European Journal of Internal Medicine 32 (2016) 13–21

Contents lists available at ScienceDirect

European Journal of Internal Medicine

journal homepage: www.elsevier.com/locate/ejim

Review Article

Evidence-based clinical practice: Overview of threats to the validity of evidence and how to minimise them

Silvio Garattini [a], Janus C. Jakobsen [b,c], Jørn Wetterslev [b], Vittorio Bertelé [a], Rita Banzi [a], Ana Rath [d], Edmund A.M. Neugebauer [e], Martine Laville [f], Yvonne Masson [f], Virginie Hivert [f], Michaela Eikermann [g], Burc Aydin [h], Sandra Ngwabyt [d], Cecilia Martinho [i], Chiara Gerardi [a], Cezary A. Szmigielski [j], Jacques Demotes-Mainard [k], Christian Gluud [b,*]

[a] IRCCS Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy
[b] The Copenhagen Trial Unit, Centre for Clinical Intervention Research, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark
[c] Department of Cardiology, Holbæk Hospital, Holbæk, Denmark
[d] Orphanet, Institut National de la Santé et de la Recherche Médicale US14, Paris, France
[e] Faculty of Health, School of Medicine, Witten/Herdecke University, Campus Cologne, Germany
[f] Centre de Recherche en Nutrition Humaine, Rhône-Alpes, Univ de Lyon, Lyon, France
[g] Institute for Research in Operative Medicine, Faculty of Health, School of Medicine, Witten/Herdecke University, Cologne, Germany
[h] Department of Medical Pharmacology, School of Medicine, Dokuz Eylul University, Izmir, Turkey
[i] Palliative Care Service, Portuguese Institute of Oncology, Porto, Portugal
[j] Department of Internal Medicine, Hypertension and Vascular Diseases, Medical University of Warsaw, Warsaw, Poland
[k] ECRIN (European Clinical Research Infrastructure Network), Paris, France

ARTICLE INFO

Article history:
Received 15 March 2016
Accepted 22 March 2016
Available online 6 May 2016

Keywords:
Evidence-based medicine
Evidence-based clinical practice
Systematic review
Randomised clinical trial
Meta-analysis
Trial Sequential Analysis

ABSTRACT

Using the best quality of clinical research evidence is essential for choosing the right treatment for patients. How to identify the best research evidence is, however, difficult. In this narrative review we summarise these threats and describe how to minimise them. Pertinent literature was considered through literature searches combined with personal files. Treatments should generally not be chosen based only on evidence from observational studies or single randomised clinical trials. Systematic reviews with meta-analysis of all identifiable randomised clinical trials with Grading of Recommendations Assessment, Development and Evaluation (GRADE) assessment represent the highest level of evidence. Even though systematic reviews are trust worthier than other types of evidence, all levels of the evidence hierarchy are under threats from systematic errors (bias); design errors (abuse of surrogate outcomes, composite outcomes, etc.); and random errors (play of chance). Clinical research infrastructures may help in providing larger and better conducted trials. Trial Sequential Analysis may help in deciding when there is sufficient evidence in meta-analyses. If threats to the validity of clinical research are carefully considered and minimised, research results will be more valid and this will benefit patients and heath care systems.

1. Introduction

James Lind conducted his controlled clinical trial on interventions for scurvy in 1747 and since then evidence-based medicine has undergone a fascinating development [1–4]. Before 1900, only a few controlled clinical trials and randomised clinical trials (RCTs) were launched. During the last century, the conduct of RCTs increased importantly and meta-analyses were introduced [1–4].

Regarding medicinal products, an international consensus has been established allowing a phased assessment of intervention effects (Table 1). Certain fields like cardiology and oncology are fortunate to produce large numbers of RCTs [5]. Other fields like neurology, nephrology, endocrinology, hepatology, and surgery are less fortunate [5]. Medical devices, nutrition, and rare diseases are considered fields especially

* Corresponding author at: The Copenhagen Trial Unit, Centre for Clinical Intervention Research, Department 7812, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark. Tel.: +45 40 40 11 82.
E-mail addresses: silvio.garattini@marionegri.it (S. Garattini), jcj@ctu.dk (J.C. Jakobsen), wetterslev@ctu.dk (J. Wetterslev), vittorio.bertele@marionegri.it (V. Bertelé), rita.banzi@marionegri.it (R. Banzi), ana.rath@inserm.fr (A. Rath), Edmund.Neugebauer@uni-wh.de (E.A.M. Neugebauer), martine.laville@chu-lyon.fr (M. Laville), yvonne.masson@clermont.inra.fr (Y. Masson), virginie.hivert@inserm.fr (V. Hivert), michaela.eikermann@uni-wh.de (M. Eikermann), burcaydin@gmail.com (B. Aydin), sandra-nadia.ngwabyt-bikeye@inserm.fr (S. Ngwabyt), cvm@aibili.pt (C. Martinho), chiara.gerardi@marionegri.it (C. Gerardi), cezary.szmigielski@wum.edu.pl (C.A. Szmigielski), jacques.demotes@ecrin.org (J. Demotes-Mainard), cgluud@ctu.dk (C. Gluud).

http://dx.doi.org/10.1016/j.ejim.2016.03.020
0953-6205/

Table 1
The phases of clinical research regarding preventive or therapeutic medical interventions.

Phases	Participants and study designs for preventive or therapeutic interventions
Phase I	Healthy participants or patients – observational studies – randomised clinical trials designed to assess the safety (pharmacovigilance), tolerability, pharmacokinetics, and pharmacodynamics of an intervention.
Phase II	Patients with disease in question – randomised clinical trials. Phase II trials are performed on larger groups (up to about 300 patients) and are designed to continue safety assessments and to assess how well the intervention works.
Phase III	Patients with disease in question – randomised clinical trials often multicentre trials on large patient groups (300 to 10,000 or more depending upon the disease and outcome studied) aimed at being the definitive assessment of how effective the intervention is, in comparison with current 'gold standard' treatment.
Phase IV	Patients with disease in question – randomised clinical trials – observational studies. These studies and trials study the impact of applying the new intervention in clinical practice. This includes large randomised clinical trials, cluster randomised trials, and observational studies (clinical databases).

For medical devices slightly different phases are described [104].

in need of better clinical research [5,6]. The European Clinical Research Infrastructures Network (ECRIN)-Integrating Activity (IA) (http://www.ecrin.org/en/cooperative-projects/ecrin-integrating-activity-clinical-research-in-europe) has therefore identified barriers for good clinical research within these fields and assessed how these barriers could be broken down in order to improve their evidence-based clinical practice [7–10].

As an integral part of these activities, we provide an overview of the hierarchy of evidence regarding interventions and consider the threats to the validity of results of RCTs and systematic reviews with meta-analyses. The threats encompass risks of systematic errors ('bias'); design errors (erroneous selection of patients, doses of medication, comparators, analyses, outcomes, etc.); and risks of random errors (misleading results due to 'play of chance') [11–16]. We suggest possible solutions to the threats including establishment of national or transnational research infrastructures like ECRIN to improve clinical research and hereby reduce research waste [17–25].

