council; British Heart Foundation; Economic and Social Research Council; European Research Council; Medical Research Council; Medtronic External Research Programme; NIHR; the UK Clinical Research Collaboration; the US NIH; the US National Institute on Aging; and Wellcome Trust. She has recently received industry support for her biomarker research from Medtronic and Roche Diagnostics.

Professor Mahesh Parmar is Professor of Medical Statistics and Epidemiology, and Director of both the MRC Clinical Trials Unit and the Institute of Clinical Trials and Methodology at University College London (UCL). Until 2014 he was an Associate Director of the National Cancer Research Network since its inception in 2001, an organisation which has more than tripled the number of f patients recruited to cancer studies in England. The Unit he directs at UCL aims to deliver swifter and more effective translation of scientific research into patient benefits, particularly in infectious diseases and cancer, by carrying out challenging and innovative studies, and by developing and implementing methodological advances in study design, conduct and analysis. The Unit receives educational grants and free or reduced priced drugs for many of the trials that it runs, including from: Abbott, Amgen, Astellas, AstraZeneca, Bayer, Baxter, Boehringer Ingelheim, Bristol-Myers Squibb, Cipla, Gilead Sciences, GlaxoSmithKline, Janssen, Janssen-Cilag, Lilly, Merck, Merck Serono, Novartis, Roche, Sanofi-Aventis, Sanofi Pasteur, Tibotec, Virco, World Health Organization (WHO)/Global drug facility (GDF). Professor Parmar has received a number of grants, either as a principal or co-applicant from: Cancer Research UK; Department of Health; European Science Foundation; MRC; and NIHR.

Professor Simon Thompson FMedSci is Director of Research in Biostatistics in the Department of Public Health and Primary Care, University of Cambridge. From 2000–2011 he was Director of the MRC Biostatistics Unit in Cambridge, an internationally acclaimed research institute in medical statistics. He held previous academic appointments at the London School of Hygiene and Tropical Medicine, and as the first Professor of Medical Statistics and Epidemiology at Imperial College London. His research interests are in meta-analysis and evidence synthesis, clinical trial methodology, health economic evaluation, and cardiovascular epidemiology; he has published widely in these areas. He has collaborated on a number of major clinical trials, including all the major recent UK national trials of screening and treatment for abdominal aortic aneurysms. He has been a strong advocate for biostatistics through his research papers, didactic articles, and contributions to courses and workshops both in the UK and abroad. He has also taken on a number of responsibilities for the MRC, the Royal Statistical Society, and other international professional societies. Professor Thompson has received a number of grants, either as a principal or co-applicant, from the MRC, British Heart Foundation, European Research Council, and NIHR.

Dr Julian Treadwell is a GP based in Wiltshire and an NIHR In-Practice Fellow at the Nuffield Department of Primary Care Health Sciences. He is Vice-Chair of the Royal College of General Practitioners (RCGP) Standing Group on Overdiagnosis and a member of the editorial board of the Drug and Therapeutics Bulletin. He has an interest in evidence informed prescribing and shared decision-making. He is a member of the RCGP and of the British Medical Association.

Professor David Webb FRSE FMedSci is the Christison Professor of Therapeutics and Clinical Pharmacology at the University of Edinburgh and leads a European Society of Hypertension (ESH) Centre of Excellence in Edinburgh. His research focuses on renal and vascular aspects of hypertension, and he runs a Scottish translational medicine and therapeutics clinical PhD programme funded by the Wellcome Trust. He is a Non-Executive Director of the MHRA and Chair of the Scientific Advisory Committee for the National Institute for Biological Standards and Control (NIBSC). He is President of the British Pharmacological Society (BPS), taking over Presidency in 2016. He is also Honorary President of the European Association for Clinical Pharmacology & Therapeutics (EACPT) and Clinical Vice-President of the International Union of Basic and Clinical Pharmacology (IUPHAR). His research is supported by grants from the British Heart Foundation, Kidney Research UK, the MRC and the Wellcome Trust.

Dr Stephen Webster is Director of the Science Communication Unit at Imperial College, which is responsible for training science graduates wishing to become media professionals, policy advisors and public engagement specialists. His academic background is in zoology and the philosophy of biology. Dr Webster was for many years a school science teacher in London. For six years he was Chair of the Wellcome Trust's public engagement award scheme, the Society

Awards. He has broadcast widely for BBC radio, including dramas and several documentaries on the nature of science. Dr Webster's books include a biography of Charles Darwin and an edited collection of essays on the role of silence in science communication.

This working group was initiated in part as a response to the public debate over the risks and benefits of statins. Because of Amgen's involvement in the development of cholesterol-lowering treatments, it was decided that the participation of **Christine Fletcher** (Executive Director of Biostatistics and a Regional Head in Global Biostatistical Science at Amgen) in the working group could be viewed as a conflict of interest. Christine Fletcher therefore stepped down from the working group. She did not participate in the project after 21 December 2015 and was not involved in discussions about emerging conclusions and recommendations. The Academy is grateful for her contributions to the first two working group meetings.

To ensure the industry perspective was appropriately considered in the report, **Dr Melanie Lee CBE FMedSci** kindly provided independent input into the report.

Observers

Representatives from the MHRA and NICE were invited to join the discussions as observers to clarify factual points. They did not input into or have sight of the study's conclusions and recommendations. The observers were:

- **Dr Sarah Branch**, Deputy Director – Vigilance and Risk Management of Medicines Division, MHRA
- **Dr Ian Hudson**, Chief Executive, MHRA
- **Professor Andrew Stevens**, Chairman of a NICE Appraisal Committee, and Professor of Public Health and former Head of Department and Division of Primary Care, Public and Occupational Health at the University of Birmingham

Secretariat

Dr Claire Cope (**Lead Secretariat**), Policy Manager, Academy of Medical Sciences
Dr Rachel Quinn, Director, Medical Science Policy, Academy of Medical Sciences
Dr Rachel Brown, Policy Officer, Academy of Medical Sciences (October 2015-October 2017)
Dr Katharine Fox, Policy Officer, Academy of Medical Sciences (January-June 2017)

We are grateful for the contributions of the Academy's policy interns: **Elizabeth Gothard** (October–December 2015; Wellcome Trust-funded PhD student), **Hannah Julienne** (January–March 2016; Wellcome Trust-funded PhD student), and **Thomas Hall** (April–June 2016; MRC-funded PhD student).

We are also grateful to **Sandra Woodhouse**, Personal Assistant to Sir Michael Rutter CBE FRS FBA, for her support.

Review group membership

The report was reviewed by a group on behalf of the Academy's Council. Reviewers were asked to consider whether the report met the terms of reference and whether the evidence and arguments presented in the report were sound and supported the conclusions. Reviewers were not asked to endorse the report or its findings. Review group members were:
Professor Christopher Day FMedSci (Chair), Vice-Chancellor and President, Newcastle University and Vice-President of the Academy of Medical Sciences
Professor Deborah Ashby OBE FMedSci, Professor of Medical Statistics and Clinical Trials, Co-Director of Clinical Trials Unit, Imperial College London
Professor Carol Dezateux CBE FMedSci, Professor of Paediatric Epidemiology, Institute of Child Health, University College London
Dr David Gillen, Vice President, Vertex Pharmaceuticals
Professor Jon Nicholl CBE FMedSci, Dean of the School of Health and Related Research, University of Sheffield

Annex II. Consultation and evidence gathering

Initial stakeholder meetings

To help frame the remit of the study and the work of the wider workstream, stakeholder meetings were held between the Chair and the individuals or organisations listed below from May to September 2015.

Meeting on 26 May 2015 with:
Professor Sir Rory Collins FRS FMedSci, British Heart Foundation Professor of Medicine and Epidemiology and Head of the Nuffield Department of Population Health, University of Oxford
Sir Michael Rawlins FMedSci, Chair, MHRA and author of the 2008 Harveian Oration[11]

First meeting on 17 June 2015 with:
Professor Dame Sue Bailey, Chair, Academy of Medical Royal Colleges
Dr Dean Marshall, General Practitioners Committee, British Medical Association
Dr Imran Rafi, Chair of Clinical Innovation and Research, Royal College of General Practitioners
Sir John Savill FRS FRSE FMedSci, Chief Executive, Medical Research Council
Dr Iain Simpson, President of the British Cardiovascular Society and the Joint Specialty Committee for Cardiology, Royal College of Physicians
Professor Tom Walley CBE, Director of the Health Technology Assessment (HTA) Programme, NIHR
Professor Peter Weissberg FMedSci, Medical Director, British Heart Foundation

Second meeting on 17 June 2015 with:
Professor David Haslam CBE, Chair, National Institute for Health and Care Excellence
Professor Gillian Leng CBE, Deputy Chief Executive, National Institute for Health and Care Excellence
Dr June Raine CBE, Director of Vigilance and Risk Management of Medicines, Medicines and Healthcare products Regulatory Agency

Meeting on 21 September 2015 with:
ABPI's Innovation Board

Call for evidence

The Academy issued an open call for evidence to inform the study and the work of the overall workstream. The call for evidence was open for eight weeks from 27 July to 21 September 2015. Those who submitted written evidence are listed below. The responses to the call for written evidence (for which permission to publish was received) are available on the Academy's website.[12]

Individuals
Professor Sheila Bird OBE FRSE, Programme Leader, MRC Biostatistics Unit, University of Cambridge and Visiting Professor, Department of Mathematics and Statistics, Strathclyde University
Professor Dame Nicky Cullum DBE FMedsci, Head of the Division of Nursing, Midwifery and Social Work and Professor of Nursing, University of Manchester
Mr Wayne Douglas
Dr Brian Edwards, Consultant Pharmacovigilance and Drug Safety and Chair, Pharmaceutical Human Factors and Ergonomic Group
Dr Ben Goldacre, Senior Clinical Research Fellow, Centre for Evidence-Based Medicine, Nuffield Department of Primary

