Evidence-based Oncology

Edited by

Chris Williams
Cochrane Cancer Network, Institute of Health Sciences, Oxford, UK

Associate editors

Vivien Bramwell
Tom Baker Cancer Centre, Calgary, Canada

Xavier Bonfill
Iberoamerican Cochrane Centre, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Jack Cuzick
Imperial Cancer Research Fund, London, UK

John F Forbes
Department of Surgical Oncology, Newcastle Mater Hospital, University of Newcastle, NSW, Australia

Robin Grant
Department of Clinical Neurosciences, University of Edinburgh, UK

David Guthrie
Derbyshire Royal Infirmary, Derby, UK

Malcolm Mason
Department of Clinical Oncology, Velindre NHS Trust, Whitechurch, Cardiff, UK

Peter Simmonds
Cancer Research UK Oncology Unit, Southampton General Hospital, UK

Arun D Singh
Wills Eye Hospital, Oncology Service, Philadelphia, USA

Rosalynne Weston
Cochrane Cancer Network, Oxford, UK

Hywel C Williams
Centre of Evidence-Based Dermatology, University of Nottingham, UK

BMJ Books

Evidence-based Oncology

Evidence-based Oncology

Edited by

Chris Williams
Cochrane Cancer Network, Institute of Health Sciences, Oxford, UK

Associate editors

Vivien Bramwell
Tom Baker Cancer Centre, Calgary, Canada

Xavier Bonfill
Iberoamerican Cochrane Centre, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Jack Cuzick
Imperial Cancer Research Fund, London, UK

John F Forbes
Department of Surgical Oncology, Newcastle Mater Hospital, University of Newcastle, NSW, Australia

Robin Grant
Department of Clinical Neurosciences, University of Edinburgh, UK

David Guthrie
Derbyshire Royal Infirmary, Derby, UK

Malcolm Mason
Department of Clinical Oncology, Velindre NHS Trust, Whitechurch, Cardiff, UK

Peter Simmonds
Cancer Research UK Oncology Unit, Southampton General Hospital, UK

Arun D Singh
Wills Eye Hospital, Oncology Service, Philadelphia, USA

Rosalynne Weston
Cochrane Cancer Network, Oxford, UK

Hywel C Williams
Centre of Evidence-Based Dermatology, University of Nottingham, UK

BMJ Books

BMJ Books is an imprint of the BMJ Publishing Group

First published in 2003
by BMJ Books, BMA House, Tavistock Square,
London WC1H 9JR

www.bmjbooks.com

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN 0 7279 1439 1

Typeset by SIVA Math Setters, Chennai, India
Printed and bound by MPG Books, Bodmin, Cornwall

Contents

Contributors

Jaffer Ajani
Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Anca Ansink
Erasmus MC Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

Ferran Ariza
Iberoamerican Cochrane Centre, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Wendy Atkin
Cancer Research UK Colorectal Cancer Unit, St Mark's Hospital, Harrow, UK

Jim Barber
Velindre NHS Trust, Cardiff, UK

Veronique Bataille
Twin Research Genetic Epidemiology Unit, St Thomas' Hospital, London, UK

Fiona Bath
Centre for Evidence-based Dermatology, Queen's Medical Centre, Nottingham, UK

MJ van den Bent
Department of Neuro-Oncology, Dr. Daniel den Hoed Cancer Center, Rotterdam, The Netherlands

William C Black
Department of Radiology, Dartmouth-Hitchcock Medical Center, Lebanon, New Hampshire, USA

Xavier Bonfill
Iberoamerican Cochrane Centre, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Michael Brada
Neuro-oncology Unit and Academic Unit of Radiotherapy & Oncology, The Institute of Cancer Research and the Royal Marsden NHS Trust, Sutton, UK

Vivien Bramwell
Tom Baker Cancer Centre, and University of Calgary, Calgary, Canada

D Brinkmann
Gynaecology Oncology Unit, Bart's and The London Queen Mary's School of Medicine and Dentistry, London, UK

Kathryn Burgon
Cochrane Unit, Velindre NHS Trust, Whitchurch, Cardiff, UK

Ian Chau
Gastrointestinal Unit, Department of Medicine, Royal Marsden Hospital, Sutton, UK

Chris Coppin
Medical Oncology, British Columbia Cancer Agency Fraser Valley Centre, Surrey, Canada

ER Copson
Cancer Research UK Medical Oncology Unit, The Royal South Hants Hospital, Southampton, UK

Thomas Crosby
Velindre Hospital, Whitechurch, Cardiff, UK

David Cunningham
Gastrointestinal Unit, Department of Medicine, Royal Marsden Hospital, Sutton, UK

Jack Cuzick
Imperial Cancer Research Fund Cancer Research UK, London, UK

RJ Davies
Department of General Surgery, Addenbrooke's Hospital, Cambridge, UK

Lena van Doorn
University Hospital Utrecht, Utrecht, The Netherlands

Stephen W Duffy
Queen Mary University of London, London, UK

Simon Everett
Department of Gastroenterology, Leeds General Infirmary, Leeds, UK

Richard M Fagerstrom
Biometry Research Group, National Cancer Institute, Bethesda, Maryland, USA

Rebecca L Faulkner
Academic Unit of Obstetrics and Gynaecology and Reproductive Health Care, St Mary's Hospital, Manchester, UK

John F Forbes
Department of Surgical Oncology, Newcastle Mater Hospital, University of Newcastle, NSW, Australia

David Forman
Unit of Epidemiology and Health Services Research, Medical School, University of Leeds, Leeds, UK

Paula Ghaneh
Department of Surgery, University of Liverpool, Liverpool, UK

Julie Glanville
NHS Centre for Reviews & Dissemination, University of York, UK

John K Gohagan
Early Detection Research Group, National Cancer Institute, Bethesda, Maryland, USA

Robin Grant
Department of Clinical Neurosciences, University of Edinburgh, UK

David Guthrie
Derbyshire Royal Infirmary, Derby, UK

Santosh G Honavar
Ocular Oncology Service, LV Prasad Eye Institute, Hyderabad, India

TJ Iveson
Cancer Research UK Medical Oncology Unit, The Royal South Hants Hospital, Southampton, UK

Ian Jacobs
Unit of Epidemiology and Health Services Research, Medical School, University of Leeds, Leeds, UK

Philip J Johnson
Division of Cancer Studies, School of Medicine, University of Birmingham, Birmingham, UK

Katharine Johnston
Health Economics Research Centre, Institute of Health Sciences, University of Oxford, UK

Henry C Kitchener
Academic Unit of Obstetrics and Gynaecology and Reproductive Health Care, St Mary's Hospital, Manchester, UK

Barnett S Kramer
Disease Prevention, National Institutes of Health, Bethesda, Maryland, USA

Julia Kumpf
Clinical Economics Group, University Hospital Ulm, Germany

Howard Kynaston
Department of Urology, University Hospital of Wales, Cardiff, UK

Tim Lancaster
Cochrane Tobacco Addiction Review Group, Institute of Health Sciences, Oxford, UK

Nannette J Liégeois
Department of Dermatology, Harvard Medical School, Lahey Clinic, Burlington, MA, USA

Imogen Locke
Department of Radiotherapy and Oncology, Royal Marsden Hospital, Sutton, UK