2. Search strategy and selection criteria

Data for this review were identified by searches of PubMed and The Cochrane Library, references from relevant articles using the search terms "evidence based clinical practice", "evidence based medicine", "evidence hierarchy", "bias risks", "design errors", and "random errors", plus personal literature files. Articles were selected with a view that they should represent important didactic efforts to increase the medical profession's understanding of the central importance that evidence quality plays in underpinning clinical practice.

3. The hierarchy of evidence

Different experimental designs have different inferential powers, hence the hierarchy of evidence (Fig. 1) [13]. Provided the methodological quality of your study is good, the higher your study is in the hierarchy, the more likely you observe something close to the 'truth'. With better inferential powers, the higher the likelihood for improving patient outcomes when one translates the research findings into clinical practice (TRIP) [13]. All levels of the hierarchy may be threatened by systematic errors; design errors; and random errors [11,13,26].

3.1. Systematic reviews and meta-analyses

The Cochrane Collaboration coined the word 'systematic review' back in 1993, and developed The Cochrane Handbook for Systematic Reviews of Interventions (http://www.cochrane.org/training/cochrane-handbook) [11]. Systematic reviews are based upon peer-reviewed protocols and follow standardised methodologies [5,11,27]. Meta-analyses conducted without a protocol run the risk of systematic, design, and random errors, which may cloud our judgement on benefits and harms of interventions, and makes it difficult to design future trials validly [26,28–30].

3.2. Systematic reviews with meta-analysis of several small RCTs compared to a single, large RCT

A heated debate about which is superior — the results of a single large RCT or the results of a systematic review of all trials on a given intervention — has been on-going since meta-analyses became widely known in the 1980s. Some claim that evidence produced in a large RCT is much more valuable than results of systematic reviews or meta-analyses [31–33]. The trial advocates consider that systematic reviews should only be viewed as hypothesis-generating research [31–33].

Systematic reviews with meta-analyses cannot always be conducted with the same scientific cogency as a RCT with pre-defined high-quality methodology, addressing an a priori hypothesised intervention effect [11,30]. Systematic review authors will often know some of the RCTs before they have prepared their protocol for the systematic review, and hence, the review methodology will be at least partly data driven [11,30]. Understanding the inherent methodological limitations of systematic reviews with consideration and implementation of an improved review methodology already at the protocol stage can minimise this limitation [30]. Hence, a cornerstone of a high quality systematic review is the application of transparent, rigorous, and reproducible methodology [34].

IntHout and colleagues used simulations to evaluate error proportions in conventionally powered RCTs (80% or 90% power) compared to random-effects model meta-analyses of smaller trials (30% or 50% power) [35]. When a treatment was assumed to have no effect and heterogeneity was present, the errors for a single trial were increased more than 10-fold above the nominal rate, even for low heterogeneity [35]. Conversely, the error rates in meta-analyses were correct [35]. Evidence from a well-conducted systematic review of several RCTs with low risk of bias therefore represents a higher level of evidence compared to the results from a single RCT [11–14,29,30]. It also appears intuitively evident that inclusion of all available data from all RCTs with low risks of bias ever conducted, should be treated as a higher level of evidence compared to the data from one single RCT [13,30].

As a relatively new approach, network meta-analyses allow comparing interventions that have never been tested head to head in RCTs [36]. Careful consideration is needed for network meta-analyses to avoid false positive results [37]. Statistical and conceptual heterogeneity of the trials combined in a network meta-analysis should be assessed to avoid incoherence and thus chance findings [36]. Reporting bias can affect the findings of a network meta-analysis and lead to incorrect conclusions about the treatments compared [38]. Due to high number of pairwise comparisons in a network analysis, the risk of type I error should be controlled (see below). To address these methodological limitations in a systematic way, a clear protocol and a concise hypothesis are needed in advance to justify the meta-analytic approach [37,39].

In order to improve the systematic review methodology, recent PRISMA guidelines have been developed for individual participant data (IPD) systematic reviews with meta-analysis [40] and for network meta-analyses [39].

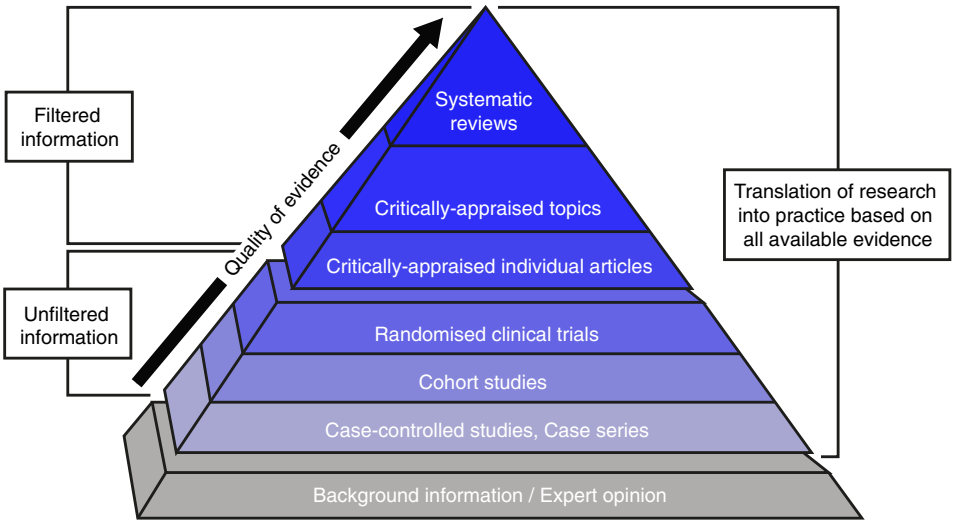

Fig. 1. The hierarchy of clinical evidence.

3.3. The results of RCTs compared to results of controlled cohort studies

Results of RCTs are generally of higher level compared to results of controlled cohort (non-randomised) studies [13,14,41]. Deeks and colleagues conducted simulations comparing results from RCTs to those of controlled cohort studies (Fig. 2) [14]. They concluded that results of controlled cohort studies often differ from results of RCTs [14]. Controlled cohort studies may still show misleading results even if the experimental and the control group appear similar in key prognostic factors. Standard methods of case-mix adjustment do not guarantee removal of undetected confounding, which may give rise to strong overestimation or underestimation of effects (Fig. 2). Residual confounding (that is any distortion that remains after controlling for confounding in the design and analysis of a study) may be high even when good prognostic data are available. Furthermore, results adjusted for baseline co-variates by logistic regression or propensity score may in some situations appear more biased than unadjusted results (Fig. 2) [14]. Other studies confirm that controlled cohort studies should not be used to validate intervention effects [13,41,42]. There are a number of real and perceived obstacles for conducting RCTs [7–10,13]. However, when new interventions are assessed we should always randomise the first patient [13,43,44]. In general, controlled cohort studies should rarely be used for assessing benefits (see GRADE below). If harmful effects are rare or appear only after long periods of time, then controlled cohort studies are needed as a supplement to RCTs to assess harmful effects [11]. Cohort studies should also be used for monitoring clinical quality and stability of treatment effects after new treatments have been introduced in clinical practice [45].