[11] Rawlins M (2008). *De Testimonio: on the evidence for decisions about the use of therapeutic interventions*. Royal College of Physicians. Available at: http://www.amcp.org/WorkArea/DownloadAsset.aspx?id=12451
[12] http://www.acmedsci.ac.uk/policy/policy-projects/how-can-we-all-best-use-evidence/evidence-repository/

Care Health Sciences, University of Oxford
Professor Sir John Grimley Evans FMedSci, Professor Emeritus, University of Oxford
Mr Barry Haslam
Mr Gordon Jarvis
Dr Huw Llewelyn, Honorary Departmental Fellow, Aberystwyth University
Professor Pascal McKeown, Director of Medical Education, Queen's University Belfast
Mr Nigel Mellor
Mr John Perrott
Professor Lawrence Phillips, Emeritus Professor of Decision Sciences, London School of Economics and Political Science
Professor Ian Roberts, Professor of Epidemiology and Public Health and Co-director of the Clinical Trials Unit, London School of Hygiene & Tropical Medicine
Professor Ian Russell, Former Senior Research Leader, Health and Care Research Wales
Sir Richard Thompson, Former President, Royal College of Physicians
Mr David Tovey, Editor in Chief, The Cochrane Collaboration
Dr Andrew Tressider, Sessional GP, Somerset NHS
Professor John Urquhart, Senior Medical Advisor, AARDEX-MWV Healthcare and Adjunct Professor; University of California, San Francisco

Organisations
Association of the British Pharmaceutical Industry
Breast Cancer Now
British Heart Foundation
British Pharmacological Society
Cancer Research UK
Centre for Evidence-Based Medicine, University of Oxford
Faculty of Pharmaceutical Medicine
Health Technology Assessment International
Healthcare Improvement Scotland – Evidence Directorate
London School of Hygiene and Tropical Medicine
Medical Research Council
Medicines and Healthcare products Regulatory Agency
MRC Network of Hubs for Trials Methodology Research and Northern Ireland Network for Trials Methodology Research
National Institute for Health and Care Excellence
Nuffield Department of Population Health, University of Oxford
Royal College of Physicians of London
University College London – Department of Science and Technology Studies
University College London – Evaluating Evidence in Medicine project team
University College London – School of Pharmacy
Wellcome Trust

Additional evidence gathering

The following individuals provided oral evidence to the working group on 5 November 2015:
Tracey Brown, Director of Sense about Science
Claire Murray, Joint Head of Operations at the Patient Information Forum

Notes of the oral evidence sessions are available on the Academy's website.[13]

We are very grateful to all those who have contributed information to the study, including anyone that we have inadvertently omitted from this list.

[13] http://www.acmedsci.ac.uk/policy/policy-projects/how-can-we-all-best-use-evidence/evidence-repository/

Annex III. Hierarchies of evidence

In an attempt to guide decision-making about the appropriate use of therapeutic interventions, 'hierarchies of evidence' have emerged which aim to provide an idea of the strength (i.e. robustness) of the underlying evidence. Details of a subset of such hierarchies are provided below.

Hierarchies of evidence are regarded as pragmatic 'rules of thumb', but their use has been the subject of debate. As put forward by Rawlins in his 2008 Harveian Oration, the idea that sources of evidence can be placed in such hierarchies is '*illusory*', as all sources of evidence on a particular topic (e.g. experimental and observational studies) should be considered and addressed to inform decision-making (Rawlins 2008). Boaz and Ashby (2003) have further argued that an evaluation of evidence should consider more than the methodological quality, including the extent to which the research is relevant and fit for purpose.

The development of the first hierarchy of evidence and subsequent permutations

The term 'hierarchy of evidence' was first used in a 1979 report by the Canadian Task Force on the Periodic Health Examination to grade the effectiveness of an intervention according to the quality of evidence obtained (Canadian Task Force on the Periodic Health Examination 1979). The report's purpose was to develop recommendations on the periodic health examination and base those recommendations on evidence in the medical literature. The effectiveness of intervention was graded according to the quality of the evidence obtained. The Canadian Task Force updated its report in 1984, 1986 and 1987 (Canadian Task Force on the Periodic Health Examination 1984, 1986, 1988). The 1987 hierarchy had the following ranking:

- I: Evidence obtained from at least one properly randomised controlled trial.
- II-1: Evidence obtained from well-designed controlled trials without randomisation.
- II-2: Evidence obtained from well-designed cohort or case-control analytic studies, preferably from more than one centre or research group.
- II-3: Evidence obtained from comparisons between times or places with or without the intervention. Dramatic results in uncontrolled experiments (such as the results of treatment with penicillin in the 1940s) could also be included in this category.
- III: Opinions of respected authorities, based on clinical experience, descriptive studies or reports of expert committees.

Various hierarchies were then developed, largely building on the Canadian Task Force ranking. Examples of such hierarchies include (see references for further detail):

- Sackett (1989)
- United States Department of Health and Human Services, Agency for Health Care Policy and Research (1993)
- Guyatt *et al.* (1995)
- US Preventive Services Task Force (1989, 1996, 2008, 2012)
- Harbour & Miller (2001)
- Centre for Evidence-based Medicine, University of Oxford (2009)

The majority of these hierarchies of evidence place either RCTs, or systematic reviews or meta-analyses of RCTs, at the top as providing the most robust forms of evidence. These are then typically followed by cohort studies, case-control studies, case series, studies with no controls, and expert opinion.

Since the early introduction of hierarchies by the Canadian Task Force and Sackett, it has become recognised that the type and level of evidence may need to be modified according to the research question under investigation (e.g. those that investigate treatments, prognoses, diagnoses, or economic/decision-making) (Burns, Rohrich & Chung 2011). The American Society of Plastic Surgeons, for example, has introduced different rating scales for therapeutic studies, diagnostic studies, and prognostic/risk studies (American Society of Plastic Surgeons 2016).

Current uses of hierarchies of evidence

Two hierarchies of evidence are currently commonly used in the UK: the NICE grading scheme and the Grading of Recommendations Assessment, Development and Evaluation (GRADE) methodology (National Clinical Guideline Centre 2010; GRADE working group 2015).

NICE grading scheme

The NICE grading scheme is used to develop guidelines on conditions and diseases and was originally adapted from the US Agency for Healthcare Policy and Research (now the Agency for Healthcare Research and Quality) Classification (National Clinical Guideline Centre 2010). Recommendations are then additionally graded A to C on the basis of the level of associated evidence, or noted as a GPP recommendation (i.e. recommended good practice based on the clinical experience of the Guideline Development Group). The NICE grading scheme is as follows:

- Ia: Evidence obtained from systematic reviews or a meta-analysis of randomised controlled trials.
- Ib: Evidence obtained from at least one randomised controlled trial.
- IIa: Evidence obtained from at least one well-designed controlled study without randomisation.
- IIb: Evidence obtained from at least one other well-designed quasi-experimental study.
- III: Evidence obtained from well-designed non-experimental descriptive studies, such as comparative studies, correlation studies.
- IV: Evidence obtained from expert committee reports or opinions and/or clinical experiences of respected authorities.

The GRADE methodology

The GRADE system was developed to help reduce confusion and inconsistencies in the other grading systems used by various organisations. It is used widely, with more than 50 organisations worldwide endorsing or using it, including the WHO, Centers for Disease Control and Prevention, Cochrane Collaboration, BMJ, National Institute for Health and Care Excellence, and Scottish Intercollegiate Guidelines Network (GRADE working group 2015). The GRADE system allows four initial levels of quality (Petrisor & Bhandari 2007; Kerwin *et al*. 2012; UpToDate 2015):

- Randomised trial = high quality
- Quasi-randomised trial = moderate quality
- Observational study = low quality
- Any other evidence = very low quality

The levels of quality can however be adjusted. They may be decreased if there is:

- A serious (-1) or very serious (-2) limitation of study quality
- Important inconsistency (-1)
- Some (-1) or major (-2) uncertainty about directness
- Imprecise or sparse data (-1)
- High probability of reporting bias (-1)

Alternatively, the grade may be increased if:

- There is strong evidence of association – significant relative risk of >2 (<0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1)
- There is ery strong evidence of association – significant relative risk of >5 (<0.2) based on direct evidence with no major threats to validity (+2)
- There is evidence of a dose response gradient (+1)
- All plausible confounders would have reduced the effect (+1)

The working group views hierarchies of evidence as useful guides, which should not be used too prescriptively nor as a substitute for judgement as part of the critical appraisal of evidence. The ‘strength’ of evidence will be dependent on the research question under investigation and the data utilised should be ‘fit for purpose’.

Annex IV. Alternative trial designs

Adaptive trials

These are trials where accumulating results are used to modify the trial during its execution, without undermining the trial or the integrity of the final results. These have sometimes been called SMARTs (sequential multiple assignment randomised trials; see Ng & Weisz 2016). For example, in a multi-armed trial testing many doses of a new therapy, interim data may help to choose the most promising dose to continue with within the trial. These studies have the potential to be more efficient, more powerful, or more informative than traditional RCTs. However, their validity relies on the protocol being firmly established before the trial begins in order to control the rate of false positive results (Berry 2011; Gallo *et al*. 2006).