Conor Magee
Department of Surgery, University of Liverpool, Liverpool, UK

Pamela M Marcus
Biometry Research Group, National Cancer Institute, Bethesda, Maryland, USA

Malcolm Mason
Department of Clinical Oncology, Velindre NHS Trust, Whitechurch, Cardiff, UK

Miquel Mateu
Thoracic Surgery Service, Hospital Mutua de Terrassa, Barcelona, Spain

Jane Melia
Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, UK

Usha Menon
Gynaecology Oncology Unit, Bart's and The London Queen Mary's School of Medicine and Dentistry, London, UK

M Michael
Division of Haematology and Medical Oncology, Peter MacCallum Cancer Institute, Victoria, Australia

R Miller
Department of General Surgery, Addenbrooke's Hospital, Cambridge, UK

Kate Misso
NHS Centre for Reviews and Dissemination, University of York, UK

Paul Moayyedi
Gastroenterology and Health Services Research, University of Birmingham, UK

Sue Moss
Cancer Screening Evaluation Unit, Institute of Cancer Research, Sutton, UK

John P Neoptolemos
Department of Surgery, University of Liverpool, Liverpool, UK

Alan Neville
McMaster University, Hamilton, Canada

ES Newlands
Charing Cross Hospitals, London, UK

Suzanne Olbricht
Department of Dermatology, Harvard Medical School, Lahey Clinic, Burlington, MA, USA

Tim Oliver
Medical Oncology Department, Royal Hospital NHS Trust, St Bartholomews Hospital, London, UK

William Perkins
Consultant Dermatologist and Dermatologic Surgeon, University Hospital, Queen's Medical Centre, Nottingham, UK

Alexandria Phan
Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Paul F Pinsky
Early Detection Research Group, National Cancer Institute, Bethesda, Maryland, USA

Ernst Pöppel
Institute of Medical Psychology and Human Studies Center, Ludwig Maximilians University Munich, Germany

Franz Porzsolt
Clinical Economics Group, University Hospital Ulm, and Evidence Based Health Care, Human Studies Center, Ludwig Maximilians University Munich, Germany

Philip C Prorok
Biometry Research Group, National Cancer Institute, Bethesda, Maryland, USA

Ramón Rami-Porta
Thoracic Surgery Service, Hospital Mutua de Terrassa, Barcelona, Spain

Ian G Rennie
Department of Ophthalmology and Orthoptics, University of Sheffield, Sheffield, UK

Dafydd Roberts
Department of Dermatology, Singleton Hospital, Swansea, UK

Trevor Roberts
Northern Centre for cancer Treatment, Newcastle General Hospital, Newcastle-upon-Tyne, UK

Paul A Rundle
Department of Ophthalmology and Orthoptics, Royal Hallamshire Hospital, Sheffield, UK

Peter Sasieni
Cancer Research UK, Department of Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, London, UK

John H Scholefield
Division of GI Surgery, University Hospital, Nottingham, UK

Mireia Serra
Thoracic Surgery Service, Hospital Mutua de Terrassa, Barcelona, Spain

Mike Shelley
Cochrane Prostatic Diseases and Urological Cancers Group, Cochrane Unit, Research Department, Velindre NHS Trust, Cardiff, UK

Peter Simmonds
Cancer Research UK Oncology Unit, Southampton General Hospital, UK

Arun D Singh
Wills Eye Hospital, Oncology Service, Philadelphia, USA

Steven Skates
Department of Medicine, Harvard Medical School and Massachusetts General Hospital, Boston, USA

Margaret F Spittle
Meyerstein Institute of Oncology, Middlesex Hospital, London, UK

Lesley Stewart
Meta-Analysis Group, MRC Clinical Trials Unit, London, UK

RP Symonds
University of Leicester, Leicester Royal Infirmary, Leicester, UK

Elinor Thompson
Iberoamerican Cochrane Centre, Hospital de la Santa Creu i Sant Pau, Barcelona, Spain

Paul A Vasey
Cancer Research UK Department of Medical Oncology, University of Glasgow, Glasgow, UK

Jacobus van der Velden
Academic Medical Center, Amsterdam, The Netherlands

Charles J Vecht
Neuro-oncology Unit, Department of Neurology, Medical Center, The Hague, The Netherlands

Shail Verma
Ottawa Regional Cancer Centre, Ottawa, Canada

Rosalynne Weston
Focus Freelance Writing and Research, Salisbury, and Cochrane Cancer Network, Oxford, UK

Sean Whittaker
St John's Institute of Dermatology, Guy's and St Thomas' Hospital, London, UK

IR Whittle
University of Edinburgh, Edinburgh, UK

Nicholas Wilcken
Department of Medical Oncology, Westmead Hospital, Westmead, Australia

Chris Williams
Cochrane Cancer Network, Institute of Health Sciences, Oxford, UK

Hywel C Williams
Centre of Evidence-Based Dermatology, University of Nottingham, UK

JR Zalcberg
Division of Haematology and Medical Oncology, Peter MacCallum Cancer Institute, Victoria, Australia

Preface

Cancer has been at the forefront of developments in clinical trials methodology for the past half-century. Some of the earliest randomised controlled trials (RCTs) were in the field of oncology and clinical research has routinely used RCTs to assess new therapies. However, before we congratulate ourselves, perhaps prematurely, we need to look to see how the evidence from RCTs and other clinical studies has been used.

The movement called evidence-based medicine has not suddenly invented the concept of using evidence from clinical experiments – Western medicine has been predicated on this concept for many hundreds of years. However, the tool of the RCT has been honed and we now have the means to synthesise information in a systematic manner that reduces the risk of bias. These developments have come at a time when electronic communication has allowed us for the first time to keep an effective track of clinical research, scientific publications and to bring all of the information together using the methodology of systematic reviews.

In the past, the need for reviews of current knowledge was met by "narrative reviews", usually written by an expert in the field. Such individuals were held to have a thorough grasp of the literature and the ability to interpret it. Readers of reviews need unbiased information and there is consistent evidence (including from cancer) that narrative reviews of healthcare interventions rarely use methods designed to reduce the risk of bias. The fact that narrative reviews are written by experts in the field who have often carried out some of the research that they are reviewing compounds the risk of bias.

Systematic reviews, the bedrock of evidence-based medicine, are designed to reduce the risk of bias. They also bring together all of the pertinent evidence from trials judged to be of good quality. Where appropriate, evidence from these trials can be pooled using the technique of meta-analysis. The result of such a process is to reduce the risk of bias and to maximise the chances of finding a useful outcome, because the intervention is being examined in the largest population possible and not in a few discreet RCTs.

This book uses an evidence-based approach to look at the strength of the underlying evidence used to support some of the key decisions in cancer care. The authors have not been commissioned to carry out systematic reviews to answer each of these questions. Each systematic review is a complex and time-consuming exercise and it would be impractical in a book of this nature. Authors were asked to use systematic searches of the medical literature and to summarise their findings. Where there are systematic reviews these were presented and discussed in the light of the rest of the literature. Where no systematic reviews were available reviewers summarised the available literature with a particular emphasis on RCTs. However, new systematic reviews were usually not carried out.

The conclusions for each review question have been graded according to the strength of the evidence underlying that conclusion. Readers may wish to use these grades in thinking about the believability of the conclusions, but should bear in mind that such grades are a crude approximation. While they provide a summary of the strength of evidence, they are also included to stimulate readers to automatically think about how believable the evidence really is.