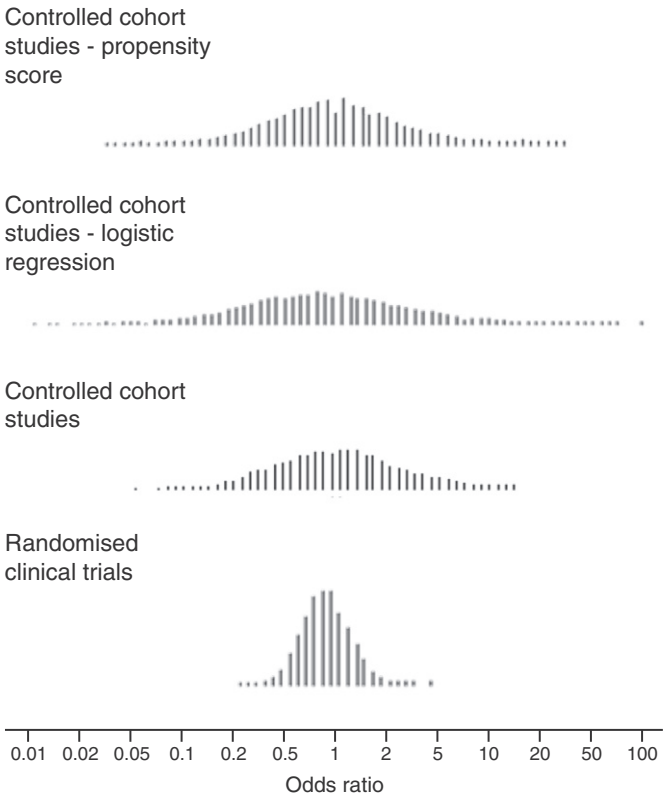

Fig. 2. Small randomised clinical trials and small controlled cohort studies sampled from a large randomised clinical trial in which the experimental intervention had no effect compared with placebo (odds ratio about 1.00) (after Deeks and colleagues 2003) [14].

4. The threats to internal validity

In the following, we focus on the threats to the internal validity of results of RCTs and systematic reviews of RCTs. Internal validity means the capability of a piece of research to provide a reliable answer to a relevant clinical question.

4.1. Threats caused by systematic errors ('bias')

Empirical evidence demonstrates that RCTs with high risks of bias lead to biased intervention effect estimates, i.e., overestimation of benefits and underestimation of harms (Table 2) [11,46–48].

Savovic et al. [46,48] combined data from seven meta-epidemiological studies and assessed how 'inadequate' or 'unclear' random sequence generation, allocation concealment, and blinding influenced intervention effect estimates, and whether these influences vary according to type of clinical area, intervention, comparison, and outcome. Outcomes were classified as 'mortality', 'other objective', or 'subjective'. Hierarchical Bayesian models were used to estimate the effect of trial characteristics on average bias (quantified as ratios of odds ratios (RORs) with 95% credible intervals (CrIs)). The analysis included 1973 trials from 234 meta-analyses. Intervention effect estimates were exaggerated by an average 11% in trials with inadequate or unclear sequence generation compared to adequate sequence generation (ROR 0.89, 95% CrI 0.82 to 0.96). Bias associated with inadequate or unclear

Table 2
Overview of domains that may bias results of randomised clinical trials and meta-analyses of such trials.

Domain	If the domain is not adequate, then bias mechanisms may lead to overestimation of beneficial intervention effects or underestimation of harmful intervention effects
Sequence generation	Systematic differences between entry characteristics of the comparison groups.
Allocation concealment	Systematic differences between entry characteristics of the comparison groups.
Blinding of participants and personnel	Systematic differences between groups in reporting of symptoms, in the provided care, or other factors that may affect the comparison groups.
Blinding of outcome assessment	Systematic differences between comparison groups in how outcomes are assessed.
Blinding of others involved (data managers, statisticians; conclusion makers; investigators)	Systematic differences between comparison groups in the handling of data, analyses, conclusions, or reporting.
Incomplete outcome data	Systematic differences between comparison groups regarding withdrawals.
Selective outcome reporting	Systematic differences between reported and unreported findings.
Vested interests (academic; industry)	Systematic conscious or unconscious manipulation of analyses, other factors, and spin.
Other domains	In each and every trial, one needs to assess other mechanisms and design factors that may bias the results (see Table 3).

Table 3
Overview of design components that may bias results of randomised clinical trials and meta-analyses of such trials.

Design components	Mechanism that leads to overestimation of beneficial intervention effects or underestimation of harmful intervention effects
Centres	If interventions are assessed in tertiary sector, the treatment effects may not be relevant in the primary or secondary sector.
Participants	If interventions are assessed in very diseased participants, the treatment effects may not be relevant for less diseased patients.
Experimental intervention	If too high or too low dosage of the experimental intervention is used, both benefits and harms may be wrongly assessed.
Control intervention	If too high or too low dosage of the control intervention is used, both benefits and harms may be wrongly assessed.
Selection of outcomes	Use of wrong outcomes (non-validated surrogates; composite outcomes, patient irrelevant outcomes).
Goal — explanatory or pragmatic	Explanatory trials will only rarely lead to treatments that can be implemented in clinical practice. Large, well-conducted pragmatic trials with long follow up needs to be conducted before implementation of prognostic, diagnostic, or therapeutic interventions in clinical practice.
Objective — superiority, equivalence, non-inferiority	See text.

sequence generation was greatest for subjective outcomes (ROR 0.83, 95% CrI 0.74 to 0.94). The effect of inadequate or unclear allocation concealment compared to adequate allocation concealment was greatest among meta-analyses with a subjectively assessed outcome intervention effect (ROR 0.85, 95% CrI 0.75 to 0.95). Lack of, or unclear, blinding compared to double blinding was associated with an average 13% exaggeration of intervention effects (ROR 0.87, 95% CrI 0.79 to 0.96). Among meta-analyses with subjectively assessed outcomes, lack of blinding appeared to cause more biased results than inadequate or unclear sequence generation or allocation concealment.

In a similar way, trials with incomplete outcome data may produce biased results, if proper intention-to-treat analyses are not conducted and valid methods are not used to handle missing data [49,50]. Chan et al. have revealed how authors of RCTs make selective outcome reporting, leading to a gross overestimation of treatment benefits [51–53]. There is, therefore, an urgent need to register all trial protocols prior to inclusion of the first participant and to publish detailed statistical analysis plans before trial data are collected [54,55].

The systematic review by Lundh and colleagues clearly demonstrated that industry involvement is associated with biased results [56]. Such bias was not explained by other bias domains [56].

In conclusion, bias associated with specific reported trial design characteristics leads to exaggeration of beneficial intervention effect estimates. For each of the domains assessed above, these effects were greatest for subjectively assessed outcomes. The average magnitude of overestimation of 10% to 20% is larger than most 'true' intervention effects.

4.2. Threats caused by design errors

A number of design errors may also lead to overestimation of benefits or underestimation of harms (Table 3). We present such threats in the following paragraphs.

4.2.1. Abuse of surrogate outcomes

Surrogate outcomes with questionable clinical relevance are frequently used instead of patient-centred outcomes. Examples of surrogate outcomes are blood cholesterol, blood glucose, sustained virological response, and blood pressure. Examples of important patient-centred outcomes are myocardial infarction, stroke, and death [16,57]. In several cases, drugs have been implemented based only on surrogate results even when similar drugs existed with proof of benefits on patient-centred outcomes [16]. Several drugs are advertised based on surrogate outcomes even though they have no effect or detrimental effects on patient-centred outcomes [13,16,29,30].