Baskerville trials

Baskerville trials are designed to evaluate therapeutic preferences of patients. In these trials, patients are randomly assigned to groups, which are each allocated a unique sequence of therapy options (e.g. a sequence of different asthma inhalers). At each appointment, the patient, in discussion with their clinician, chooses whether to continue the therapy or move to the next option. Both the patient and the clinician are blind to the sequence of therapeutic options. This design provides an ethical and quantitative way to study which, if any, therapy is preferred by patients, and the results can be compared to the original safety and efficacy findings from conventional RCTs (Baskerville *et al*. 1984).

Basket (bucket) trials

These are studies that test the efficacy of therapies that are targeted to the underlying make-up of the disease, typically the genetic make-up (for example, a trial could include patients with a common mutation arising in lung, colon, and bone cancer). In the field of oncology, the genetic findings (or the biomarker findings) from the tumour are often used to help define the genetic make-up (Redig & Jänne 2015; Lopez-Chavez *et al*. 2015). Currently, these trial designs are most developed and used in the field of oncology. In some situations, these designs may have more limited applicability, for example where tissue samples (such as brain tissue) cannot easily be examined during life and where the causes of disease are multi-factorial and not driven largely by genetic mutations (Academy of Medical Sciences 2007).

Cluster trials

In these trials, the unit of randomisation is at a group (cluster) level rather than at the individual participant level, for example hospitals (as opposed to individual participants) might be randomised to a new intervention or control group. They are useful in situations when it is impracticable to direct interventions towards individuals, where the intervention would be given at a cluster level, or where there is likely to be a psychological 'contamination' effect across individuals. They retain many of the strengths of individual-level RCTs; however, the clustered nature of the study needs to be taken into account when calculating the sample size and they generally have less statistical power than a trial of the same number of participants who are randomised directly rather than in clusters. Analyses of cluster RCTs must take account of the clustering (Bland & Kerry 1997; Brugha *et al*. 2016; Campbell *et al*. 2004).

Crossover trials

Randomised crossover studies are a powerful means of assessing responses to treatment, particularly in chronic stable conditions (see Williams *et al*. 2015). The trial is a repeated measures design in which each patient is randomly assigned to a sequence of at least two treatments (e.g. current best treatment or placebo, and novel treatment). In most crossover trials, all subjects receive the same number of treatments and participate for the same number of periods, and each subject receives all treatments. Crossover studies have two major advantages over a parallel group design (where two treatments are compared side by side): 1) the influence of confounding covariates is reduced because each patient serves as their own control; and 2) they are statistically efficient, so require fewer subjects than non-crossover designs. However, they have three potential limitations: 1) the order in which treatments are administered may affect the outcome; 2) 'carry-over' effects between treatments can confound the estimate of the effect of the subsequent treatment; and 3) many treatments/diseases do not lend themselves to the crossover design as, by the time of the

subsequent treatment, the patient might have become ineligible to receive it.

Equivalence trials

Equivalence trials are used to test whether a novel intervention is no better or worse than a standard intervention (Greene *et al*. 2008).[14] They are often used to test novel interventions that are expected to have equal efficacy but also have other benefits, such as ease of use, fewer side effects, or reduced cost. A typical example is testing whether a cheaper generic drug is as effective as the original patented drug (Lesaffre 2008). To test for equivalence, a clearly delineated zone within which the intervention is considered equivalent should be defined before the onset of the trial. However, there is no common method for defining this. Because equivalence trials do not have a placebo controlled group, it is also imperative that a consistently effective standard treatment exists and that it is administered correctly to avoid erroneous conclusions about the efficacy of the novel intervention (Greene *et al*. 2008).

Factorial trials

Factorial trials test two or more interventions and whether those interventions interact (Montgomery, Peters & Little 2003). The simplest example is a 2 x 2 design where participants are given either no intervention, one of the interventions on their own, or both interventions (**Table 1**). The advantage of factorial design over traditional parallel group design – where two treatments are compared side by side (e.g. a new treatment vs. standard of care) – is that fewer participants are needed to achieve the same statistical power. Factorial trials can also be used to investigate whether interventions interact, although larger sample sizes are required to detect interactions. Factorial trials are an efficient method of investigating multiple interventions simultaneously, as long as the interventions do not interact significantly. If interventions do interact, then the analysis of these trials becomes more complicated and trials can be incorrectly analysed and interpreted.

Table 1. Example of a 2 x 2 factorial trial design (Higgins & Green 2011)

	Behavioural intervention (B)	Standard care (C)
Aspirin (A)	A and B	A and C
Placebo (P)	P and B	P and C

For example, dietary conditions were assessed in a randomised double-blind prevention trial with a factorial design to determine whether folic acid taken during pregnancy prevented neural tube defects (NTDs) in newborn babies. Four dietary conditions were compared: no supplementation; folic acid supplements; other vitamins; and a combination of folic acid and other vitamins. From 1,195 completed pregnancies, there were 27 with a NTD, six in the two groups with folic acid and 21 in the other two groups. In this study, folic acid prevented about 75% of NTDs (Medical Research Council Vitamin Study Research Group, 1991).

Ring trials

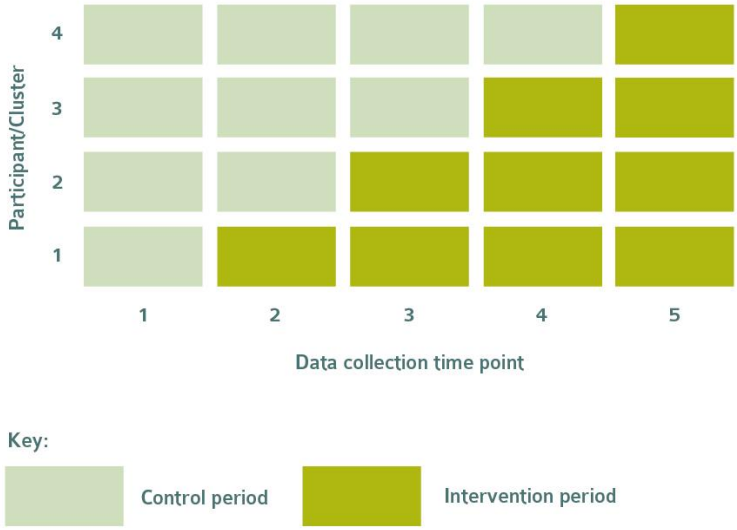
A ring trial, which is often used for vaccinations, is one when a 'ring' of individuals (at high risk) connected to a known case receives an intervention. In the context of immunisation, individuals at high risk of contracting a disease due to their social or geographical proximity to a known case are vaccinated. This approach was central to the eradication of smallpox in the 1970s (Fenner *et al*. 1988). More recently, it was adopted in vaccination trials for Ebola (Ebola Ça Suffit Ring Vaccination Trial Consortium 2015).

Stepped wedge trials

These are characterised by a randomised, sequential roll-out of an intervention over time (Hussey & Hughes 2007) and can be considered a special form of cluster randomised trial (see above). The roll-out is typically unidirectional, from control to intervention, meaning that all participants ultimately end up receiving the intervention. The name originates from a schematic illustration of the trial design (**Figure 3**) (The Gambia Hepatitis Study Group 1987).

[14] Equivalence trials should not be confused (as is commonly the case) with non-inferiority trials, which only test whether an intervention is no worse than a standard intervention.

Figure 3. A schematic illustration of the stepped wedge trial design



Umbrella trials

These constitute a variant of basket trials, and an example is trials in which patients with a given cancer type are assigned to one of a number of treatment arms based on the molecular make-up of their tumour. This could lead to more efficient evaluation of treatments and increased targeted personalised treatments (Catenacci 2015; Kaplan *et al*. 2013).

Annex V. References

Academy of Medical Sciences (2005). *Safer medicines*. http://www.acmedsci.ac.uk/viewFile/publicationDownloads/SaferMed.pdf

Academy of Medical Sciences (2007). *Identifying the environmental causes of disease: how should we decide what to believe and when to take action?* https://www.acmedsci.ac.uk/viewFile/publicationDownloads/119615475058.pdf

Academy of Medical Sciences (2013a). *Realising the potential of stratified medicine.* http://www.acmedsci.ac.uk/viewFile/51e915f9f09fb.pdf

Academy of Medical Sciences (2013b). *Clinical trial transparency*. http://www.acmedsci.ac.uk/policy/policy-projects/clinical-trials-and-data-disclosure/

Academy of Medical Sciences (2013c). *Academy of Medical Sciences' FORUM.* https://acmedsci.ac.uk/about/objectives/linking-academia-industry-NHS/forum

Academy of Medical Sciences (2013d). *Antimicrobial resistance*. http://www.acmedsci.ac.uk/policy/policy-projects/antimicrobial-resistance/

Academy of Medical Sciences (2015a). *How can we all best use evidence?* http://www.acmedsci.ac.uk/policy/policy-projects/how-can-we-all-best-use-evidence/

Academy of Medical Sciences (2015b). *Submission to the 'Science in Emergencies: Lessons from Ebola' inquiry*. http://www.acmedsci.ac.uk/policy/policy-projects/academy-responds-to-science-in-emergencies-lessons-from-ebola-inquiry/

Academy of Medical Sciences (2015c). *Reproducibility and reliability of biomedical research*. http://www.acmedsci.ac.uk/policy/policy-projects/reproducibility-and-reliability-of-biomedical-research/

Academy of Medical Sciences (2016a). *Evidence repository.* http://www.acmedsci.ac.uk/policy/policy-projects/how-can-we-all-best-use-evidence/evidence-repository/

Academy of Medical Sciences (2016b). *Perspectives on 'Evaluating evidence in health' – a workshop report.* www.acmedsci.ac.uk/evidence/evaluating-evidence-in-health

Academy of Medical Sciences & Wellcome Trust (2015). *Use of neuraminidase inhibitors in influenza*. http://www.acmedsci.ac.uk/viewFile/561595082cd83.pdf

Adebamowo C, *et al*. (2014). *Randomised controlled trials for Ebola: practical and ethical issues*. Lancet **384(9952)**, 1423-1424.