Inevitably, in a book of this type there will be many questions that were not included. The limited list of key questions were selected by the authors with the section editor. Often this was driven by those areas where there was known to be RCTs and sometimes systematic reviews. There is an emphasis on questions deemed to be of significance by clinicians because this is where the research has been carried out. Questions of particular importance to patients and their families have often been much less well researched and because of this are even harder to systematically review.

Users of this book should be able to read about the evidence underlying many of the decisions that underpin our current approach to treating the common cancers. Often clear conclusions elude us because there is a paucity of data and some of it is of doubtful quality. In these circumstances systematic reviews are often an essential starting point of new trials.

I hope that readers will find that this book is a useful starting place when looking for evidence for how we currently treat cancer.

Abbreviations

5-FU	5-fluorouracil
AA	anaplastic astrocytoma
AD	dissection of axillary nodes
AFP	alpha-fetoprotein
AIDS	acquired immune deficiency syndrome
AJCC	American Joint Committee on Cancer
AMED	Allied and Complementary Medicine Database
AMS	atypical mole syndrome
APC	adenomatous polyposis coli
APR	abdominoperineal resection
AR	anterior resection
ASCO	American Society of Clinical Oncology
ATLL	adult T-cell leukaemia lymphoma
BCC	basal cell carcinoma
BED	biologically effective dose
BMI	Body Mass Index
BNI	British Nursing Index
BP	breast preservation
BPH	benign prostatic hyperplasia
CBA	cost-benefit analysis
CCTR	Cochrane Controlled Trials Register
CDSR	Cochrane Database of Systematic Reviews
CEA	carcinoembryonic antigen
CEA	cost-effectiveness analysis
CHART	continuous hyperfractionated accelerated radiotherapy
CHRPE	congenital hypertrophy of the retinal pigment epithelium
CI	confidence interval
CIN	cervical intraepithelial neoplasia
CLE	complete local excision of the primary tumour
CR	cumulative risk
CRC	colorectal cancer
CT	computerised tomography
CTCL	cutaneous T-cell lymphoma
CUA	cost-utility analysis
DARE	Database of Abstracts of Reviews of Effects
DCIS	duct carcinoma *in situ*
DDFS	distant disease-free survival
DFS	disease-free survival
DPD	dihydropyrimidine dehydrogenase
DRE	digital rectal examination
DRS	disease-related symptoms
DSS	disease-specific survival
EBM	evidence-based medicine
EBRT	external beam radiotherapy
EBT	endobronchial brachytherapy
ECG	electrocardiogram
ECP	extracorporeal photopheresis
ED&C	electrodesiccation and curettage
EFS	event-free survival
EGFR	epidermal growth factor receptor
ERCP	endoscopic cholangiopancreatography
ERM	extended radical mastectomy
EUS	endoscopic ultrasound
FAMMM	familial atypical multiple mole melanoma
FAP	familial adenomatous polyposis
FBC	full blood count
FNA	fine needle aspiration
FOBT	faecal occult blood test
GDEPT	gene-directed enzyme prodrug therapy
GORD	gastro-oesophageal reflux disease
GTD	gestational trophoblastic disease
HAART	highly active antiretroviral therapy
HAI	hepatic arterial influsional
HBV	hepatitis B virus
HCC	hepatocellular carcinoma
hCG	human chorionic gonadotrophin
HCV	hepatitis C virus
HGG	high grade glioma
HHV8	human herpes virus 8
HIV	human immunodeficiency virus
HNPCC	hereditary non-polyposis colorectal carcinoma
HPV	human papillomavirus
HTA	Health Technology Assessment
HTLV-I	human T-lymphotrophic virus type-1
IARC	International Agency for Research on Cancer
ICER	incremental cost-effectiveness ratio
IFNα	interferon alfa
IGF-1	insulin-like growth factor 1
IL-2	interleukin-2
IMH	immunohistochemistry
IORT	intraoperative radiotherapy
IPD MA	individual patient data meta-analyses
IPF	independent prognostic factor
IPSID	immunoproliferative small intestinal disease
IRA	ileorectal anastomosis
KPS	Karnofsky Performance Status
KS	Kaposi's sarcoma
KSHV	KS-associated herpesvirus
LAVH	laparoscopic-assisted vaginal hysterectomy
Lev	levamisole
LFT	liver function tests
LGG	low grade glioma
LHRH	luteinising hormone-releasing hormone
LINAC	linear accelerator
LLETZ/LEEP	large loop excision of the transformation zone
LM	lentigo maligna
LMM	lentigo maligna melanoma
LPA	lysophosphatidic acid
LR	local recurrence
LSS	limb salvage surgery
LV	leucovorin

MALTomas	mucosa-associated lymphoid tissue lymphomas
MeSH	Medical Subject Headings
MF/SS	mycosis fungoides/Sézary syndrome
MFH	malignant fibrous histiocytoma
MM	malignant melanomas
MMP	matrix metalloproteinase
MMS	Mohs micrographic surgery
MRCP	magnetic resonance cholangiopancreatography
MSI	microsatellite instability
MSI-H	microsatellite instability – high
NCI	National Cancer Institute
NICE	National Institute for Clinical Excellence
NMSC	non-melanomic skin cancer
NR	not reported
NSCLC	non-small cell lung cancer
OMNI	Organising Medical Networked Information
OR	odds ratio
OS	overall survival
OSSN	ocular surface squamous neoplasia
PA	para-aortic
PCI	prophylactic cranial irradiation
PCNA	proliferating cell nuclear antigen
PD	progressive disease
PDECGF	platelet derived endothelial cell growth factor
PDT	photodynamic therapy
PEI	percutaneous alcohol injection
PET	positron emission tomography
PICO/PIOC	population, intervention, comparison and outcome
PJS	Peutz–Jeghers syndrome
PORT	postoperative radiotherapy
PSA	prostate specific antigen
PSTT	placental site trophoblastic tumours
PTC	percutaneous transhepatic cholangiography
PVI	portal vein infusion
QALYs	quality adjusted life-years
QoL	quality of life
Q-TWIST	quality adjusted time without symptoms or toxicity
RCC	renal cell cancer
RM	radical mastectomy
RP	radical prostatectomy
RPC	restorative proctocolectomy
RR	relative risk
SCC	squamous cell carcinoma
SCLC	small cell lung cancer
SD	stable disease
SERM	selective oestrogen receptor modulator
SIL	squamous intraepithelial lesions
SLNB	sentinel lymph node biopsy
SR	systematic reviews
SRS	stereotactic radiotherapy
SSS	sphincter sparing surgery
STS	soft tissue sarcoma
TACE	transarterial chemoembolisation
Tc-MAA	technetium-99m macroaggregated albumin scan
TIMP	tissue inhibitor of matrix metalloproteinase
TM	total mastectomy
TME	total mesorectal excision
TNM	Tumour, Node, Metastasis classification
TP	thymidine phosphorylase
TRUS	transrectal ultrasound
TS	thymidylate synthase
TSEB	total skin electron beam therapy
TUR	transurethral resection
U&E	urea and electrolytes
VAIN	vaginal intraepithelial neoplasia
VEGF	vascular endothelial growth factor
WBRT	whole brain radiotherapy
WW	watchful waiting
XRR	external beam radical radiotherapy
XRT	external beam radiotherapy

Levels of evidence and grades of recommendation used in *Evidence-based Oncology*

Levels of evidence and grades of recommendation appear within the text in the clinical chapters, for example, **Evidence Level Ia** and **Grade A**.