RCTs ought to assess if an intervention is safe and effective. Surrogate outcomes may neither be meaningful for patients nor sufficient evidence for implementing an intervention into clinical practice. If the findings of pragmatic RCTs are to benefit health-care decision-making, then careful selection of appropriate outcomes is crucial to the design of RCTs. These issues could be addressed with the development and application of agreed sets of outcomes, known as core outcome sets [58].

The COMET (Core Outcome Measures in Effectiveness Trials) Initiative (http://www.comet-initiative.org) brings together people interested in the development and application of core outcome sets. The objective of COMET is to design core outcome sets for each specific condition, which represent the minimum that should be measured and reported in all studies, trials, and systematic reviews. This would allow the results of trials and other studies to be compared, contrasted and combined as appropriate, thus ensuring that all trials contribute with usable information. This does not imply that outcomes in a particular study should be restricted to those in the core outcomes, and researchers would still continue to assess other relevant outcomes as well as those in the core outcome sets.

The development and application of core outcome sets would make research more likely to measure and report appropriate patient-centred outcomes [57]. A large proportion of RCTs fails to include all outcomes that patients, clinicians, and decision makers need when deciding if an intervention should be used or not [59,60]. Despite increasing recognition of the importance of incorporating patients' opinions in the development of core outcome sets, the patients involvement has been limited.

4.2.2. Abuse of composite outcomes

To reduce the required sample size, RCTs often adopt composite outcomes [29,61,62]. However, composite outcomes make it difficult to interpret the clinical significance of the results [29,61]. Any benefit on a composite outcome may be presumed to relate to all its components [61], but evidence shows that intervention effects on composite

outcomes often apply to a single component, most likely the less relevant [61,62]. Moreover, proper statistical analyses of composite outcomes require an analysis of each single outcome in the composite outcome which creates problems with multiplicity and each single component will often not have sufficient power to confirm or refute the anticipated intervention effect [29,63]. Composite outcomes may be used, but only if results on their single components are reported so the clinical implications of the results can be thoroughly interpreted [29]. Patient-centred single outcomes (e.g., all-cause mortality) should always be preferred to composite outcomes if power is sufficient using the single outcome.

4.2.3. Abuse of non-inferiority trials

Non-inferiority trials are designed to establish whether a new intervention is not worse than a standard treatment. Non-inferiority trials frequently accept an intervention as being, e.g., 20% inferior compared with the standard treatment [64]. If the new intervention is inferior to the standard treatment but within a given limit, it is then considered non-inferior, even though possibly worse. So conceived, this trial design is not ethical because RCTs should generally be designed to test superiority of an intervention, not just its non-inferiority [64]. Non-inferiority trials often allow substantial harm to patients.

4.2.4. Abuse of poor reporting or no reporting

Trials with significant results are more likely to be published than those with neutral or negative results [55]. Random errors ('play of chance') cause especially small trials to indicate both benefit and harm, when there is none [35]. Therefore, publication bias will increase the risk of erroneous conclusions about intervention effects [65]. The AllTrials initiative is campaigning for the publication of the results from all past, present, and future clinical trials [66]. This initiative is of utmost importance but the reporting of each single trial must also be thorough and valid [20]. Studies have shown that the poor description of trial interventions resulted in 40% to 89% of trials being non-replicable [20]. Comparisons of protocols with publications showed that most trials had at least one primary outcome changed, introduced, or omitted; and investigators of new trials rarely set their findings in the context of a systematic review [20]. Reporting guidelines such as CONSORT and PRISMA aim to improve the quality of research reports, but these guidelines should be followed much more thoroughly [67]. Adequate reports of research should clearly describe which questions were addressed and why, what was done, what was shown, and what the clinical implications of the findings were [20]. The Nordic Trial Alliance has called for full transparency of all clinical research [55], and the WHO has also called for public disclosure of all trials (http://www.who.int/ictrp/results/WHO_Statement_results_reporting_clinical_trials.pdf).

4.2.5. Additional threats caused by design errors

A number of additional design errors need to be considered which may affect either the internal validity of the RCT results or their external validity, meaning their actual clinical implication in the every-day clinical practice. In this respect the following should be taken into consideration: (1) is the dose, form, length, etc. of both the experimental and control intervention adequate?; (2) is the trial population similar to a clinical population so trial results can apply to it?; (3) is the trial designed as a pragmatic trial so the effects of the trial interventions can be reproduced in a clinical setting?; and (4) was the initial research question valid [68–71]? We have summarised the threats caused by design errors in Table 3.

Regarding the impact of design errors on the external validity of RCT, one further issue should be considered. According to the EU directive on drugs, new drugs should be approved on the basis of "quality, efficacy and safety" [72]. This gives the industry the possibility to avoid head to head comparisons between two treatment options. This in turn will potentially make it difficult to know which intervention is most effective in a given clinical condition. 'Efficacy' is an ambiguous word. In the best interest of patients the legislation should rather request 'comparative therapeutic value to the patient' [73]. Furthermore, regulatory agencies do not take into consideration studies that are not presented by industry and most regulatory authorities have not yet started to require systematic reviews assessing benefits and harms. In order to avoid obvious conflicts of interests, RCTs and systematic reviews should be conducted by independent non-profit organisations and results should not be handled by ghost authors [74–76].

4.3. Risk of random errors 'play of chance'

Both SPIRIT and CONSORT endorse that any result of a RCT ought to be related to a sample size [77,78]. The inclusion of an adequate number of participants in RCTs aims at avoiding two possible drawbacks, i.e., to let the RCT show an effect that does not actually exists (type I error) or not show an effect that exists (type II error). The estimation of the sample size in a RCT requires four components: a maximally allowed risk of type I error (α) and type II error (β); an anticipated intervention effect (μ) on the primary outcome; and the variance of the primary outcome (ϑ) [29]. Given these four components, the formula provides an estimate of the sample size (N) needed to detect or reject the anticipated intervention effect (μ) with the chosen error risks in trials with equal group size $N = 4 \cdot \frac{(Z_{1-\alpha/2}+Z_{1-\beta})\cdot\vartheta}{\mu^2}$ where $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are the corresponding $(1-\alpha/2)$ and $(1-\beta)$ fractiles from the normal distribution [79].

4.3.1. Interim-analysis in a single RCT

If the primary outcome in a RCT is planned to be evaluated before the estimated total sample size has been reached there is international consensus for employing a data monitoring and safety committee (DMSC) [29,80]. The DMSC should recommend stopping for benefit only when the P-value is less than an adjusted threshold for statistical significance related to the acquired number of randomised participants. The thresholds for significance should be adjusted according to the fraction that the accrued number of participants constitutes of the required sample size, i.e., the P-value has to reach a value lower than the α used in the sample size calculation (usually 0.05) [29,80]. The reasons for more restrictive stopping thresholds at an interim-analysis are dual: testing on sparse data adds uncertainty to the actual estimate of the intervention effect (due to the larger risk of having unequal distribution of prognostic factors in smaller samples), and repetitive testing on accumulating data requires adjustment for 'longitudinal' multiplicity [29,81,82]. Before the fixed sample size has been reached, more strict thresholds for significance (e.g., about 99.5% confidence intervals when half of the sample size has been reached and 99% confidence intervals when three-quarters of the sample size has been reached) have to be used to assess whether the thresholds for significance have been crossed or not [29,30]. The procedure for interim-analysis of a RCT is called group sequential analysis. Often the O'Brien-Fleming α-spending function is chosen [83–85]. If the cumulative z-score breaks a group sequential monitoring boundary it is reliable to trust the results even though the planned sample size is not reached (Fig. 3). The methodology has been further developed by Lan-DeMets monitoring boundaries allowing one to test whenever wanted [83–85].

