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

Klinička praksa temeljena na dokazima: pregled prijetnji valjanosti dokaza i kako ih spriječiti

Silvio Garattini1, Janus C. Jakobsen2,3, Jørn Wetterslev2, Vittorio Bertele4,1, Rita Banzi1, Ana Rath4, Edmund A.M. Neugebauer5, Martine Laville6, Yvonne Masson6, Virginie Hivert6, Michaela Eikermann7, Burc Aydin8, Sandra Ngwabyt4, Cecilia Martinho9, Chiara Gerardi1, Cezary A Szmigielski10, Jacques Demotes-Maynard11, Christian Gluud2*

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 health care systems.

Key words: evidence-based clinical practice; evidence-based medicine; meta-analysis; randomised clinical trial; systematic review; trial sequential analysis

*Corresponding author:
Christian Gluud, M.D., Dr. Med. Sci.
The Copenhagen Trial Unit, Centre for Clinical Intervention Research, Department 7812, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark e-mail: cgluud@ctu.dk

**Sažetak.** Primjena najkvalitetnijih dokaza kliničkih istraživanja ključna je u odabiru ispravnog liječenja pacijenata. No, način na koji će se odabrati najbolji dokazi predstavlja često poteškoću. Ovim preglednim člankom prikazujemo opasnosti navedenog odabira, kao i načine kako ih umanjiti. Relevantni izvori razmatrani su pretragom literature u kombinaciji s osobnim datotekama. Izbor liječenja uglavnom se ne bi smio temeljiti isključivo na opservacijskim ili pojedinačnim rando-miziranim kliničkim studijama. Sustavni pregledi s metaanalizom svih identificiranih randomiziranih kliničkih studija procijenjenih sustavom stupnjevanja procjene, razvoja i evaluacije preporuka (engl. Grading of Recommendations Assessment, Development and Evaluation; GRADE) predstavljaju najvišu razinu dokaza. Iako su sustavni pregledi pouzdaniji od drugih vrsta dokaza, sve razine hijerarhije dokaza ugrožene su sustavnim pogreškama (engl. bias); pogreškama dizajna studije (zloupotreba surogatnih ishoda, složenih ishoda itd.) i slučajnim pogreškama (igra slučaja). Kliničke istraživačke infrastrukture mogu pomoći u pružanju većih i adekvatnije provedenih ispitivanja. Sekvencijska analiza studija može pomoći pri odlučivanju kada postoji dovoljna razina dokaza u metaanalizama. Ako se prijetnje valjanosti kliničkih istraživanja pažljivo razmatraju i minimiziraju, rezultati istraživanja bit će vrijedniji i korisniji pacijentima i zdravstvenim sustavima.

**Ključne riječi:** klinička praksa temeljena na dokazima; medicina temeljena na dokazima; meta-analiza; randomizirana klinička studija; sustavni pregled

---

**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\(^5\).

All levels of the hierarchy of evidence may be threatened by systematic errors; design errors; and random errors.

Other fields like neurology, nephrology, endocrinology, hepatology, and surgery are less fortunate\(^5\). Medical devices, nutrition, and rare diseases are considered fields especially 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

---

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

<table>
<thead>
<tr>
<th>Phases</th>
<th>Participants and study designs for preventive or therapeutic interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase I</td>
<td>Healthy participants or patients</td>
</tr>
<tr>
<td></td>
<td>– observational studies</td>
</tr>
<tr>
<td></td>
<td>– randomised clinical trials</td>
</tr>
<tr>
<td></td>
<td>designed to assess the safety (pharmacovigilance), tolerability, pharmacokinetics, and pharmacodynamics of an intervention.</td>
</tr>
<tr>
<td>Phase II</td>
<td>Patients with disease in question</td>
</tr>
<tr>
<td></td>
<td>– randomised clinical trials</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td>Phase III</td>
<td>Patients with disease in question</td>
</tr>
<tr>
<td></td>
<td>– randomised clinical trials</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td>Phase IV</td>
<td>Patients with disease in question</td>
</tr>
<tr>
<td></td>
<td>– randomised clinical trials</td>
</tr>
<tr>
<td></td>
<td>– observational studies</td>
</tr>
<tr>
<td></td>
<td>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).</td>
</tr>
</tbody>
</table>

For medical devices slightly different phases are described\(^10,^4\)
these fields and how these barriers could be broken down in order to improve their evidence-based clinical practice.\textsuperscript{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’).\textsuperscript{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.\textsuperscript{17-25}

**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.

**THE HIERARCHY OF EVIDENCE**

Different experimental designs have different inferential powers, hence the hierarchy of evidence (Fig. 1).\textsuperscript{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).\textsuperscript{13} All levels of the hierarchy may be threatened by systematic errors; design errors; and random errors.\textsuperscript{11,13,26}

**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).\textsuperscript{11} Systematic reviews are based upon peer-reviewed protocols and fol-
low standardised methodologies. 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.