Levels of evidence

Ia	Meta-analysis of randomised controlled trials (RCTs)
Ib	At least 1 RCT
IIa	At least 1 non-randomised study
IIb	At least 1 other well designed quasi-experimental study
III	Non-experimental, descriptive studies
IV	Expert committee reports or opinions/experience of respected authorities

Grades of recommendations

A	At least one RCT as part of body of literature of overall good quality and consistency addressing recommendation **Evidence levels Ia, Ib**
B	No RCT but well conducted clinical studies available **Evidence levels IIa, IIb, III**
C	Expert committee reports or opinions/experience of respected authorities in the absence of directly applicable good quality clinical studies **Evidence level IV**

From *Clinical Oncology* (2001)**13**:S212
Source of data: MEDLINE, *Proceedings of the American Society of Medical Oncology* (ASCO).

Section I

Principles and practice of “critical appraisal”

Chris Williams, Editor

1 Appraising clinical literature in cancer

Chris Williams

We are all of us, whether we are consumers, researchers, or policy makers, inundated with unmanageable and increasing amounts of information on health care. This chapter discusses how we can best appraise and use this information, whether it be from clinical trials or clinical reviews.

Sackett and Haynes[1] have defined evidence-based medicine (EBM) as "the practice [of EBM] is a process of life-long, problem-based learning in which care for our own patients creates the need for evidence about diagnosis, prognosis, therapy and other clinical and health care issues". EBM has been criticised partly because it can be taken to suggest that evidence has just been discovered – it would be more accurate to say that EBM expresses that we can base care on better evidence than we were able to in the past. This is because we have better evidence from trials and better ways of synthesising this evidence. We also now have the technology to transfer large amounts of data easily.

This book is based on the premise that patients and their professional carers need the best available evidence when making clinical decisions.[2] It suggests that we need systematic reviews to efficiently integrate valid information and provide a basis for rational decision making.[3] It is also acknowledged that careful review of the literature is a complicated and time-consuming business and that clinicians are not in a position to carry out systematic reviews on all of the questions that they encounter in clinical practice. As well as understanding how to appraise reports of clinical research, clinicians also need to understand how to appraise systematic reviews. Studies have shown that evidence to support decision making is required more often that most clinicians realise – partly because so many decisions are taken to be routine and the evidence underlying them is not questioned (Covell 1985).[4]

The use of explicit, systematic methods in reviews limits bias (systematic errors) and reduces chance effects by increasing the number of participants, thus providing more reliable results upon which to draw conclusions and make decisions.[5,6] Systematic reviews can establish where effects of health care are consistent and can be applied across populations and in different settings. They can also show where effects may vary significantly.

Meta-analysis, the use of statistical methods to summarise the results of independent studies, can provide more precise estimates of the effects of health care than those derived from the individual studies included in a systematic review.[7–10] Systematic reviews ideally include meta-analysis, but often this is not possible because the questions, trial populations and method of delivering therapy were too variable to allow meaningful pooling of results in a meta-analysis.

Recognition of the key role of reviews in synthesising and disseminating the results of research has prompted people to consider the validity of narrative reviews. Social science and psychology led this field and it was not until the late 1980s that people drew attention to the poor scientific quality of healthcare review articles.[11–13] The first survey of the quality of narrative reviews in cancer was not published until 1997.

This chapter focuses on appraisal of and systematic review of randomised controlled trials (RCTs) because they are likely to provide more reliable information than other sources of evidence on the effectiveness of different therapies.[14] Systematic reviews of other types of evidence can be useful to those wanting to make better decisions about health care, when RCTs are not available. The basic principles of reviewing non-RCT research are the same, although meta-analysis is often not appropriate and care should be taken not to overinterpret the results.

What is the evidence that RCTs are the best way to test new treatments?

Although it has been long accepted that RCTs are the best way of testing clinical effectiveness, there are few systematic studies testing this hypothesis. The historical data from a variety of conditions supports the contention that randomised trials are more reliable than historically controlled or uncontrolled trials.[15] Sacks *et al.*[16] examined the outcomes in six different clinical questions that had been tested in both RCTs and historically controlled trials (HCTs). Box 1.1 shows that HCTs grossly overestimated the potential benefit of treatment compared with RCTs. Importantly, the differences in outcomes between RCTs and HCTs lay in the outcomes in the control group (where HCT patients fared worse than RCT patients) and not in the experimental arm where the results were similar in HCT and RCT patients.

Box 1.1 A study of comparative results of RCTs and HCTs asking the same question

- 6 Therapies, 50 RCTs, 56 HCTs
- 44 of 56 HCTs (79%) found the "new" therapy to be significantly better than the control
- 10 of 50 RCTs (20%) found the "new" therapy to be significantly better than the control
- The outcomes for new treatments were similar regardless of whether they were from RCTs or HCTS
- Outcomes were clearly worse for control patients in HCTs when compared with control patients in RCTS

Abbreviations: RCTs, randomised control trials; HCTs, historically controlled trials

This might seem academic, but failure to identify effective treatments may delay their use by years and ineffective treatments may be recommended when they are toxic or where there are other genuinely effective therapies. Such misinformation can cause real harm. One of the questions included in the paper by Sacks *et al.* was the use of diethylstilbestrol (DES) in women who have had recurrent miscarriages. Four HCTs were published in the 1960s that appeared to show that DES was highly effective in preventing habitual abortion. However, three RCTs showed that DES had no effect and that the outcome in the control group of the four HCTs was particularly poor (Table 1.1). On the basis of the HCT evidence millions of women worldwide erroneously received DES during pregnancy in an attempt to reduce the chance of miscarriage. Long-term follow up of the RCTs have revealed major toxicity of the DES given during pregnancy when the fetus is vulnerable. The finding of a major excess of vaginal clear cell cancers in the daughters of the DES-treated women was devastating. In addition follow up has shown that the male and female offspring of DES women have an increased incidence of depression and that male offspring are much less likely to form stable long-term relationships.[17–19]

RCTs have become the accepted way of testing therapies because the process of randomisation helps minimise the risk of bias. Where there is doubt that randomisation was adequately concealed, there is strong evidence that the outcome is biased.[20] In this observational study they assessed the methodological quality of 250 controlled trials from 33 meta-analyses (from the Cochrane Pregnancy and Childbirth Database) and then analysed, using multiple logistic regression models, the associations between those assessments and estimated treatment effects. The main outcome measures included associations between estimates of treatment effects and inadequate allocation concealment. Compared with trials in which authors reported adequately concealed treatment allocation, trials in which concealment was either inadequate or unclear (did not report or incompletely reported a concealment approach) yielded larger estimates of treatment effects ($P < 0{\cdot}001$). Odds ratios were exaggerated by 41% for inadequately concealed trials and by 30% for unclearly concealed trials (adjusted for other aspects of quality). They concluded that there is empirical evidence that inadequate concealment of randomisation is in controlled trials associated with bias.