4.3.2. Required information size in a meta-analysis of RCTs

Contrary to RCTs [29], risks of random errors in systematic reviews have received relatively limited attention [30]. Most of the RCTs in Cochrane systematic reviews are underpowered to detect even large intervention effects [86]. Almost 80% of the meta-analyses in Cochrane reviews are underpowered to detect or reject a 30% relative risk reduction taking the observed between trial variance in a random-effects meta-analysis into consideration [86]. Therefore, most meta-analyses may be considered as interim-analyses of intervention effects on the way to the required information size (RIS) [85,87,88].

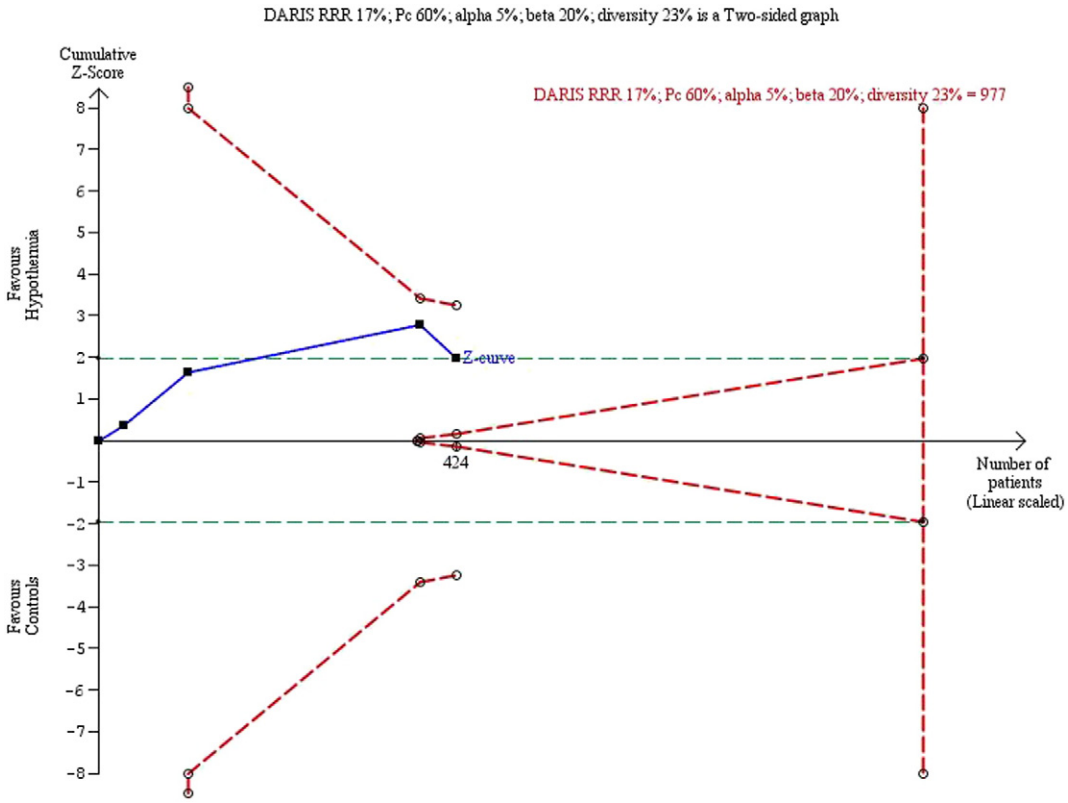

Fig. 3. Trial Sequential Analysis of a meta-analysis including four randomised clinical trials. The Z-value is the test statistic and $|Z| = 1.96$ corresponds to a $P = 0.05$, the higher the Z-value the lower the *P*-value. The Trial Sequential Analysis assesses all-cause mortality after out of hospital cardiac arrest randomising patients to cooling to 33 °C versus no temperature control in the four trials. The required information size, to detect or reject a 17% relative risk reduction found in the random-effects meta-analysis, is calculated to 977 participants using the diversity found in the meta-analysis of 23%, with a double-sided α of 0.05, a power of 80%, and based on a proportion of patients with the outcome of 60% in the control group (Pc). The cumulative Z-curve (blue full line with quadratic indications of each trial) surpasses the traditional boundary for statistical significance during the third trial and touches the traditional boundary after the fourth trial (95% confidence interval 0.70–1.00; $P = 0.05$). However, none of the trial sequential monitoring boundaries for benefits or harms (etched red curves above and below the traditional horizontal lines for statistical significance) or for futility (etched red wedge) has been surpassed. The result is therefore inconclusive when adjusted for sequential testing on an accumulating number of participants and the fact that the required information size has not yet been achieved. The TSA adjusted confidence interval is 0.63–1.12 after inclusion of the fourth trial.

We used simulations to assess the risk of overestimating an intervention effect to >20% or >30% relative risk reduction, when there was in fact no intervention effect, and found it to be considerably greater than 5% if the RIS (meta-analytic sample size) is not reached [15]. First when the number of outcomes was above 200 and the cumulated sample was above 2000, assuming moderate heterogeneity, the risk of overestimation declined towards 5% [15]. Our study also showed that surpassing a RIS to detect or reject a realistic intervention effect of 20% in a random-effects meta-analysis reduced the risk of overestimating the intervention effect (by 20% and 30%) to the nominal 2.5% [15]. Estimating a RIS therefore seems crucial for the interpretation of the statistical significance of results of meta-analyses [30,85,87–89].

4.3.3. Trial Sequential Analysis

As most cumulative meta-analyses may be regarded as interim-analyses in the process of reaching a RIS they should be analysed as such using sequential meta-analysis methodology [85,87,90]. Trial Sequential Analysis is a sequential meta-analysis methodology of cumulative meta-analysis using Lan-DeMets monitoring boundaries [91]. Lan-DeMets monitoring boundaries offer the possibility to demonstrate if adjusted statistical thresholds for benefit, harm, or futility are crossed [84,85,92].

To assess a meta-analysis transparently, a pre-planned sequential meta-analysis with a priori chosen anticipated intervention effect μ, α, β, and a model based variance of the outcome should be part of any protocol for a systematic review [11]. Trial Sequential Analysis offers such a transparent platform and a programme with a manual is available for free at: http://www.ctu.dk/tsa [84].

4.4. Other threats to the validity of systematic reviews and meta-analyses

Different types of biases hamper the conduct and interpretation of systematic reviews [87,93]. Selective reporting of completed studies leads to publication bias because positive trials with impressive findings are more likely to be published [11]. The simplest method to detect possible publication bias is visual inspection of a funnel plot. Other methods might contribute, including Egger test, [94] Begg-Mazumdar test [95], and 'trim-and-fill' method [96]. In 2000, Sutton et al. found that about half of the Cochrane meta-analyses may be subject to some level of publication bias and about 20% had a strong indication of missing trials [97]. The authors concluded that around 5% to 10% of meta-analyses might be interpreted incorrectly because of publication bias. Only few reviews report assessment of publication bias [98]. There are a number of problems when publication bias is assessed with the available methods, e.g., asymmetry of the funnel plot might be due to other factors than publication bias [11], any type of bias might cause funnel plot asymmetry, and lack of symmetry may be due to lack of power. Outcome reporting bias within individual trials (see 'Abuse of poor or no reporting') is another type of bias important to be considered when conducting a systematic review [53]. Also selective reporting of other studies occurs often [99].