Agnew-Blais J, *et al*. (2016). *Evaluation of the persistence, remission and emergence of ADHD in young adulthood*. Journal of the American Medical Association Psychiatry **73**, 1-8.

AllTrials (2014). AllTrials calls for all past and present clinical trials to be registered and their full methods and summary results reported. http://www.alltrials.net/find-out-more/all-trials/

Altman DG (2005). *Adjustment for covariate imbalance*. In Armitage P & Colton T eds. (2005). Encyclopedia of biostatistics. 2nd ed. John Wiley, Chichester.

Altman DG, *et al*. (2001). *The revised CONSORT statement for reporting randomized trials: explanation and elaboration*. Annals of Internal Medicine **134**, 663–694.

American Society of Plastic Surgeons (2016). *Evidence-based Clinical Practice Guidelines.* http://www.plasticsurgery.org/for-medical-professionals/quality-and-health-policy/evidence-based-medicine-guidelines/description-and-development-of-evidence-based-practice-guidelines.html

Angrist JD, Imbens GW & Rubin DB (1996). *Identification of causal effects using instrumental variables.* Journal of the American Statistical Association **91(434)**, 444-455.

Arango C, *et al.* (2014). *Second-generation antipsychotic use in children and adolescents: a six-month prospective cohort study in drug-naïve patients.* Journal of the American Academy of Child & Adolescent Psychiatry, **53**, 1179-1190.

Arthritis Research UK (2016). *Clinical studies*. http://www.arthritisresearchuk.org/research/research-funding-and-policy/our-clinical-study-groups/our-current-clinical-trials.aspx

Ashby D (2006). *Bayesian statistics in medicine: a 25 year review*. Statistics in Medicine **25**, 3589-3631.

Association of the British Pharmaceutical Industry (2013). *Clinical Trial Transparency: Technical standards for data sharing for old, current and future clinical trials.* http://www.abpi.org.uk/industry-info/Documents/Clinical Trial Transparency.pdf

ASTERIX project (2016). *Welcome to the ASTERIX project*. http://www.asterix-fp7.eu/

Atladóttir H, *et al.* (2007). *Time trends in reported diagnoses of childhood neuropsychiatric disorders: a Danish cohort study*. Archives of Paediatric and Adolescent Medicine **161**, 193-198.

Banaschewski T, *et al*. (2006). *Long-acting medications for the hyperkinetic disorders: a systematic review and European treatment guideline*. European Child & Adolescent Psychiatry **15**, 476-495.

Banks E, *et al*. (2016). *Absolute risk of cardiovascular disease events, and blood pressure- and lipid-lowering therapy in Australia*. The Medical Journal of Australia **204(8)**, 320e1-8.

Barkley RA (2000). *Taking charge of ADHD*. Guilford Press, New York.

Barter P, *et al*. (2007). *Effects of torcetrapib in patients at high risk for coronary events*. The New England Journal of Medicine **357**, 2109–2122.

Baskerville JC, *et al*. (1984). *Clinical trials designed to evaluate therapeutic preferences*. Statistics in Medicine **3(1)**, 45-55.

Bassler D, *et al.* (2010). *Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis*. Journal of the American Medical Association **303**, 1180-1187.

Begley CG & Ellis LM (2012). *Drug development: raise standards for preclinical cancer research*. Nature **483**, 531–533.

Berry DA (2011). *Adaptive clinical trials: The promise and the caution*. Journal of Clinical Oncology **29(6)**, 606-609.

Bland JM & Kerry SM (1997). *Trials randomised in clusters*. BMJ **315(7108)**, 600.

Boaz A & Ashby D (2003). *Fit for purpose? Assessing research quality for evidence based policy and practice*. ESRC UK Centre for Evidence-Based Policy and Practice, London. https://www.kcl.ac.uk/sspp/departments/politicaleconomy/research/cep/pubs/papers/paper-11.aspx

Boers M, *et al*. (1998). *The OMERACT filter for outcome measures in rheumatology*. Journal of Rheumatology **25**, 198-199.

British Heart Foundation (2016). *Clinical Study Grants.* https://www.bhf.org.uk/research/information-for-researchers/what-we-fund/clinical-study

Brugha TS, *et al*. (2016). *Can community midwives prevent antenatal depression? An external pilot study to test the feasibility of a cluster randomized controlled universal prevention trial.* Psychological Medicine **46(2)**, 345-56.

Burns PB, Rohrich RJ & Chung KC (2011). *The Levels of Evidence and their role in Evidence-Based Medicine*. Journal of Plastic & Reconstructive Surgery **128**, 305-310.

Byrne D & Ragin CC (2009). *The SAGE Handbook of Case-Based Methods*. Sage, London.

Cahan S & Cohen N (1989). *Age versus schooling effects on intelligence development*. Child Development **60**, 1239-1249.

Campbell M, Cohen IL & Small AM (1982). *Drugs in aggressive behavior*. Journal of the American Academy of Child Psychiatry **21**, 107-117.

Campbell M, *et al*. (1983). *Long-term therapeutic efficacy and drug-related abnormal movements: a prospective study of haloperidol in autistic children*. Psychopharmacology Bulletin **19**, 80-83.

Canadian Task Force on the Periodic Health Examination (1979). *The periodic health examination*. Canadian Medical Association Journal **121,** 1193-254.

Canadian Task Force on the Periodic Health Examination (1984). *Task Force Report: The periodic health examination: 2. 1984 update*. Canadian Medical Association Journal **130**, 1278–1285.

Canadian Task Force on the Periodic Health Examination (1986). *Task Force Report: The periodic health examination: 2. 1985 update.* Canadian Medical Association Journal **134**, 724–727.

Canadian Task Force on the Periodic Health Examination (1988). *Task Force Report: The periodic health examination*: 2. 1987 update. Canadian Medical Association Journal **138**, 618–626.

Cancer Research UK (2016). *Funding for researchers*. http://www.cancerresearchuk.org/funding-for-researchers

Cape J, *et al*. (2016). *Group cognitive behavioural treatment for insomnia in primary care: a randomized controlled trial.* Psychological Medicine **46**, 1015-1025.

Cartwright N (2007). *Are RCTs the gold standard?* BioSocieties **2**, 11-20.

Catenacci, DVT (2015). *Next-generation clinical trials: Novel strategies to address the challenge of tumor molecular heterogeneity*. Molecular Oncology **9(5)**, 967-996.

Centre for Evidence-Based Medicine (2009). *Oxford Centre for Evidence-based Medicine – Levels of Evidence* http://www.cebm.net/oxford-centre-evidence-based-medicine-levels-evidence-march-2009/

Cholesterol Treatment Trialists' (CTT) Collaboration (2005). *Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins*. Lancet **366**, 1267-1278.

Cholesterol Treatment Trialists' (CTT) Collaboration (2010). *Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials*. Lancet **376(9753)**, 1670-1681.

Cholesterol Treatment Trialists' Collaboration (2012a). *The effects of lowering LDL cholesterol with statin therapy in people at low risk of vascular disease: meta-analysis of individual data from 27 randomised trials.* Lancet **380(9841)**, 581-590.

Cholesterol Treatment Trialists' Collaboration (2012b). *Lack of effect of lowering LDL cholesterol on cancer: meta-analysis of individual data from 175,000 people in 27 randomised trials of statin therapy.* PLoS One **7(1)**, e29849.

Cholesterol Treatment Trialists' Collaboration (2015). *Efficacy and safety of LDL-lowering therapy among men and women: meta-analysis of individual data from 174,000 participants in 27 randomised trials*. Lancet **385(9976)**, 1397-1405.
Clinical Practice Research Datalink (2016). *Welcome to the Clinical Practice Research Datalink.* https://www.cprd.com/intro.asp

ClinicalTrials.gov (2012).*ClinicalTrials: A service of the U.S. National Institutes of Health.* https://clinicaltrials.gov/

Cochrane Community Archive (2015). *Systematic reviews & meta-analyses*. https://community-archive.cochrane.org/about-us/evidence-based-health-care/webliography/books/sysrev

Colin-Jones DG, *et al*. (1985). *Postmarketing surveillance of the safety of cimetidine: mortality during second, third, and fourth years of follow up*. BMJ (Clinical Research Edition) **291(6502)**, 1084-1088.

Collins R, *et al*. (2016). *Interpretation of the evidence for the efficacy and safety of statin therapy*. Lancet **388 (10059)**, 2532–2561.

Committee on Publication Ethics (2011). *A short guide to ethical editing for new editors.* http://publicationethics.org/files/short%20guide%20to%20ethical%20editing%20for%20new%20editors.pdf

CONSORT (2010). *The CONSORT Statement.* http://www.consort-statement.org/

Corbyn Z (2012). *Promising new era dawns for cystic fibrosis treatment*. Lancet **379(9825)**, 1475–1476.

Correll CU, *et al*. (2009). *Cardiometabolic risk of second-generation antipsychotic medications during first-time use in children and adolescents*. Journal of the American Medical Association **302**, 1765-1773.

Cuervo LG & Clarke M (2003). *Balancing benefits and harms in health care*. BMJ **327(7406)**, 65-66.

Cumming G (2014). *The new statistics: Why and how*. Psychological Science **25**, 7-29.

Davey Smith G, *et al*. (2008). *Clustered environments and randomized genes: a fundamental distinction between conventional and genetic epidemiology*. PLoS Medicine **4**, e352.

Deane BR & Sivarajah J (2017). *Clinical trial transparency update: an assessment of the disclosure of results of company-sponsored trials associated with new medicines approved in Europe in 2013*. Current Medical Research and Opinion **33(3)**, 473-478.