Although safe randomisation is the key to a reliable RCT, attention also needs to be paid to other features. Among these, sufficient power (number of events) and appropriate endpoints are very important. An individual patient data meta-analysis of 52 RCTs of chemotherapy for non-small cell cancer found fewer than 10 000 patients who had been treated over three decades.[21] During this time period many millions would have died of this disease. None of the trials was powered to answer the questions being asked, the mean size of the treatment arms being less than 100 at a time when only a small benefit was plausible. Although chemotherapy is largely palliative in this setting, there was no usable outcome data on symptom control or quality of life.

Also of paramount importance is the question itself. In addition to the 52 RCTs included in this review a large number of RCTs were identified where the comparison was between two different types of chemotherapy. This was in spite of a lack of evidence that any chemotherapy could provide benefit to patients with advanced non-small cell lung cancer. Clearly, a key to a good RCT is to make the appropriate comparison.

The development of the CONSORT statement and its subsequent iterations will hopefully increase the quality of current and future trials and their reporting.[22] Check lists

Table 1.1 Comparison of the results of RCTs and HCTs testing the ability of diethylstilbestrol to prevent recurrent abortion

Type of trial	No. of trials	No. of patients	% Live infants DES – treated	Control
RCT	3	2175	87·3	87·6
HCT	4	2358	85·3	56
HCT [matched]	1	216	45	8

Table 1.2 Items that should be included in reports of randomised trials

Heading	Subheading	Descriptor
Title		Identify the study as randomised trial
Abstract		Use a structured format
Introduction		State prospectively defined hypothesis, clinical objectives, and planned subgroup or covariate analysis
Methods	Protocol	*Describe the:* Planned study population, together with inclusion or exclusion criteria Planned interventions and their timing Primary and secondary outcome measure(s) and the minimum important difference(s), and indicate how the target sample size was projected Rationale and methods for statistical analyses, detailing the main comparative analyses and whether they were completed on an intention-to-treat basis Prospectively defined stopping rules (if warranted)
	Assignment	*Describe the:* Unit of randomisation (for example individual, cluster, geographic) Method used to generate allocation schedule Method of allocation concealment and timing of assignment Method to separate the generator from the executor of assignment
	Masking (blinding)	*Describe the:* Mechanism (for example capsules, tables) Similarity of treatment characteristics (for example appearance, taste) Allocation schedule control (location of code during trial and when broken) Evidence for successful blinding among participants, person doing intervention, outcome assessors, and data analysts
Results	Participant flow and follow up	Provide a trial profile summarising participant flow, numbers and timing of randomisation assignment, interventions and measurements for each randomised group State estimated effect of intervention on primary and secondary outcome measures, including a point estimate and measure of precision (confidence interval)
	Analysis	State results in absolute numbers when feasible (for example 10/20 not 50%) Present summary data and appropriate descriptive and interferential statistics in sufficient detail to permit alternative analyses and replication Describe prognostic variables by treatment group and any attempt to adjust them Describe protocol deviations from the study as planned, together with the reasons
Discussion		State specific interpretation of study findings, including sources of bias and imprecision (internal validity) and discussion of external validity, including appropriate quantitative measures when possible State general interpretation of the data in light of the totality of the available evidence

(Table 1.2) should help improve the quality of published reports of clinical trials and this will aid in synthesising the literature.

In addition to the potential benefit to be gained from improving medical knowledge in general, there is some evidence that inclusion in an RCT is beneficial to patients regardless of the outcome of the trial. Braunholtz *et al.*[23] carried out a systematic review of the literature. They found only 14 research articles (covering more than 21 trials) with relevant primary data. They found that the evidence available was limited in breadth (coming largely from cancer trials) and quality, as well as quantity. There was weak evidence to suggest that clinical trials have a positive effect on the outcome of participants. This does not appear to depend strongly on the trial demonstrating that an experimental treatment is superior. However, benefit to participants is less evident where scope for a "protocol/Hawthorne effect" (benefit from improved routine care within a trial) was apparently limited (because there was no effective routine treatment or because the comparison group also received protocol care). A form of bias, arising if clinicians who tend to recruit to trials also tend to be better clinicians, could also explain these results. They concluded that, while the evidence is not conclusive, it is more likely that clinical trials have a positive rather than a negative effect on the outcome of patients. They found that the effect seems to be larger in trials where an effective treatment already exists and is included in the trial protocol.

Currently very few patients are entered into RCTs. In the UK the current NHS Cancer Plan aims to double recruitment from 3% to 6%. There are complex factors that stop patients being recruited into cancer trials and research into how to improve recruitment is sorely needed.

Why do we need reviews?

Apart from the need to find time-efficient means of using the literature to help make decisions, there is good evidence that a systematic approach can produce results that change practice. Systematic reviews of therapy for acute myocardial infarction[5] show how careful review of all of the evidence can change thinking (Figure 1.1). Early experience with thrombolytic therapy was largely ignored and narrative reviews and textbooks failed to routinely recommend such treatment for 10–15 years after meta-analysis would have shown these treatments to be effective. Conversely, lidocaine (lignocaine) has been consistently recommended for use in myocardial infarction by narrative reviews and textbooks, when there was no evidence of benefit. Thus, systematic reviews could, in this situation, change practice and help researchers to develop new trials.

Reviews are useful because they:

- are an efficient use of time
- can help support individual patient decisions
- can help in preparing guidelines and treatment protocols
- can help in developing and planning new clinical research.

What is wrong with narrative reviews?

Reviews are not new, so what is wrong with the classical or narrative review that has been used for many generations? Mulrow[11] was the first to examine the methodological quality of narrative reviews in general medicine. Since then a number of similar studies have examined the methods used in different branches of medicine, including cancer. The findings have been uniformly similar. Bramwell and Williams[24] reported on the methodological quality of reviews published in the *Journal of Clinical Oncology* from its inception in 1983 through to 1995. In the areas that are regarded as key to reducing the risk of bias (data identification, selection of data to be included, assessment of the validity of that data, quantitative synthesis of the data), less than 10% of the reviews used methods designed to reduce bias.

The outcome of this is that narrative reviews may often be unreliable. In the example above,[5] narrative reviews and textbook reviews failed to identify the true situation, as they were often selective in their use of the literature. In order to address this problem the concept of systematic reviews has been developed.

What are the main elements of a systematic review?

Systematic reviews aim to address the weaknesses identified in narrative reviews by paying careful attention to those areas where bias may be evident in the process of finding, selecting, extracting data from, and synthesising the results of trials asking similar questions. This essentially means writing a protocol setting out how the review is to be carried out in order to minimise bias. The key steps in preparing a systematic review are briefly discussed in the following sections. Users of systematic reviews should be looking to see if the reviewers have done a thorough job in each of these areas.