The PRISMA authors listed 27 items to be included when reporting a systematic review or meta-analysis [54]. It includes assessment of the risk of bias in individual trials and across trials. Reporting an assessment of possible publication bias was stated as a marker of the thoroughness

of the conduction of the systematic review, and accordingly, failure to report the assessment of the risk of bias in included trials can be seen as a marker of lower quality of conduct [100].

Systematic reviews should primarily base their conclusions on results of trials with low risk of bias and not mix trials at low risk of bias with trials at high risk of bias [101].

5. Grading the quality of evidence (GRADE)

Judgments about the quality of evidence and recommendations of interventions in healthcare are complex. The hierarchy of evidence is a good framework for evaluating the effects of interventions. Sometimes, however, you need to downgrade or upgrade the inferential powers of a piece of research. If a systematic review includes several trials with high risk of bias and random errors, then the inferential power of the systematic review needs to be downgraded. If a cohort study is well conducted and shows an extraordinary large intervention effect (e.g., insulin for diabetic coma or drainage of an abscess), then that evidence may be upgraded. However, such extraordinary effective interventions are very rare in clinical practice — and can never be identified in advance [13]. As clinical research is a forward moving process, it is important that the most valid research design is chosen from the very beginning — the RCT [11,13,43].

A systematic and explicit approach may prevent wrong recommendations. During the 2000s, a working group developed Grading of Recommendations Assessment, Development and Evaluation (GRADE; http://www.gradeworkinggroup.org/index.htm) [102]. Recommendations to administer or not administer an intervention should be based on the trade-offs between benefits on the one hand, and harms, burdens, and costs on the other [102]. If benefits outweigh harms, experts will recommend that clinicians offer a treatment to a given specified patient group [102]. After going through the process of grading evidence, the overall quality will be categorised as high, moderate, low, or very low [102]. The uncertainty associated with the trade-off between the benefits and harms will determine the strength of recommendations. GRADE has only two levels, strong and weak recommendations:

- Review authors will make a strong recommendation if they are very certain that benefits do, or do not, outweigh harms.
- Review authors should only offer weak recommendations if they believe that benefits and harms are closely balanced, or appreciable uncertainty exists about their magnitude.

In addition, the importance of patient values and preferences in clinical decision making should also be considered (see 'Abuse of surrogate outcomes'). When fully informed patients are liable to make different choices, guideline panels should offer weak recommendations.

The hierarchy of evidence will apply to the vast majority of interventions. However, we have in the past witnessed some interventions with dramatic effects. When such interventions are at hand, lower levels of the hierarchy may be used for proving benefits. The problem is, however, that we have hardship in predicting when we have such an intervention at hand [13].

When developing new interventions, investigators and industry are wise in conducting their research in different phases in which the scientific evaluation of the benefits and harms is adjusted to the level of knowledge obtained (Table 1). The different research designs used in the different phases will depend on the intervention one wants to examine [11,13]. Such designs ought always to be based on up-to-date systematic reviews of the available evidence [11,29,30,43,54,69–71,77,78].

6. Discussion

Clinical research has undergone a dramatic development since James Lind, but due to many threats to the validity of RCTs and other studies, this development has to continue [20,26,87]. The threats to validity and the associated waste of clinical research affect all interventions and all diseases [17–22]. However, the threats are especially daunting in fields with less accumulated experience in conducting RCTs and more difficulties in identifying rare patients. As patients' lives depend on properly conducted RCTs as well as valid assessments of such RCTs, improvements of the methodology are urgently needed. We have reviewed the threats to internal and external validity and mentioned a number of ways in which these threats can be prevented or minimised. Our mention of amendments is not exhaustive. Other improvements of methodology that needs mention are the Human Genome Epidemiology Network (HuGENet) and the EQUATOR Network (Enhancing the Quality and Transparency of Health Research, http://www.equator-network.org).

7. Conclusions

We have in this paper considered threats to the validity of evidence in general. We feel that the chance to introduce the required amendments into all fields of medicine would be greatly enhanced by forming national and regional infrastructures that could support clinical research [75,103]. We will in four connected papers discuss the common and special bottlenecks for conducting clinical research on medical devices, nutrition, and rare diseases [7–10]. Through identifying the threats to internal validity and through providing solutions for these problems, it is our hope that more and better quality clinical research may be achieved and used.

Contributors

CG coordinated the project. CG, JCJ, JW, SG, and VB wrote the first drafts. JDM coordinated the application for the EU. All authors provided feedback on subsequent drafts of the paper.

Role of funding sources

The ECRIN-IA grant from the EU FP7 (GA 284395) provided support for meetings and for the conduct of this review. The Mario Negri Institute housed the ECRIN-IA meeting in February, 2013. The funding sources had no influence on data collection, design, analysis, interpretation; or any aspect pertinent to the study.

Conflicts of interests

All authors are involved in conducting randomised clinical trials and are members of ECRIN. CG and JW are members of the Copenhagen trial unit's task force to develop theory and software for doing Trial Sequential Analysis. No additional conflicts are known.

Acknowledgements

The ECRIN-IA grant from the EU FP7 is thanked for support for meetings and for the conduct of this review.

The Mario Negri Institute is thanked for housing the ECRIN-IA meeting in February, 2013.

All participants of ECRIN-IA are thanked for participating in discussions identifying the bottlenecks of clinical research and the threats to internal validity of clinical research (and hence threats to external validity of clinical research) suggesting ways to blow up the bottlenecks and erase the threats.

References

[1] The Library and Information Services Department. The Royal College of Physicians of Edinburgh: James Lind Library. [Available online at] http://wwwjameslindlibraryorg/; 2003.