**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 ongoing 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. The trial advocates consider that systematic reviews should only be viewed as hypothesis-generating research.

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. 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. 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.

Hence, a cornerstone of a high-quality systematic review is the application of transparent, rigorous, and reproducible methodology.

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). 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. Conversely, the error rates in meta-analyses were correct. Evidence from a well-conducted systematic review of several RCTs with low risk of bias therefore generally represents a higher level of evidence compared to the results from a single RCT. 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.

As a relatively new approach, network meta-analyses allow comparing interventions that have never been tested head to head in RCTs. Careful consideration is needed for network meta-analyses to avoid false positive results. Statistical and conceptual heterogeneity of the trials combined in a network meta-analysis should be assessed to avoid incoherence and thus chance findings. Reporting bias can affect the findings of a network meta-analysis and lead to incorrect conclusions about the treatments compared. Due to high number of pairwise comparisons in a network analysis, the risk of chance findings should be reduced (see below type I error). 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.

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

**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. Deeks and colleagues conducted simulations comparing results from RCTs to those of controlled cohort studies. They concluded that results of controlled cohort studies often differ from results of RCTs. 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 over- or underestimation of effects. 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). Other studies confirm that controlled cohort studies should not be used to validate intervention effects. There are a number of real and perceived obstacles for conducting RCTs. However, when new interventions are assessed, we should always randomise the first patient. 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. Cohort studies should also be used for monitoring clinical quality and stability of treatment effects after new treatments have been introduced in clinical practice.

**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.

**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). Savovic et al. 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 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 val-
id methods are not used to handle missing
data\textsuperscript{49,50}. Chan et al. have revealed how authors
of RCTs make selective outcome reporting, lead-
ing to a gross overestimation of treatment bene-
fits\textsuperscript{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\textsuperscript{54,55}.
The systematic review by Lundh and colleagues
clearly demonstrated that industry involvement
is associated with biased results\textsuperscript{56}. Such bias was
not explained by other bias domains\textsuperscript{56}.
In conclusion, bias associated with specific re-
ported trial design characteristics leads to exag-
geration of beneficial intervention effect
estimates. For each of the domains assessed
above, these effects were greatest for subjective-
ly assessed outcomes. The average magnitude of
overestimation of 10% to 20% is larger than most
‘true’ intervention effects.

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

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

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

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

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. 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. Several drugs are advertised based on surrogate outcomes even though they have no effect or detrimental effects on patient-centred outcomes.

RCTs ought to assess if an intervention is safe and effective. Surrogate outcomes may not necessarily be meaningful for patients or 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.

The COMET (Core Outcome Measures in Effectiveness Trials) Initiative 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.

Abuse of composite outcomes

To reduce the required sample size, RCTs often adopt composite outcomes. However, composite outcomes make it difficult to interpret the clinical significance of the results. Any benefit on a composite outcome may be presumed to relate to all its components, but evidence shows that intervention effects on composite outcomes often apply to a single component, most likely the less relevant. 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. 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.

Patient-centred single outcomes (e.g., all-cause mortality) should always be preferred to composite outcomes if power is sufficient using the single outcome.

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. 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. Non-inferiority trials often allow substantial harm to patients.
Abuse of poor reporting or no reporting

Trials with significant results are more likely to be published than those with neutral or negative results\textsuperscript{55}. Random errors (‘play of chance’) cause especially small trials to indicate both benefit and harm, when there is none\textsuperscript{35}. Therefore, publication bias will increase the risk of erroneous conclusions about intervention effects\textsuperscript{55}. The AllTrials initiative is campaigning for the publication of the results from all past, present, and future clinical trials\textsuperscript{66}. This initiative is of utmost importance but the reporting of each single trial must also be thorough and valid\textsuperscript{20}. Studies have shown that the poor description of trial interventions resulted in 40% to 89% of trials being non-replicable\textsuperscript{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\textsuperscript{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\textsuperscript{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\textsuperscript{35}. The Nordic Trial Alliance has called for full transparency of all clinical research\textsuperscript{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).

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 everyday 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 may apply to it?; (3) is the trial designed as a pragmatic trial so the effects of the trial interventions may be reproduced in a clinical setting?; and (4) was the initial research question valid\textsuperscript{48,72}? 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”\textsuperscript{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’\textsuperscript{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\textsuperscript{74-76}.