De Angelis C, *et al*. (2004). *Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors*. The New England Journal of Medicine **351**, 1250-1251.

Department for Education (2010). *The impact of Sure Start Local Programmes on seven year olds and their families. Research Report DFE-RR220*. http://www.ness.bbk.ac.uk/impact/documents/DFE-RR220.pdf

Diller LH (1998). *Running on Ritalin: A physician reflects on children, society and performance in a pill*. Bantam Books, New York.

Dumville JC, Torgerson DJ & Hewitt CE (2006). *Reporting attrition in randomised controlled trials*. BMJ **332**, 969-971.

Ebola Ça Suffit Ring Vaccination Trial Consortium (2015). *The ring vaccination trial: a novel cluster randomised controlled trial design to evaluate vaccine efficacy and effectiveness during outbreaks, with special reference to Ebola*. BMJ **351**, h3740.

Ebrahim S & Taylor FC (2014). *Statins for the primary prevention of cardiovascular disease*. BMJ **348**, g280.

Ebrahim S & Davey Smith G (2015). *N-of-1 approach to determine when adverse effects are caused by statins*. BMJ **351**, h5281.

Egger M, Schneider M & Davey Smith G (1998). *Spurious precision? Meta-analysis of observational studies*. BMJ **316(7125)**, 140-144.

EU Clinical Trials Register (2012). *News update.* https://www.clinicaltrialsregister.eu/

European Commission (2014). *Clinical trials – Regulation EU No 536/2014.* http://ec.europa.eu/health/human-use/clinical-trials/regulation/index_en.htm

European Medicines Agency (2014). *European Medicines Agency policy on publication of clinical data for medicinal products for human use*. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf

European Medicines Agency (2016). *Scientific guidelines: paediatrics.* http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/general/general_content_000404.jsp&mid=WC0b01ac0580926186

Fedorowicz VJ & Fombonne E (2005). *Metabolic side effects of atypical antipsychotics in children: a literature review*. Journal of Psychopharmacology **19**, 533-550.

Felson DT, *et al*. (1993). *The American College of Rheumatology preliminary core set of disease activity measures for rheumatoid arthritis clinical trials.* Arthritis and Rheumatology **36**, 729-740.

Fenner F, *et al*. (1988). *Smallpox and its eradication*. World Health Organization, Geneva.

Finegold JA, *et al*. (2014). *What proportion of symptomatic side effects in patients taking statins are genuinely caused by the drug? Systematic review of randomized placebo-controlled trials to aid individual patient choice*. European Journal of Preventative Cardiology **21(4)**, 464-474.

Fleming TR & DeMets DL (1996). *Surrogate end points in clinical trials: are we being misled?* Annals of Internal Medicine **125**, 605-613.

Frank JD (1961). *Persuasion and Healing: A Comparative Study of Psychotherapy*. John Hopkins University Press, Baltimore.

Friedman LM & Schron EB (2015). *Methodology of intervention trials in individuals*. In Detels R, *et al*. eds. (2015) Oxford Textbook of Global Public Health. 6th ed. Oxford University Press, Oxford.

Furu K, *et al*. (2010). *The Nordic countries as a cohort for pharmacoepidemiological research*. Basic & Clinical Pharmacology & Toxicology **106**, 86–94.

Gagne JJ, *et al*. (2014). *Innovative research methods for studying treatments for rare diseases: methodological review*. BMJ **349**, g6802.

Gallo P, *et al*. (2006). *Adaptive designs in clinical drug development – An Executive Summary of the PhRMA Working Group*. Journal of Biopharmaceutical Statistics **16(3)**, 275-283.

Garbe E, *et al*. (2013). *High-dimensional versus conventional propensity scores in a comparative effectiveness study of coxibs and reduced upper gastrointestinal complications*. European Journal of Clinical Pharmacology **69(3)**, 549-557.

Glasziou P, *et al*. (2007). *When are randomised trials unnecessary? Picking signal from noise*. BMJ **334**, 349-351.

Glymour MM, Tchetgen TEJ & Robins JM (2012). *Credible Mendelian randomization studies: approaches for evaluating the instrumental variable assumptions.* American Journal of Epidemiology **175(4)**, 332-339.

Goss CH, *et al.* (2002). *The cystic fibrosis therapeutics development network (CF TDN): a paradigm of a clinical trials network for genetic and orphan diseases.* Advanced Drug Delivery Review **54(11)**, 1505-1528.

GRADE working group (2015). *Organizations that have endorsed or that are using GRADE.* http://www.gradeworkinggroup.org/

Greene C, *et al*. (2008). *Noninferiority and equivalence designs: issues and implications for mental health research.* Journal of Traumatic Stress **21(5)**, 433–439.

Greenland S (2000). *An introduction to instrumental variables for epidemiologists*. International Journal of Epidemiology **29**, 722-729.

Gupta S, *et al*. (2011). *A framework for applying unfamiliar trial designs in studies of rare diseases*. Journal of Clinical Epidemiology **64(10)**, 1085-1094.

Guyatt GH, *et al.* (1995). *Users' guides to the medical literature. IX. A method for grading health care recommendations.* Journal of the American Medical Association **274**, 1800-1804.

Haerskjold A, *et al.* (2015). *The Danish National Prescription Registry in studies of a biological pharmaceutical: palivizumab – validation against two external data sources*. Clinical Epidemiology **7**, 305–312.

Hampton JR (2015). *Therapeutic fashion and publication bias: the case of anti-arrhythmic drugs in heart attack*. Journal of the Royal Society of Medicine **108(10)**, 418–420.

Harbour R & Miller J, for the Scottish Intercollegiate Guidelines Network Grading Review Group (2001). *A new system for grading recommendations in evidence based guidelines.* BMJ **323**, 334-336.

Harden A, Weston R & Oakley A (1999). *A review of the effectiveness and appropriateness of peer-delivered health promotion interventions for young people*. EPPI-Centre, University of London, London. http://www.healthevidence.org/view-article.aspx?a=review-effectiveness-appropriateness-peer-delivered-health-promotion-19064

Hazell P, *et al*. (2002). *Tricyclic drugs for depression in children and adolescents.* Cochrane Database of Systematic Reviews **2**, CD002317.

Health Research Authority (2015). *Transparency, registration and publication.* http://www.hra.nhs.uk/resources/during-and-after-your-study/transparency-registration-and-publication/

Heckman J (1979). *Sample selection as a specification error*. Econometrica **47(1)**, 153–161.

Hernan MA, *et al*. (2008). *Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease*. Epidemiology **19(6)**, 766-779.

Higgins JPT & Green S, eds (2008). *Cochrane Handbook for Systematic Reviews of Interventions.* Wiley, Chichester.

Higgins JPT & Green S (2011a). *Introduction to sources of bias in clinical trials*. Cochrane Handbook for Systematic Reviews of Interventions. http://handbook.cochrane.org/chapter_8/8_4_introduction_to_sources_of_bias_in_clinical_trials.htm

Higgins JPT & Green S, eds (2011b). *Cochrane Handbook for Systematic Reviews of Interventions.* Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. www.handbook.cochrane.org

Higgins JP, *et al.* (2011). *The Cochrane Collaboration's tool for assessing risk of bias in randomised trials*. BMJ **343**, d5928.

Hill AB (1965). *The environment and disease: association or causation?* Proceedings of the Royal Society of Medicine **58**, 295-300.

Hill AB (2015). *The environment and disease: association or causation?* Journal of the Royal Society of Medicine **108**, 32-37.

Hingorani A & Humphries S (2005). *Nature's randomised trials*. Lancet **366(9501)**, 1906-1908.

Honda H, Shimizu Y & Rutter M (2005). *No effect of MMR withdrawal on the incidence of autism: a total population study*. Journal of Child Psychology and Psychiatry **46**, 572-579.

Hsia J, *et al*. (2006). *Conjugated equine estrogens and coronary heart disease: the Women's Health Initiative.* Archives of Internal Medicine **166(3)**, 357-365.

Hussey MA & Hughes JP (2007). *Design and analysis of stepped wedge cluster randomized trials*. Contemporary Clinical Trials **28(2)**, 182–191.

Imbens GW & Angrist JD (1994). *Identification and estimation of local average treatment effects*. Econometrica **62(2)**, 467-475.

IMI PROTECT (2009). *About PROTECT*. http://www.imi-protect.eu/about.shtml

Innovative Medicines Initiative (2013). *GETREAL: Incorporating real-life clinical data into drug development.* http://www.imi.europa.eu/content/getreal

Innovative Medicines Initiative (2014). *WEB-RADR: Recognising Adverse Drug Reactions.* http://www.imi.europa.eu/content/web-radr

InSPiRe (2016). *Innovative methodology for small populations research*. http://www2.warwick.ac.uk/fac/med/research/hscience/stats/currentprojects/inspire/

Integrated DEsign and AnaLysis of small population group trials (2016). *Integrated DEsign and AnaLysis of clinical trials in SPG*. http://www.ideal.rwth-aachen.de/

International Committee of Medical Journal Editors (2015*). Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals.* http://www.icmje.org/icmje-recommendations.pdf

Ioannidis JPA (2005). *Why most published research findings are false*. Plos Medicine **8(4)**, 40-47.

Ioannidis JPA, *et al*. (2014). *Increasing value and reducing waste in research design, conduct, and analysis.* Lancet **383(9912)**, 166–175.