[2] Opinel A, Trohler U, Gluud C, Gachelin G, Smith GD, Podolsky SH, et al. Commentary: the evolution of methods to assess the effects of treatments, illustrated by the development of treatments for diphtheria, 1825–1918. Int J Epidemiol 2013; 42(3):662–76.
[3] Heiberg P. Studier over den statistiske undersøgelsesmetode som hjælpemiddel ved terapeutiske undersøgelser [Studies on the statistical study design as an aid in therapeutic trials] (http://www.jameslindlibrary.org/illustrating/records/studier-over-den-statistiske-undersogelsesmetode-som-hjaelpemid/key_passages). Bibl Laeger 1897;89:1–40.
[4] Gluud C, Nikolova D. Likely country of origin in publications on randomised controlled trials and controlled clinical trials during the last 60 years. Trials 2007;8:7.
[5] The Cochrane Collaboration. The Cochrane Library. [Avaiable at] http://wwwthecochranelibrarycom/view/0/indexhtml; 2014.
[6] ECRIN-IA. Description of the project. [Available from] http://www.ecrin.org/index.php?id=141.
[7] Durisic S, Garattini S, Rath A, Neugebauer EAM, Laville M, Jakobsen JC, Kubiac C, DeMotes-Mainard J, Gluud C: Common barriers to the conduct of randomised clinical trials - the Europeean Clinical Research Infrastructure (ECRIN) perspective Trials, 2016 [to be submitted].
[8] Rath A, Ngwabyt S, Durisic S, Garattini S, Neugebauer EAM, Laville M, Jakobsen JC, Kubiac C, DeMotes-Mainard J, Gluud C: Specific barriers to the conduct of randomised clinical trials within rare diseases — the Europeean Clinical Research Infrastructure (ECRIN) perspective Trials, 2016 [to be submitted].
[9] Neugebauer EAM, Rath A, Durisic S, Garattini S, Laville M, Jakobsen JC, Kubiac C, DeMotes-Mainard J, Gluud C: Specific barriers to the conduct of randomised clinical trials on medical devices — the Europeean Clinical Research Infrastructure (ECRIN) perspective Trials, 2016 [to be submitted].
[10] Laville M, Neugebauer EAM, Rath A, Durisic S, Garattini S, Jakobsen JC, Kubiac C, DeMotes-Mainard J, Gluud C: Specific barriers to the conduct of randomised clinical trials on nutrition - the Europeean Clinical Research Infrastructure (ECRIN) perspective Trials, 2016 [to be submitted].
[11] Higgins JPT, Green S. The Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Cochrane Collaboration. 2011 [Available from] http://www.cochrane-handbook.org/.
[12] Keus F, Wetterslev J, Gluud C, van Laarhoven CJ. Evidence at a glance: error matrix approach for overviewing available evidence. BMC Med Res Methodol 2010;10:90.
[13] Jakobsen JC, Gluud C. The necessity of randomized clinical trials. Br J Med Res 2013; 3(4):1453–68.
[14] Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27):1–173 [iii-x].
[15] Thorlund K, Imberger G, Walsh M, Chu R, Gluud C, Wetterslev J, et al. The number of patients and events required to limit the risk of overestimation of intervention effects in meta-analysis-a simulation study. PLoS One 2011;6, e25491.
[16] Gluud C, Brok J, Gong Y, Koretz RL. Hepatology may have problems with putative surrogate outcome measures. J Hepatol 2007;46(4):734–42.
[17] Al-Shahi Salman R, Beller E, Kagan J, Hemminki E, Phillips RS, Savulescu J, et al. Increasing value and reducing waste in biomedical research regulation and management. Lancet 2014;383(9912):176–85.
[18] Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gulmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet 2014;383(9912):156–65.
[19] Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet 2009;374(9683):86–9.
[20] Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014; 383(9913):267–76.
[21] Macleod MR, Michie S, Roberts I, Dirnagl U, Chalmers I, Ioannidis JP, et al. Biomedical research: increasing value, reducing waste. Lancet 2014;383(9912):101–4.
[22] Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? Lancet 2016;387: 1573–86.
[23] Ioannidis JP. How to make more published research true. PLoS Med 2014;11(10), e1001747.
[24] Ioannidis JP. Clinical trials: what a waste. BMJ 2014;349:g7089.
[25] Ioannidis JP, Greenland S, Hlatky MA, Khoury MJ, Macleod MR, Moher D, et al. Increasing value and reducing waste in research design, conduct, and analysis. Lancet 2014;383(9912):166–75.
[26] Ioannidis JPA. Why most published research findings are false. PLoS Med 2005; 2(8), e124.
[27] Sackett DL. How to read clinical journals. Can Med Assoc J 1982;126(12):1373.
[28] Jørgensen AW, Hilden J, Gøtzsche P. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs. Syst Rev 2006;333(782).
[29] Jakobsen JC, Gluud C, Winkel P, Lange T, Wetterslev J. The thresholds for statistical and clinical significance — a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC Med Res Methodol 2014;14(34).
[30] Jakobsen JC, Wetterslev J, Winkel P, Lange T, Gluud C. Thresholds for statistical and clinical significance in systematic reviews with meta-analytic methods. BMC Med Res Methodol 2014;14(1):120.
[31] Borzak S, Ridker PM. Discordance between meta-analyses and large-scale randomized, controlled trials. Examples from the management of acute myocardial infarction. Ann Intern Med 1995;123(11):873–7.
[32] Hennekens CH, DeMets D. The need for large-scale randomized evidence without undue emphasis on small trials, meta-analyses, or subgroup analyses. JAMA 2009;302(21):2361–2.
[33] Stegenga J. Is meta-analysis the platinum standard of evidence? Stud Hist Philos Biol Biomed Sci 2011;42(4):497–507.
[34] Thayer KA, Wolfe MS, Rooney AA, Boyles AL, Bucher JR, Birnbaum LS. Intersection of systematic review methodology with the NIH reproducibility initiative. Environ Health Perspect 2014;122(7):A176–7.
[35] Inthout J, Ioannidis JP, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Methods Med Res 2016;25:538–52.
[36] Mills EJ, Thorlund K, Ioannidis JP. Demystifying trial networks and network meta-analysis. BMJ 2013;346:f2914.
[37] Del Re AC, Spielmans GI, Flückiger C, Wambold BE. Efficacy of new generation antidepressants: differences seem illusory. PLoS One 2013;8(6), e63509.
[38] Trinquart L, Abbe A, Ravaud P. Impact of reporting bias in network meta-analysis of antidepressant placebo-controlled trials. PLoS One 2012;7(4), e35219.
[39] Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med 2015;162(11):777–84.
[40] Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, Stewart G, et al. Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. JAMA 2015;313(16):1657–65.
[41] Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA 2001;286(7):821–30.
[42] Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ 2016;352:i493.
[43] Chalmers TC. Randomize the first patient. N Engl J of Med 1977;296(2):107.
[44] Ioannidis JP. Are mortality differences detected by administrative data reliable and actionable? JAMA 2013;309(13):1410–1.
[45] Winkel P, Zhang NF. Statistical Development of Quality in Medicine. Wiley; 2007.
[46] Savović J, Jones H, Altman D, Harris R, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized controlled trials: combined analysis of meta-epidemiologic studies. Health Technol Assess 2012;16(35):1–82.
[47] Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ 2008;336(7644):601–5.
[48] Savovic J, Jones HE, Altman DG, Harris RJ, Juni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med 2012;157(6):429–38.
[49] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med 2012;367(14): 1355–60.
[50] Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338.
[51] Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One 2008;3(8), e3081.
[52] Dwan K, Gamble C, Williamson PR, Kirkham JJ. Reporting bias group: systematic review of the empirical evidence of study publication bias and outcome reporting bias — an updated review. PLoS One 2013;8(7), e66844.
[53] Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 2004;291(20):2457–65.
[54] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 2009; 339:b2700.
[55] Skoog M, Saarimäki JM, Gluud C, Sheinin M, Erlendsson K, Aamdal S, et al. Transaprency and registration in clinical research in the Nordic countries. Nordic Trial Alliance, NordForsk; 2015 1–108.
[56] Lundh A, Sismondo S, Lexchin J, Busuioc OA, Bero L. Industry sponsorship and research outcome. Coch Database Syst Rev 2012;12, MR000033.
[57] Williamson PR, Altman DG, Blazeby JM, Clarke M, Devane D, Gargon E, et al. Developing core outcome sets for clinical trials: issues to consider. Trials 2012;13:132.
[58] Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. Driving up the quality and relevance of research through the use of agreed core outcomes. J Health Serv Res Policy 2012;17(1):1–2.
[59] Serrano-Aguilar P, Trujillo-Martin MM, Ramos-Goni JM, Mahtani-Chugani V, Perestelo-Perez L, Posada-de la Paz M. Patient involvement in health research: a contribution to a systematic review on the effectiveness of treatments for degenerative ataxias. Soc Sci Med 2009;69(6):920–5.
[60] Mease PJ, Arnold LM, Crofford LJ, Williams DA, Russell IJ, Humphrey L, et al. Identifying the clinical domains of fibromyalgia: contributions from clinician and patient Delphi exercises. Arthritis Rheum 2008;59(7):952–60.
[61] Griffin NF, Melanie C, John W, Joanne E, Carl. Composite outcomes in randomized trials - greater precision but with greater uncertainty? JAMA 2003;289:2554–9.
[62] Cordoba G, Schwartz L, Woloshin S, Bae H, Gotzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ 2010;341:c3920.
[63] Phillips A, Haudiquet V. ICH E9 guideline 'statistical principles for clinical trials': a case study. Stat Med 2003;22(1):1–11 [discussion 13-17].
[64] Garattini S, Bertele V. Non-inferiority trials are unethical because they disregard patients' interests. Lancet 2007;370(9602):1875–7.