Locating and selecting studies

A comprehensive, unbiased search of the literature is one of the key differences between a systematic review and a narrative review. While electronic databases such as

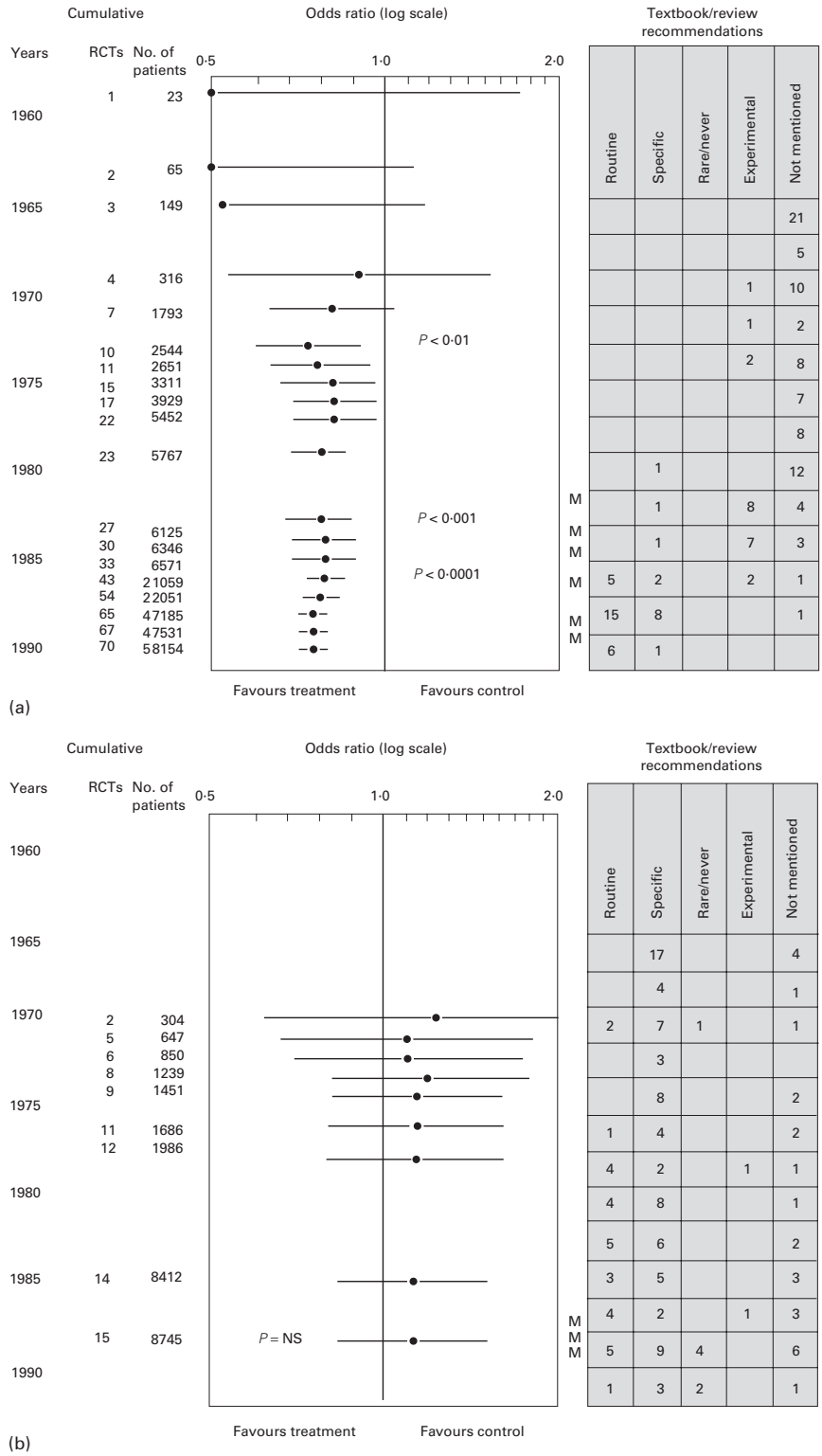

Figure 1.1 Cumulative meta-analysis of two treatments for acute myocardial infarction compared with treatment recommendations from reviews and textbooks. (a) This summarises the situation for thrombolytic therapy and (b) the use of lidocaine (lignocaine). The meta-analysis for thrombolytic therapy shows that there was good evidence for the use of this treatment from the early 1970s, but it was only in the late 1980s and early 1990s that it was recommended at all or used routinely. In contrast, there has never been any evidence to support the use of lidocaine (if anything, the evidence suggests that it may be harmful), but it was routinely recommended throughout this period. (With permission from *Oxford Textbook of Oncology*)

MEDLINE and Embase are powerful tools, they only include a subset of all biomedical journals. A study by the Cochrane Cancer Network compared the results of hand searching leading cancer journals with an optimal electronic search. The electronic search only found about 50% of the RCTs found by hand searching, even though the journals were in MEDLINE and Embase. If relevant records are in such databases it is still difficult to retrieve them easily. In addition, these databases do not contain the totality of published medical literature and, even if they did, a significant proportion of studies are never published,[25] and abstracts never turned into full peer review publications.[26] Failure to identify all of the available literature would not matter if this failure were a random event. However, there is good evidence that bias is acting and that there is a strong tendency for "positive" trials to be found and "negative" trials to be lost. As well as a bias regarding whether or not a report of a trial is published, there is good evidence that "positive" trials are published several years earlier than those with "negative" results (Box 1.2).[27]

Non-English-language references are underrepresented in MEDLINE and Embase and published articles only are included, so there is the potential for a review to be influenced by publication bias (which means that studies with positive results are selectively published) if one relies on studies identified using MEDLINE and Embase.[25,28–35] There is also some evidence to suggest that there is language bias, with bilingual researchers preferring to publishing "positive" results in English and "negative" results in their own language.[36–38]

In order to reduce the risk of bias it is important to use a variety of sources to identify studies and to have a systematic approach to selecting studies for inclusion in a review. The potential for reference bias (a tendency to preferentially cite studies supporting one's own views) is reduced by using multiple search strategies.[39,40] It should also be remembered that strongly "positive" trials are more likely to be published on multiple occasions, sometimes with different authors and different results.[41]

Box 1.2 Publication record of trials submitted to the Royal Prince Alfred Hospital Ethics Committee between 1979 and 1988, correlation with significance outcome

- 784 Eligible studies
- 520 (70%) Replied to study
- 218 Trials included tests of significance
- Those with positive outcomes were significantly more likely to have been published than negative results (HR, 2·32; 95% CI, 1·47–3.66; $P = 0·0003$)
- This result was even stronger for the 130 clinical trials (HR, 3·13; 95% CI, 1·76–5·58; $P = 0·0001$)
- Time to publication of the 218 trials was shorter for those with positive outcomes than those with negative results (median 4·8 *v* 8·0 years)
- The results for time to publication for the 130 clinical trials was similar (median 4·7 *v* 8·0 years)

Quality assessment of studies

Quality assessment of individual studies summarised in a systematic review is required to:

- limit bias in conducting the systematic review
- gain insight into potential comparisons
- guide interpretation of findings.

This quality assessment should look at those factors related to:

- applicability of the findings (also called external validity or generalisability). This is related to the definition of the key components of the question being addressed. Specifically, whether the findings of the trial are applicable to a particular population, intervention strategy and how the people, interventions and outcomes of interest were defined by these studies and the reviewers;
- validity of individual studies – interpretation of results is dependent upon the validity of the included studies, addressed in more detail in the following sections.

Validity

When a systematic review (or trial report) is being prepared or read, the validity of an individual study is the extent to which its design and conduct are likely to prevent systematic errors, or bias.[42] An issue that should not be confused with validity is precision.