ISD Scotland (2010). *Use of the NSS National Safe Haven*. http://www.isdscotland.org/Products-and-Services/EDRIS/Use-of-the-National-Safe-Haven/

ISD Scotland (2016). *CHI number*. http://www.ndc.scot.nhs.uk/Dictionary-A-Z/Definitions/index.asp?ID=128&Title=CHI%20Number

ISIS-2 Collaborative Group (1988). *Randomised trial of intravenous streptokinase, oral aspirin, both, or neither among 17,187 cases of suspected acute myocardial infarction: ISIS-2*. Lancet **332(8607)**, 349-60.

Jacobs A & Wager E (2005). *European Medical Writers Association (EMWA) guidelines on the role of medical writers in developing peer-reviewed publications*. Current Medical Research and Opinion **21(2)**, 317–321.

Jadad AR & Enkin MW (2007). *Randomized Controlled Trials: Questions, Answers, and Musings*. Blackwell Publishing, Oxford. http://onlinelibrary.wiley.com/doi/10.1002/9780470691922.fmatter/pdf

Jensen PS & MTA Group (2002). *Treatments: the case of the MTA study*. In Sandberg S ed. Hyperactivity and attention disorders of childhood. 2nd ed. Cambridge University Press, Cambridge.

Joy TR, *et al*. (2014). *N-of-1 (single-patient) trials for statin-related myalgia*. Annals of Internal Medicine **160(5)**, 301-310.

Kaiser PK, *et al*. (2007). *Ranibizumab for predominantly classic neovascular age-related macular degeneration: subgroup analysis of first-year ANCHOR results*. American Journal of Ophthalmology **144**, 850-857.

Kaplan R, *et al*. (2013). *Evaluating many treatments and biomarkers in oncology: a new design*. Journal of Clinical Oncology **31(36)**, 4562-4568.

Kennedy SB, *et al*. (2016). *Implementation of an Ebola virus disease vaccine clinical trial during the Ebola epidemic in Liberia: Design, procedures, and challenges*. Clinical Trials **13(1)**, 49-56.

Kerwin A, *et al*. (2012). *The Eastern Association of the Surgery of Trauma approach to practice management guideline development using Grading of Recommendations, Assessment, Development, and Evaluation (GRADE) methodology.* Journal of Trauma and Acute Care Surgery **73**, S283-S287.

Khan A & Brown W (2015). *Antidepressants versus placebo in major depression: an overview*. World Psychiatry **14(3)**, 294–300.

Korn EL & Baumrind S (1998). *Clinician Preferences and the Estimation of Causal Treatment Differences*. Statistical Science **13**, 209-235.

Kraemer HC (2015). *Evaluating interventions*. In Thapar A, *et al*. eds. Rutter's Child and Adolescent Psychiatry. 6th ed. Wiley Blackwell, Oxford.

Lachmann HJ, *et al*. (2009). *Use of canakinumab in the cryopyrin-associated periodic syndrome*. The New England Journal of Medicine **360(23)**, 2416–2425.

Laupacis A, *et al*. (1988). *An assessment of clinically useful measures of the consequences of treatment*. The New England Journal of Medicine **318**, 1728-1736.

Lawlor DA, *et al*. (2004a). *Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence?* Lancet **363**, 1724-1727.

Lawlor DA, *et al*. (2004b). *Observational versus randomised trial evidence.* Lancet **364**, 754-755.

Lawlor DA & Davey Smith G (2006). *Cardiovascular risk and hormone replacement therapy.* Current Opinion in Obstetrics and Gynaecology **18**, 658-665.

Lawlor DA, *et al*. (2008). *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology*. Statistics in Medicine **27**, 1133-1163.

Leibenluft E & Dickstein DP (2015). *Bipolar disorder in childhood*. In Thapar A, *et al*. eds. Rutter's Child and Adolescent Psychiatry. 6th edition. Wiley Blackwell, Oxford.

Lesaffre E (2008). *Superiority, equivalence, and non-inferiority trials.* Bulletin of the NYU Hospital for Joint Diseases **66(2)**, 150-154.

Leucht S, *et al*. (2013). *Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis*. Lancet **382**, 951-962.

Li T, *et al*. (2011). *Network meta-analysis-highly attractive but more methodological research is needed*. BioMed Central Medicine **9**, 79.

Lindsay DS (2015). *Replication in Psychological Science*. Psychological Science **26(12)**, 1827-1832.

Lone F (2003). Epidemiology. *The Epidemiologist's Dream: Denmark*. Science **301(5630)**, 163.

Lopez-Chavez A, *et al*. (2015). *Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology Phase II basket trial*. Journal of Clinical Oncology **33(9)**, 1000-1007.

MacKinnon DP & Fairchild AJ (2009). *Current Directions in Mediation Analysis*. Current Directions in Psychological Science **18(1)**, 16–20.

Mamdani M, *et al*. (2005). *Reader's guide to critical appraisal of cohort studies: 2. Assessing potential for confounding*. BMJ **330**, 960–962.

Manson J, *et al*. (2003). *Estrogen plus Progestin and the Risk of Coronary Heart Disease.* The New England Journal of Medicine **349**, 523-534.

Matheson SL, Shepherd AM & Carr VJ (2014). *How much do we know about schizophrenia and how well do we know it? Evidence from the Schizophrenia Library*. Psychological Medicine **44**, 3387-3405.

McBride WG (1961). *Thalidomide and congenital abnormalities*. Lancet **2(7216)**, 1358-1359

McClellan M, McNeil BJ & Newhouse JP (1994). *Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality? Analysis using instrumental variables*. Journal of the American Medical Association **272**, 859-866.

McKee M, *et al*. (1999). *Interpreting the evidence: choosing between randomised and non-randomised studies*. BMJ **319**, 312–315.

McNeil J, *et al*. (2010). *The Value of Patient-Centred Registries in Phase IV Drug Surveillance*. Pharmaceutical Medicine **24(5)**, 281-288.

Medical Research Council (2008). *Developing and evaluating complex interventions: new guidance*. https://www.mrc.ac.uk/documents/pdf/complex-interventions-guidance/

Medical Research Council (2016). *Translation and clinical trials*. http://www.mrc.ac.uk/funding/science-areas/translation/translation-and-clinical-trials/

Medical Research Council Vitamin Study Research Group (1991). *Prevention of neural tube defects: results of the Medical Research Council Vitamin Study*. Lancet **338**, 131-137.

Mellin GW & Katzenstein M (1962). *The saga of thalidomide. Neuropathy to embryopathy, with case reports of congenital anomalies*. The New England Journal of Medicine **267**, 1184-1193.

Mills EJ, Thorlund K & Ioannidis J (2013). *Demystifying trial networks and network meta-analysis.* BMJ **346**, f2914.

Moffitt TE, *et al*. (2015). *Is adult ADHD a childhood-onset neurodevelopmental disorder? Evidence from a four-decade longitudinal cohort study*. American Journal of Psychiatry **172**, 967-977.

Montgomery A, Peters T & Little P (2003). *Design, analysis and presentation of factorial randomised controlled trials.* BioMed Central Medical Research Methodology **3**, 26-30.

Moriarty PM, *et al.* (2014). *ODYSSEY ALTERNATIVE: Efficacy And Safety of the Proprotein Convertase Subtilisin/kexin Type 9 Monoclonal Antibody, Alirocumab, versus Ezetimibe, in Patients With Statin Intolerance as Defined by a Placebo Run-in and Statin Rechallenge Arm*. http://circ.ahajournals.org/content/130/23/2105.full#T116

Moriarty PM, *et al*. (2015). *Efficacy and safety of alirocumab vs ezetimibe in statin-intolerant patients, with a statin rechallenge arm: The ODYSSEY ALTERNATIVE randomized trial*. Journal of Clinical Lipidology **9**, 758-769.

Murray R (2014). *On collecting meta-analyses of schizophrenia and postage stamps.* Psychological Medicine **44**, 3407-3408.

National Clinical Guideline Centre UK (2010). *Hierarchy of evidence and grading of recommendations.* In National Clinical Guideline Centre UK (2010) Chronic Obstructive Pulmonary Disease: Management of Chronic Obstructive Pulmonary Disease in Adults in Primary and Secondary Care

(NICE Clinical Guidelines, No. 101). Royal College of Physicians, London. http://www.ncbi.nlm.nih.gov/books/NBK65032/

National Institute for Health and Care Excellence (2008). A*ttention deficit hyperactivity disorder: diagnosis and management of ADHD in children, young people and adults. CG72*. https://www.nice.org.uk/Guidance/cg72

National Institute for Health and Care Excellence (2013). *Attention deficit hyperactivity disorder. QS39*. https://www.nice.org.uk/guidance/qs39

National Institute for Health and Care Excellence (2014a). *Cardiovascular disease: risk assessment and reduction, including lipid modification. CG181*. https://www.nice.org.uk/guidance/cg181

National Institute for Health and Care Excellence (2014b). *Concerns about the latest NICE draft guidance on statins*. https://www.nice.org.uk/Media/Default/News/NICE-statin-letter.pdf

National Institute for Health and Care Excellence (2016). *Glossary.* https://www.nice.org.uk/glossary?letter=n

National Institute for Health Research (2015). *Funding and support.* http://www.nihr.ac.uk/funding-and-support/

National Institutes of Health (2015). *Liberia-U.S. clinical research partnership opens trial to test Ebola treatments*. https://www.nih.gov/news-events/news-releases/liberia-us-clinical-research-partnership-opens-trial-test-ebola-treatments

Nelson CA, Fox NA & Zeanah CH (2014). *Romania's Abandoned Children: Deprivation, Brain Development, and the Struggle for Recovery*. Harvard University Press, Cambridge, Massachusetts.

Newcorn J, *et al*. (2016). *Extended-release guanfacine hydrochloride in 6–17 year olds with ADHD: a randomised-withdrawal maintenance of efficacy study*. Journal of Child Psychology and Psychiatry **57(6)**, 717-728.