[65] Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 2010;14(8):iii [ix-xi, 1-193].
[66] +AllTrials. All Trials Registered, All Results Reported. [avaiable at] http://www.alltrialsnet/.
[67] Stevens A, Shamseer L, Weinstein E, Yazdi F, Turner L, Thielman J, et al. Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review. BMJ 2014;348:g3804.
[68] Ioannidis JP. Some main problems eroding the credibility and relevance of randomized trials. Bull NYU Hosp Jt Dis 2008;66(2):135–9.
[69] Clarke M, Hopewell S, Chalmers I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. Lancet 2010;376(9734):20–1.
[70] Clarke M, Horton R. Bringing it all together: Lancet-Cochrane collaborate on systematic reviews. Lancet 2001;357(9270):1728.
[71] Young C, Horton R. Putting clinical trials into context. Lancet 2005;366(9480):107–8.
[72] Directive 2001/83/EC of the European Parliament and of the Council of. on the Community code relating to medicinal products for human use (Consolidated version: 16/11/2012); 6 November 2001.
[73] Garattini S, Bertele V. How can we regulate medicines better? BMJ 2007;335(7624):803–5.
[74] Garattini S, Bertele V. The scientific community should lobby to be able to apply for drug licences. BMJ 2012;344, e3553.
[75] Light DW, Maturo AF. Good Pharma. Palgrave Macmillan; 2015.
[76] Wislar JS, Flanagin A, Fontanarosa PB, Deangelis CD. Honorary and ghost authorship in high impact biomedical journals: a cross sectional survey. BMJ 2011;343:d6128.
[77] Chan A-W, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K. Berlin JA et al: SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med 2013;158:200–7.
[78] Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomized trials. Ann Int Med 2010;152(11):726–32.
[79] Chow SC, Wang H, Shao J. Sample size calculations in clinical research. Chapman & Hall/CRC Biostatistics Series. Chapman and Hall/CRC; 2007.
[80] DeMets DL, Lan KK. Interim analysis: the alpha spending function approach. Stat Med 1994;13(13–14):1341–56.
[81] Lindley DV. A statistical paradox. Biometrika 1957;44(1/2):187–92.
[82] Jennison C, Turnbull BW. Repeated confidence intervals for group sequential clinical trials. Control Clin Trials 1984;5(1):33–45.
[83] Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analysis. J Clin Epidemiol 2008;61:763–9.
[84] Thorlund K, Engstrøm J, Wetterslev J, Brok J, Imberger G, Gluud C. User manual for trial sequential analysis (TSA). Copenhagen, Denmark: Copenhagen Trial Unit, Centre for Clinical Intervention Research; 2011 1–115 [Available from http://www.ctu.dk/tsa].
[85] Wetterslev J, Thorlund K, Brok J, Gluud C. Trial sequential analysis may establish when firm evidence is reached in cumulative meta-analysis. J Clin Epidemiol 2008;61(1):64–75.
[86] Turner RM, Bird SM, Higgins JP. The impact of study size on meta-analyses: examination of underpowered studies in Cochrane reviews. PLoS One 2013;8(3), e59202.
[87] Roberts I, Ker K, Edwards P, Beecher D, Manno D, Sydenham E. The knowledge system underpinning healthcare is not fit for purpose and must change. BMJ 2015;350:h2463.
[88] Wetterslev J, Thorlund K, Brok J, Gluud C. Estimating required information size by quantifying diversity in random-effects model meta-analyses. BMC Med Res Methodol 2009;9:86.
[89] Gluud C, Jakobsen JC, Imberger G, Lange T, Wetterslev J. Re: the knowledge system underpinning healthcare is not fit for purpose and must change — reponses to the opposing viewpoints of Roberts and colleagues and Tovey and colleagues. BMJ 2015.
[90] Higgins JPT, Whitehead A, Simmonds M. Sequential methods for random-effects meta-analysis. Stat Med 2011;30(9):903–21.
[91] Lan GKK, DeMets DL. Discrete sequential boundaries for clinical trials. Biometrika 1983;70(3):659–63.
[92] Brok J, Thorlund K, Gluud C, Wetterslev J. Trial sequential analysis reveals insufficient information size and potentially false positive results in many meta-analysis. J Clin Epidemiol 2008;61:763–9.
[93] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D, Stroup DF. Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. Lancet 1999;354(9193):1896–900.
[94] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315(7109):629–34.
[95] Begg CB, Mazumdar M. Operating characteristics of a rank correlation test for publication bias. Biometrics 1994;50(4):1088–101.
[96] Duval S, Tweedie R. Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics 2000;56(2):455–63.
[97] Sutton AJ, Duval SJ, Tweedie RL, Abrams KR, Jones DR. Empirical assessment of effect of publication bias on meta-analyses. BMJ 2000;320(7249):1574–7.
[98] Moher D, Tetzlaff J, Tricco AC, Sampson M, Altman DG. Epidemiology and reporting characteristics of systematic reviews. PLoS Med 2007;4(3), e78.
[99] Pieper D, Antoine SL, Mathes T, Neugebauer EA, Eikermann M. Systematic review finds overlapping reviews were not mentioned in every other overview. J Clin Epidemiol 2014;67(4):368–75.
[100] Moja LP, Telaro E, D'Amico R, Moschetti I, Coe L, Liberati A. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005;330(7499):1053.
[101] Bjelakovic G, Nikolova D, Gluud LL, Simonetti RG, Gluud C. Antioxidant supplements for prevention of mortality in healthy participants and patients with various diseases. Cochrane Database Syst Rev 2012;3, CD007176.
[102] Puhan MA, Schunemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ 2014;349:g5630.
[103] Gluud C, Sorensen TI. New developments in the conduct and management of multi-center trials: an international review of clinical trial units. Fundam Clin Pharmacol 1995;9(3):284–9.
[104] McCulloch P, Altman DG, Campbell WB, Flum DR, Glasziou P, Marshall JC, et al. No surgical innovation without evaluation: the IDEAL recommendations. Lancet 2009;374(9695):1105–12.