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. The inclusion of an adequate number of participants in RCTs aims at avoiding two\textsuperscript{77,78} possible drawbacks, i.e., to let the RCT show an effect that does not actually exist (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 ($\alpha$) and type II error ($\beta$); an anticipated intervention effect ($\mu$) on the primary outcome; and the variance of the primary outcome ($\sigma^2$). Given these four components, the formula provides an estimate of the sample size ($N$) needed to detect or
reject the anticipated intervention effect \((\mu)\) with the chosen error risks in trials with equal group size \(N = 4(Z_{1-\alpha/2} + Z_{1-\beta})\sqrt{\mu^2}\), where \(Z_{1-\alpha/2}\) and \(Z_{1-\beta}\) are the corresponding \((1 - \alpha/2)\) and \((1 - \beta)\) fractiles from the normal distribution\(^7\).

**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 a \(P\)-value is related to the acquired number of randomised participants. The thresholds for significance should then be adjusted according to the information size, i.e., the adjusted \(P\)-value has to reach a value lower than the \(\alpha\) 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 'longitudi-
nal’ multiplicity. 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. The procedure for interim-analysis of a RCT is called group sequential analysis. Often the O’Brien-Fleming $\alpha$-spending function is chosen. 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.

Required information size in a meta-analysis of RCTs

Contrary to RCTs, risks of random errors in systematic reviews have received relatively limited attention. Most of the RCTs in Cochrane systematic reviews are underpowered to detect even large intervention effects. 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. Therefore, most meta-analyses may be considered as interim-analyses of intervention effects on the way to the required information size (RIS). 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. first when the number of outcomes was above 200 and the cumulated sample was above 2000 with moderate heterogeneity, the risk of overestimation declined towards 5%. 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%. Estimating a RIS therefore seems crucial for the interpretation of the statistical significance of results of meta-analyses.

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. Trial Sequential Analysis is a sequential meta-analysis methodology of cumulative meta-analysis using Lan-DeMets monitoring boundaries. Lan-DeMets monitoring boundaries offer the possibility to demonstrate if information size adjusted thresholds for benefit, harm, or futility are crossed.

To assess a meta-analysis transparently, a preplanned sequential meta-analysis with a priori chosen anticipated intervention effect $\mu$, $\alpha$, $\beta$, and should be part of any protocol for a systematic review. Trial Sequential Analysis offers such a transparent platform and a programme with a manual is available for free at: www.ctu.dk/tsa.

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

Different types of biases hamper the conduct and interpretation of systematic reviews. Selective reporting of completed studies leads to publication bias because positive trials with impressive findings, are more likely to be published. The simplest method to detect possible publication bias is visual inspection of a funnel plot. Other methods might contribute, including Egger test, Begg-Mazumdar test, and ‘trim-and-fill’ method. 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. 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. 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, 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. Also, selective reporting of other studies occurs often.

The PRISMA authors listed 27 items to be included when reporting a systematic review. 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.

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

**GRADING THE EVIDENCE WITH 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. 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.

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). Recommendations to administer or not administer an intervention should be based on the trade-offs between benefits on the one hand, and risks, burdens, and costs on the other. If benefits outweigh risks, experts will recommend that clinicians offer a treatment to a given specified patient group. After going through the process of grading evidence, the overall quality will be categorised as high, moderate, low, or very low. 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 risks 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 and hopefully we will meet such interventions in the future. 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.

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. Such designs ought always to be based on up-to-date systematic reviews of the available evidence.

**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. The threats to validity and the associated waste of clinical research affects all interventions and all diseases. 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, www.equator-network.org).

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. We will in four connected papers discuss the common and special bottlenecks for conducting clinical research on medical devices, nutrition, and rare diseases. 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.

ACKNOWLEDGEMENTS

The ECRIN-IA grant from the EU FP7 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. 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 RCTs and other studies, and hence threats to external validity of clinical research) suggesting ways to blow up the bottlenecks and erase the threats.

Conflicts of interests: CG and JW are members of the Copenhagen Trial Unit’s task force to develop theory and software for doing Trial Sequential Analysis. All authors are involved in conducting randomised clinical trials and members of ECRIN. No additional conflicts of interests.

REFERENCES

11. Higgins JPT, Green S. The Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0. The Co-


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;302:2361-2.


89. Gluud C, Jakobsen JC, Imberger G, Lange T, Wetterslev J. The knowledge system underpinning healthcare is not fit for purpose and must change – Responses to the opposing viewpoints of Roberts and colleagues and Tovey and colleagues. BMJ 2015;350:h2463.