Newhouse JP & McClellan M (1998). *Econometrics in outcomes research: the use of instrumental variables*. Annual Review of Public Health **19**, 17-34.

Ng MY & Weisz JR (2016). *Annual Research Review: Building a science of personalized intervention for youth mental health.* Journal of Child Psychology and Psychiatry **57**, 216-236.

NHS Choices (2015*). Clinical trials – What happens in a clinical trial?* http://www.nhs.uk/Conditions/Clinical-trials/Pages/Introduction.aspx#what-happens

Nicklin J, *et al*. (2010). *Collaboration with patients in the design of patient-reported outcome measures: capturing the experience of fatigue in rheumatoid arthritis*. Arthritis Care Research (Hoboken) **62(11)**, 1552-1558.

Normand SL, *et al*. (2005). *Readers guide to critical appraisal of cohort studies: 3. Analytical strategies to reduce confounding*. BMJ **330**, 1021–1023.

Nosek, BA *et al*. (2015). *Promoting an open research culture*. Science **348(6242)**, 1422-1425.

Nuffield Council on Bioethics (2007). *Public health: ethical issue*s. Cambridge Publishers Ltd, Cambridge. http://nuffieldbioethics.org/wp-content/uploads/2014/07/Public-health-ethical-issues.pdf

OMERACT (2015). *The OMERACT Handbook.* http://www.omeract.org/pdf/OMERACT_Handbook.pdf

Open Science Collaboration (2015). *Estimating the reproducibility of psychological science*. Science **349(6251)**, aac4716.

Open Science Framework (2011). *Open Science Framework: a scholarly commons to connect the entire research cycle*. https://osf.io/

Open Science Network (2013). *Reproducibility Project: Cancer Biology.* https://osf.io/e81xl/

Parmar MKB, Carpenter J & Sydes MR (2014). *More multiarm randomised trials of superiority are needed.* Lancet **384(9940)**, 283-284.

Patil P, Peng RD & Leek JT (2016). *What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science*. Perspectives on Psychological Science **11(4)**, 539-544.
Pearson H (2016). *The Life Project: The extraordinary story of our ordinary lives*. Penguin Random House, New York.

Petrisor BA & Bhandari M (2007). *The hierarchy of evidence: Levels and grades of recommendation.* Indian Journal of Orthopaedics **41**, 11-15.

Pfizer (2016). *Independent Grants for Learning & Change.* http://www.pfizer.com/responsibility/grants_contributions/independent_grants

Planès S, Villier C & Mallaret M (2016). *The nocebo effect of drugs*. Pharmacology Research & Perspectives **4(2)**, e00208.

Pocock SJ (1992). *When to stop a clinical trial*. BMJ **305(6847)**, 235–240.

Pocock SJ, *et al*. (2002). *Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems*. Statistics in Medicine **21(19)**, 2917-2930.

Pocock SJ & Hughes MD (1989). *Practical problems in interim analyses, with particular regard to estimation.* Controlled Clinical Trials **10 (suppl 4)**, 209-221.

Prayle AP, Hurley MN & Smyth AR (2012). *Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross-sectional study*. BMJ **344**, d7373.

Prinz F, *et al*. (2011). *Believe it or not: how much can we rely on published data on potential drug targets?* Nature Reviews Drug Discovery **10**, 712-713.

PRISMA (2015). *Welcome to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)* http://www.prisma-statement.org/

Rapoport JL, *et al.* (1978). *Dextroamphetamine: cognitive and behavioral effects in normal prepubertal boys*. Science **199**, 560-563.

Rapoport JL, *et al.* (1980). *Dextroamphetamine. Its cognitive and behavioral effects in normal and hyperactive boys and normal men*. Archives of General Psychiatry **37**, 933-943.

Rasmussen N (2008). *On Speed: The many lives of amphetamine*. NYU Press, New York.

Ravaud P, *et al*. (2009). *ARTIST (osteoarthritis intervention standardized) study of standardised consultation versus usual care for patients with osteoarthritis of the knee in primary care in France: pragmatic randomised controlled trial*. BMJ **338**, b421.

Rawal B & Deane BR (2014). *Clinical trial transparency: an assessment of the disclosure of results of company-sponsored trials associated with new medicines approved recently in Europe*. Current Medical Research and Opinion **30(3)**, 395-405.

Rawlins MD (2008). *De Testimonio. On the evidence for decisions about the use of therapeutic interventions*. The Harveian Oration of 2008. Royal College of Physicians, London. http://www.amcp.org/WorkArea/DownloadAsset.aspx?id=12451

Ray KK, *et al*. (2010). *Statins and all-cause mortality in high-risk primary prevention: a meta-analysis of 11 randomized controlled trials involving 65,229 participants.* Archives of Internal Medicine **170**, 1024–1031.

Redig AJ & Jänne PA (2015). *Basket trials and the evolution of clinical trial design in an era of genomic medicine*. Journal of Clinical Oncology **33(9)**, 975-977.

Registered Reports (2014). *Registered Reports: A step change in scientific publishing.* https://www.elsevier.com/reviewers-update/story/innovation-in-publishing/registered-reports-a-step-change-in-scientific-publishing

For an evolving list of journals that have adopted Registered Reports: https://osf.io/8mpji/wiki/home/

Review on Antimicrobial Resistance (2016). *Tackling drug-resistant infections globally: Final report and recommendations*. http://amr-review.org/sites/default/files/160525_Final%20paper_with%20cover.pdf

Risch N, *et al*. (2009). *Interaction between the serotonin transporter gene (5-HTTLPR), stressful life events, and risk of depression: a meta-analysis*. Journal of the American Medical Association **301**, 2462-2471.

Robins LN (1978). *Sturdy childhood predictors of adult antisocial behaviour: replications from longitudinal studies*. Psychological Medicine **8(4)**, 611-622.

Roche (2015). *Grants and Donations*. https://www.roche.co.uk/home/corporate-responsibility/grants-and-donations.html

Rokx C, *et al*. (2016). *Virological responses to lamivudine or emtricitabine when combined with tenofovir and a protease inhibitor in treatment-naïve HIV-1-infected patients in the Dutch AIDS Therapy Evaluation in the Netherlands (ATHENA) cohort.* HIV Medicine **17(8)**, 571-580.

Roose SP, *et al.* (2016). *Practising evidence-based medicine in an era of high placebo response: number needed to treat reconsidered*. British Journal of Psychiatry **208**, 416-420.

Rosenbaum PR (2001). *Stability in the absence of treatment*. Journal of the American Statistical Association **96**, 210-219.

Rosenbaum PR & Rubin DB (1983). *The central role of the propensity score in observational studies for causal effects*. Biometrika **70 (1)**, 41-55.

Ross JS, *et al.* (2012). *Publication of NIH-funded trials registered in ClinicalTrials.gov: cross sectional analysis*. BMJ **344**, d7292.

Rothwell PM, *et al.* (2005). *Treating Individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy*. Lancet **365(9455)**, 256–265.

Rowe SM, *et al*. (2012). *Progress in cystic fibrosis and the CF Therapeutics Development Network*. Thorax **67(10)**, 882–890.

Rubins HB, *et al*. (2002). *Diabetes, plasma insulin, and cardiovascular disease: subgroup analysis from the Department of Veterans Affairs high-density lipoprotein intervention trial (VA-HIT)*. Archives of Internal Medicine **162**, 2597-2604.

Rutter M (2006). *Genes and behaviour: Nature-nurture interplay explained*. Wiley Blackwell, London.

Rutter M (2007). *Proceeding From Observed Correlation to Causal Inference: The Use of Natural Experiments.* Perspectives on Psychological Science **2**, 377-395.

Rutter M & Pickles A (2016). *Annual Research Review: Threats to the validity of child psychiatry and psychology*. Journal of Child Psychology and Psychiatry **57(3)**, 398-416.

Rutter M & Thapar A (2015). *Using natural experiments and animal models to study causal hypotheses in relation to child mental health problems*. In Thapar A, *et al*. eds. Rutter's Child and Adolescent Psychiatry. 6th ed. Wiley Blackwell, Oxford.

Sackett DL (1989). *Rules of evidence and clinical recommendations on the use of antithrombotic agents.* Chest **95**, 2S-4S. http://www.ncbi.nlm.nih.gov/pubmed/2914516

Saeed MA, Vlasakakis G & Della Pasqua O (2015). *Rational use of medicines in older adults: Can we do better during clinical development*? Clinical Pharmacology and Therapeutics **97(5)**, 440-443.

SAIL Databank (2016). *FAQ: How are data provided, anonymised, linked and accessed?* http://www.saildatabank.com/faq

Sales AE, *et al*. (2004). A*ssessing response bias from missing quality of life data: the Heckman method*. Health and Quality of Life Outcomes **2**, 49-59.

Salsburg D (2002). *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company, New York.

Sampson RJ, Laub JH & Wimer C (2006). *Does Marriage Reduce Crime? A Counterfactual Approach To Within-Individual Causal Effects*. Criminology **44(3)**, 465–508.

Savarese G, *et al*. (2013). *Benefits of statins in elderly subjects without established cardiovascular disease: a meta-analysis*. Journal of American College Cardiology **62(22)**, 2090-2099.

Schneeweiss S, *et al*. (2009). *High-dimensional propensity score adjustment in studies of treatment effects using health care claims data*. Epidemiology **20(4)**, 512-522.

Schofield P, *et al*. (2016). *Does depression diagnosis and antidepressant prescribing vary by location? Analysis of ethnic density associations using a large primary-care dataset.* Psychological Medicine **46**, 1321-1329.