Precision is a measure of the likelihood of chance effects leading to random errors. It is reflected in the confidence interval around the estimate of effect from each study and the weight given to the results of each study when an overall estimate of effect or weighted average is derived. Thus more precise results are given more weight.

Variation in validity can explain variation in the results of the studies included in a systematic review. More rigorous studies designed to avoid bias should be more likely to yield results that are closer to the "truth". Quantitative analysis of results from studies with varying degrees of validity can result in "false positive" conclusions if the less rigorous studies are biased toward overestimating treatment effectiveness. They can also come to "false negative" conclusions if less rigorous studies are biased towards underestimating an intervention's effect.[43]

It is important to critically appraise all studies in a review, even if there is no variability in either the validity or results

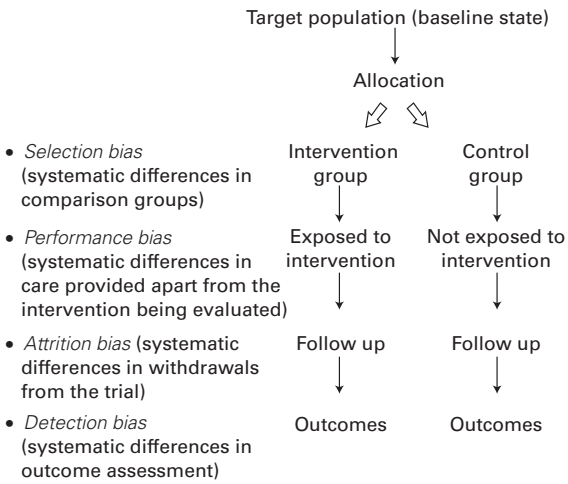

Figure 1.2 Sources of systematic bias

of the included studies. For instance, the results may be consistent among studies, but, if all the studies are flawed, the review's conclusions would not be as strong as in a series of rigorous studies yielding consistent results.

For readers of systematic reviews the key is to look to see if the reviewers made a systematic and prospective attempt to evaluate the validity of included trials.

Sources of bias in trials of healthcare interventions

There are four sources of systematic bias (Figure 1.2) in trials of health care.

- selection bias
- performance bias
- attrition bias
- detection bias.

Unfortunately, we do not have strong empirical evidence of a relationship between trial outcomes and the risk of these biases,[42,44] but there is a logical basis for suspecting such relationships and good reason to consider these potential biases when assessing studies for a review.[45]

Users of systematic reviews need to ask whether the reviewers have assessed the risk of each of these potential biases when preparing their review.

Selection bias

The way that comparison groups are assembled may lead to bias.[14]

- *Using an appropriate method to prevent foreknowledge of treatment assignment is crucially important in trial design*

When assessing a potential participant's eligibility for a trial, researchers and participants themselves should remain unaware of the next assignment in the sequence until after the decision about eligibility has been made. The ideal is for the process to be entirely independent of the individuals making the allocation. This is best achieved if assignment is by someone who is not responsible for recruiting subjects, such as someone based in a central trials office or pharmacy.

Concealing assignment should not be confused with "blinding" of patients, researchers, outcome assessors, and analysts. The reason for concealing the assignment schedule is to eliminate selection bias. In contrast, blinding (used after the allocation of the intervention) reduces performance and detection biases (see below).

Empirical research has shown that lack of adequate allocation concealment is clearly associated with bias.[46–48]

- *Concealment has been found to be more important in preventing bias than other components of allocation, such as the generation of the allocation sequence (for example, computer, random number table, alternation)*

The validity of studies can be judged on the method of allocation concealment. The method for assigning participants to interventions should be robust against patient and clinician bias and its description should be clear. The following approaches may be used to ensure adequate concealment schemes. Opaque numbered, sealed envelopes may be less secure that the other methods and a centralised method or pharmacy-controlled randomisation is always preferable (for example, allocation by a central office unaware of subject characteristics) using:

- prenumbered or coded identical containers administered serially to participants;
- on-site computer system combined with allocations kept in a locked unreadable computer file that can be accessed only after the characteristics of an enrolled participant have been entered.

Inadequate approaches to allocation concealment include:

- alternation
- the use of case record numbers (odd or even)
- date of birth or day of the week (odd or even)
- any procedure that is transparent before allocation, such as an open list of random numbers.

When studies do not report any concealment approach, adequacy should be considered unclear. An adequate description of the method of allocation and its concealment is frequently not reported.

Performance bias

Performance bias refers to systematic differences in the care provided to the participants in the comparison groups other than the intervention under investigation.

To protect against unintended differences in care and placebo effects, those providing and receiving care can be "blinded" so that they do not know the group to which participants have been allocated. Evidence suggests that such blinding is important in protecting against bias.[20,49,50] Studies have shown that contamination (provision of the intervention to the control group) and cointervention (provision of unintended additional care to either comparison group) can affect study results.[51,52]

There is evidence that participants who know their assignment report more symptoms, leading to biased results.[49] For these reasons, readers of trial reports and systematic reviews may want to consider the use of "blinding" as a criterion for validity. The key points are:

- Were the recipients of care unaware of their assigned intervention?
- Were those providing care unaware of the assigned intervention?
- Were those responsible for assessing outcomes unaware of the assigned intervention? This addresses detection bias (see below).

Attrition bias

Attrition bias refers to systematic differences in loss of participants between the comparison groups in the study. Because of inadequacies in reporting on loss of participants (for example, withdrawals, dropouts, protocol deviations), reviewers should be cautious about implicit accounts of follow up. The approach to handling losses has great potential for biasing the results and reporting inadequacies cloud this problem.

Detection bias

Detection bias refers to systematic differences between the comparison groups in outcome assessment. Trials that blind the people who will assess outcomes to the intervention allocation should logically be less likely to be biased than trials that do not. Blinding is likely to be particularly important in research with subjective outcome measures such as pain.[20,49,50] Despite this, at least two empirical studies have failed to demonstrate a relationship between blinding of outcome assessment and study results. This may be due to inadequacies in the reporting of studies.[53]

Bias from the selective reporting of results is different from bias in outcome assessment. This source of bias may be important in areas where multiple outcome measures are used.[54] Specification of predefined primary outcomes and analyses by the investigators can be useful indicators of validity.

Approaches to summarising the validity of studies

Because there is no "gold standard" for the validity of a trial, the possibility of validating any proposed scoring system or scale is limited.[55] While there are a number of scoring systems available, none can be recommended without reservation. They may carry a greater risk of confusing the issue and may not be transparent to readers. For these reasons, it is generally preferable to report how each trial scored on each criterion. Readers should assess whether a review has systematically gathered and reported information on the various aspects of validity discussed above.

Applying quality assessment criteria

It is preferable that there are multiple reviewers – this may limit bias, minimise errors, and improve reliability of findings. Reviewers should have complementary areas of expertise, such as medical content knowledge and review methodology experience.

Although experts in medical content may have preformed opinions that can bias their assessments,[6] they may also give more consistent assessments of the validity of studies than those without content expertise.[56] Content expertise is important in interpreting the subtleties of the clinical material.