Schwartz D & Lellouch J (1967). *Explanatory and pragmatic attitudes in therapeutical trials*. Journal of Chronic Diseases **20(8)**, 636-648.

Schwartz G, *et al*. (2012). *Effects of dalcetrapib in patients with a recent acute coronary syndrome*. New England Journal of Medicine **367**, 2089–2099.

Science and Technology Committee (Commons) (2015). *Science in emergencies: UK lessons from Ebola inquiry*. http://www.parliament.uk/business/committees/committees-a-z/commons-select/science-and-technology-committee/inquiries/parliament-2015/science-in-emergencies/

Shadish WR, Cook TD & Campbell DT (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company, Boston & New York.

Sharma T, *et al*. (2016). *Suicidality and aggression during antidepressant treatment: systematic review and meta-analyses based on clinical study reports*. BMJ **352**, i65.

Simmons JP, Nelson LD & Simonsohn U (2011). *False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant*. Psychological Science **22(11)**, 1359–1366.

Singal AG, Higgins PDR & Waljee AK (2014). *A primer on effectiveness and efficacy trials*. Clinical and Translational Gastroenterology **5**, e45.

Smeeth L, *et al*. (2009). *Effect of statins on a wide range of health outcomes: a cohort study validated by comparison with randomized trials*. British Journal of Clinical Pharmacology **67**, 99-109.

Spiegelhalter DJ, *et al*. (1999). *An introduction to Bayesian methods in health technology assessment*. BMJ **319(7208)**, 508-512.

Spiegelhalter DJ, *et al*. (2000). *Bayesian methods in health technology assessment: a review*. Health Technology Assessment **4(38)**, 1-130.

Spirtes P (2010). *Introduction to Causal Inference*. Journal of Machine Learning Research **11**, 1643-1662.

Stigler K, *et al*. (2004). *Weight gain associated with atypical antipsychotic used in children and adolescents: prevalence, clinical relevance, and management*. Pediatric Drugs **6**, 33-44.

STROBE Statement (2009). *STROBE Statement: STrengthening the Reporting of OBservational studies in Epidemiology*, http://www.strobe-statement.org/

Taft AJ, *et al*. (2011). *Mothers' AdvocateS In the Community (MOSAIC)-non-professional mentor support to reduce intimate partner violence and depression in mothers: a cluster randomised trial in primary care*. BioMed Central Public Health **11**, 178-188.

Tajika A, *et al*. (2015). *Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up*. The British Journal of Psychiatry **207(4)**, 357–362.

Taylor D, Paton C & Kapur S (2015). *The Maudsley Prescribing Guidelines in Psychiatry*. 12th edition. Wiley Blackwell, Oxford.

Taylor F, *et al*. (2013). *Statins for the primary prevention of cardiovascular disease.* Cochrane Database Systematic Reviews **1**, CD004816.

The Farr Institute (2016). *About the Farr Institute*. http://www.farrinstitute.org/

The Gambia Hepatitis Study Group (1987). *The Gambia Hepatitis Intervention Study*. Cancer Research **47(21)**, 5782-5787.

Lancet (2014*). Research: increasing value, reducing waste*. http://www.lancet-journals.com/researchseries/

The Scottish Government (2012). *Joined-up data for better decisions: Guiding Principles for Data Linkage*. http://www.gov.scot/Resource/0040/00407739.pdf

The Scottish Government (2014). *Data Management Board: A Data Vision for Scotland.* http://www.gov.scot/Resource/0044/00448438.pdf

The World Association of Medical Editors (2005). *Ghost writing initiated by commercial companie*s. Journal of General Internal Medicine **20(6)**, 549-551.

Thompson SG & Higgins JPT (2005). *Treating Individuals 4: can meta-analysis help target interventions at individuals most likely to benefit?* Lancet **365(9456)**, 341–346.

Tolmie JL (2002). *Down syndrome and other autosomal trisomies*. In Rimoin DL, *et al*. eds. Emery and Rimoin's Principles and Practice of Medical Genetics. 4th Edition. Churchill Livingstone, London & New York.

Toschke AM, *et al.* (2011). *Antihypertensive treatment after first stroke in primary care: results from the General Practitioner Research Database (GPRD)*. Journal of Hypertension **29**, 154-160.

Trivedi MH, *et al.* (2006). *Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice*. American Journal of Psychiatry **163**, 28-40.

Turner RM, *et al*. (2009). *Bias modelling in evidence synthesis*. Journal of the Royal Statistical Society **172**, 21-47.

Uher R & McGuffin P (2010). *The moderation by the serotonin transporter gene of environmental adversity in the etiology of depression: 2009 update*. Molecular Psychiatry **15**, 18-22.

Uman LS (2011) *Systematic Reviews and Meta-Analyses*. Journal of the Canadian Academy of Child and Adolescent Psychiatry **20(1)**, 57–59.

United States Department of Health and Human Services, Agency for Health Care Policy and Research (1993). *Acute pain management: operative or medical procedures and trauma*. https://archive.ahrq.gov/clinic/medtep/acute.htm

UpToDate (2015). *Grading Tutorial. Grading Recommendations in UpToDate.* http://www.uptodate.com/home/grading-tutorial#

US Food and Drug Administration (2015). *FDA's Sentinel Initiative.* http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm

US Preventive Services Task Force (1989). *Guide to Clinical Preventive Services*. Appendix A. DIANE Publishing.

US Preventive Services Task Force (1996). *Guide to clinical preventive services: report of the U.S. Preventive Services Task Force.* 2nd ed. Williams & Wilkins, Baltimore.

US Preventive Services Task Force (2008). *Procedure Manual*. Appendix VII. AHRQ Publication No. 08-05118-EF.

US Preventive Services Task Force (2012). *Grade Definitions*. http://www.uspreventiveservicestaskforce.org/Page/Name/grade-definitions

Vandenbroucke JP (2004). *When are observational studies as credible as randomised trials?* Lancet **363**, 1728–1731.

Van Marwijk HWJ, *et al*. (2008). *Primary care management of major depression in patients aged ≥55 years: outcome of a randomised clinical trial*. British Journal of General Practice **58**, 680–687.

Visser J & Jehan Z (2009). *ADHD: a scientific fact or a factual opinion? A critique of the veracity of ADHD.* Emotional and Behavioural Difficulties **14(2)**, 127-140. http://www.tandfonline.com/doi/abs/10.1080/13632750902921930

Voight BF, *et al.* (2012). *Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study*. Lancet **380**, 572-580.

Wang AT, *et al.* (2010). *Association between industry affiliation and position on cardiovascular risk with rosiglitazone: cross sectional systematic review*. BMJ (Clinical Research Edition.) **340**, c1344.

Wathen JK & Thall PF (2008). *Bayesian adaptive model selection for optimizing group sequential clinical trials*. Statistics in Medicine **27**, 5556-5604.

Wellcome Trust (2016). *Funding for clinical trials.* http://www.wellcome.ac.uk/Funding/Biomedical-science/Application-information/WTX022708.htm

Wellcome Trust Case Control Consortium (2007). *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls*. Nature **447(7145)**, 661-678.

Wieseler B, *et al.* (2012). *Impact of document type on reporting quality of clinical drug trials: a comparison of registry reports, clinical study reports, and journal publications*. BMJ **344**, d8141.

Wilks DC, *et al*. (2011). *Objectively measured physical activity and fat mass in children: a bias-adjusted meta-analysis of prospective studies.* PLoS ONE **6(2)**, e17205.

Williams B, *et al*. (2015). *Spironolactone versus placebo, bisoprolol, and doxazosin to determine the optimal treatment for drug-resistant hypertension (PATHWAY-2): a randomised, double-blind, crossover trial.* Lancet **386(10008)**, 2059-2068.

Woolcock M (2013). *Using case studies to explore the external validity of 'complex' development interventions*. Evaluation **19(3)**, 229-248.

Writing Group for the Women's Health Initiative Investigators (2002). *Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial*. Journal of the American Medical Association **288(3)**, 321-333.

Würtz P, *et al*. (2016). *Metabolomic profiling of statin use and genetic inhibition of HMG-CoA reductase*. Journal of the American College of Cardiology **67**, 1200-1210.

Yellow Card (2016). *About Yellow Card*. https://yellowcard.mhra.gov.uk/the-yellow-card-scheme/

Zhang J (2008). *Causal reasoning with ancestral graphs*. The Journal of Machine Learning Research **9**, 1437-1474.

Zimmerman M (2016).*The FDA's failure to address the lack of generalisability of antidepressant efficacy trials in product labelling*. The British Journal of Psychiatry **208**, 512-514.

Zwarenstein M, *et al.* (2008). *Improving the reporting of pragmatic trials: an extension of the CONSORT statement.* BMJ **337**, a239.

The Academy of Medical Sciences is the independent body in the UK representing the diversity of medical science. Our elected Fellows are the UK's leading medical scientists from hospitals, academia, industry and the public service. Our mission is to advance biomedical and health research and its translation into benefits for society. We are working to secure a future in which:

- UK and global health is improved by the best research.
- The UK leads the world in biomedical and health research, and is renowned for the quality of its research outputs, talent and collaborations.
- Independent, high quality medical science advice informs the decisions that affect society.
- More people have a say in the future of health and research.

Our work focusses on four key objectives, promoting excellence, developing talented researchers, influencing research and policy, and engaging patients, the public and professionals.

www.acmedsci.ac.uk

The Academy of Medical Sciences

Academy of Medical Sciences
41 Portland Place
London, W1B 1QH

@acmedsci

+44(0)20 3141 3200
info@acmedsci.ac.uk
www.acmedsci.ac.uk

Registered Charity No. 1070618
Registered Company No. 3520281