Limitations of quality assessment

There are two major difficulties when assessing the validity of studies:

- The first is inadequate reporting of trials.[47,57,58,59] Because something was not reported, it does not mean that it was not done. Attempts to obtain additional data from investigators are sometimes necessary, but this may be difficult with no response from the original researchers.
- A second limitation (partly is a consequence of the first) is limited empirical evidence of a relationship between criteria thought to measure validity and actual study outcomes. While there is empirical evidence suggesting that inadequate concealment of allocation and lack of double blinding result in overestimates of the effects of treatment, research is needed to establish which criteria are key determinants of study results. Improved reporting of methods will facilitate such research.

Summarising effects across studies

An aim of a systematic review is to provide a reliable estimate of the effects of an intervention, based on a weighted average of the results of all the available relevant studies. Typically, the weight given to each study is the inverse of its variance, that is, more precise estimates (from larger studies with more events) are given more weight.[60] It is also possible to give studies more or less weight based on other factors such as their methodological quality, but this is rarely done.[43]

If it makes practical sense to combine the results of a group of studies and the observed differences between the results of the studies are not statistically significant (there is no statistical heterogeneity), it is relatively straightforward to combine the results. Each study is summarised using a measure of effect (such as an odds ratio, a relative risk, or a mean difference) that represents the within study comparison of the intervention and control groups. In this way participants in each study are only compared with other participants in the same study.

It is not the intention of this chapter to summarise current thinking on the methodology of systematic reviews. For those wishing to pursue this, *Systematic Reviews in Health Care: Meta-analysis in Context*,[61] is a useful starting point. Systematic reviews are relatively new and there remains much that is controversial or requiring further work. This includes statistical/methodological issues (such as how to estimate heterogeneity), as well as generic problems, (such as how to review data from non-randomised studies), studies of diagnostic techniques, and prognostic/predictive factors.

Currently methodology for non-randomised trial evidence is a major issue. There is little likelihood that sophisticated methods can make up for what may be deficiencies in the original research methods, but systematic reviews that carefully review the whole literature may help improve future research methods and may identify questions suitable for new research. Not all questions in medicine can be addressed by RCTs and there is a need for better ways of synthesising evidence from unrandomised studies.

Interpreting results

Although it can be argued that the results of a systematic review should stand on their own, many readers need help interpreting the results. Users of systematic reviews should look for consideration of the following points:

- the strength and reliability of the evidence
- the applicability of the results
- implication of costs and current practice
- clarification of any important trade-offs between the expected benefits and harms.

The primary purpose of a systematic review should be to present information, rather than to offer conclusions. Readers should look to the discussion and conclusions as an aid to understanding the implications of the evidence when making practical decisions.

Strength of evidence

This should start with a discussion of any important methodological problems in the included trials and the methods used in the review that might affect making practical decisions or future research.

It is often helpful to discuss how the included studies fit into the context of other evidence that is not included in the review. For example, for reviews of drug therapy it may be relevant to refer to dosage studies or non-randomised studies of the risk of adverse events – particularly those that are rare or delayed.

Because conclusions regarding the strength of inferences about the effectiveness of an intervention are essentially causal inferences, readers might want to consider guidelines for assessing the strength of a causal inference, such as those put forward by Hill.[62] In the context of a systematic review of clinical trials, these considerations might include:

- How good is the quality of the included trials?
- How large and significant are the observed effects?
- How consistent are the effects across trials?
- Is there a clear dose–response relationship?
- Is there indirect evidence that supports the inference?
- Have other plausible competing explanations of the observed effects (for example, bias or cointervention) been ruled out?

A variety of approaches to grading strength of evidence is available,[63–67] but none is universally appropriate for a wide range of reviews. Thus grading of evidence (as used in this book) can lack transparency, and should be interpreted with caution, but it may be useful in helping readers think about the reliability of the evidence.

Applicability

When interpreting evidence from RTCs or systematic reviews, users must decide how applicable the evidence is to their particular question. To do this, they must first decide whether the review provides valid information about potential benefits and harms that are important to them. They then need to decide whether the participants and settings in the included studies are reasonably similar to their own situation. In addition, it is important to consider the characteristics of the interventions and additional care provided during the research. Such consideration requires a

difficult extrapolation and Friedman has characterised this as: "A leap of faith is always required when applying any study findings to the population at large … In making that jump, one must always strike a balance between making justifiable broad generalisations and being too conservative in one's conclusions."[68]

Rather than rigidly applying the inclusion and exclusion criteria of the studies in particular clinical circumstances, it is generally better to ask whether there are compelling reasons why the evidence should not be used in those circumstances.[69] Such reasons, where difference from the original trials might limit applicability of results, include:

- biologic (for example, age, sex, genetic variability)
- cultural variation (local attitudes to disease and its treatment)
- variation in compliance with the therapy
- variation in baseline risk(for example, risk of recurrence in breast cancer).

Variation in the results of the included studies

As well as identifying limitations of the applicability of results, readers should look for important variation in results within the circumstances to which the results are applicable. Is there predictable variation in the relative effects of the intervention, and are there identifiable factors that may cause effects to vary? These might include:

- patient features, such as age, sex, biochemical markers
- intervention features, such as the timing or intensity of the intervention
- disease features, such as hormone receptor status.

Even in the absence of statistical heterogeneity, these features should be examined by testing whether there is an interaction with treatment and not by subgroup analysis. Differences between subgroups, particularly those that correspond to differences between studies, need to be interpreted cautiously. Chance variation between subgroups is inevitable, so unless there is strong evidence of an interaction then it should be assumed there is none.

Common errors in reaching conclusions

Common mistakes made in drawing conclusions include:

- confusing "no evidence of effect" with "evidence of no effect";
- describing a positive but statistically non-significant trend as "promising", whereas a "negative" effect of the same magnitude is not commonly described as a "warning sign";
- framing the conclusion in wishful terms, for example "the included studies were too small to detect a reduction in mortality" when the included studies showed a statistically non-significant increase in mortality. (One way of avoiding such errors is to consider the results "blinded"; that is, consider how the conclusions would be presented and framed if you reverse the direction of the results. If the confidence interval for the estimate of the difference in the effects of the interventions overlaps the null value, the analysis is compatible with both a true beneficial effect and a true harmful effect. If one of the possibilities is mentioned in the conclusion, the other possibility should be mentioned as well.);
- reaching conclusions that go beyond the evidence. (Often this is done implicitly, without referring to the additional information or judgements used in reaching the conclusions. Even when conclusions about the implications of a review for practice are supported by additional information and explicit judgements, the additional information that is considered is rarely systematically reviewed.).

Users of reviews need to be alert to the potential that the authors will have fallen into one of these traps.

Trade-offs

In addition to considering the strength of evidence underlying any conclusions that are drawn, reviewers should be as explicit as possible about any judgements about preferences (the values attached to different outcomes) that they make. Healthcare interventions generally entail costs and risks of harm, as well as expectations of benefit. Drawing conclusions about the practical usefulness of an intervention includes making trade-offs, either implicitly or explicitly, between the estimated benefits and the estimated costs and harms.[2] It is beyond the scope of most systematic reviews to incorporate formal economic analyses – although they might well be used for such analyses.[70,71] However, reviewers should consider all of the potentially important outcomes of an intervention when drawing conclusions, including ones for which there may be no reliable data from the included trials. They should also be cautious about any assumptions that they make about the relative value of the benefits, harms, and costs of an intervention.

Are all systematic reviews equal?

Systematic reviews, as is all clinical research, are subject to potential bias and poor methodology. The results of a systematic review should be interpreted with caution and