Eur Respir J 2009; 34: 552–558
DOI: 10.1183/09031936.00095408
Copyright©ERS Journals Ltd 2009

PERSPECTIVE

How to increase the value of randomised trials in COPD research

M.A. Puhan*,# and H.J. Schünemann¶,+

ABSTRACT: **Methodological criteria that increase the validity of randomised trials are often not considered in respiratory research, even in large chronic obstructive pulmonary disease (COPD) trials. We describe four important aspects in the design, analysis and reporting of randomised trials, selected based on their relevance to current COPD research and based on our judgments of importance for researchers and users of the literature.**

First, to optimally control for confounding, where confounding refers to a factor that is associated with an exposure or intervention and influences the outcome, a clear definition of the main relationship between treatment and the primary outcome as well as identification of measurable confounders is required. In addition to randomisation *per se* as the key method to protect against confounding, restriction (excluding patients with specific characteristics that may introduce confounding), stratification (separate randomisation of patients with specific characteristics) and statistical adjustment are means to be considered to optimally control for confounding that simple randomisation may not achieve.

Secondly, the selection of the primary outcome should be guided by the importance to patients. Secondary outcomes provide hypotheses about the effects observed for the primary outcome and can provide important data for systematic reviews and meta-analyses, but should be interpreted with caution in single trials.

Thirdly, in study power calculations, not only the actual sample size, but the number of events, has a large influence on the power of the study and, often, unrealistic assumptions about event rates are made to increase the feasibility of trials.

Finally, essential steps to transfer results from research to practice include complete reporting of trials and developing tools, such as decision aids, to support patients and physicians in their shared decision making.

KEYWORDS: **Chronic obstructive pulmonary disease, evidence-based medicine, study design**

AFFILIATIONS
*Dept of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD, USA.
#Horten Centre, University of Zurich, Zurich, Switzerland.
¶Clarity Research Group, Dept of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, ON, Canada.
+Dept of Epidemiology, Italian National Cancer Institute ''Regina Elena'', Rome, Italy.

CORRESPONDENCE
H.J. Schünemann
Dept of Clinical Epidemiology and Biostatistics, Faculty of Health Sciences
McMaster University
1200 Main Street W
Hamilton
ON
Canada
E-mail: schuneh@mcmaster.ca

Received:
June 24 2008
Accepted after revision:
March 23 2009

Randomised controlled trials provide the least biased estimates of treatment effects. They support healthcare decisions by providing the evidence that is necessary to balance benefits and downsides of treatments. However, our confidence in these decisions requires consideration of important methodological concepts in the design, conduct and reporting of randomised controlled trials that even large trials sometimes omit. For example, the decision regarding primary and secondary outcomes in the design phase not only influences the interpretation of trial results, but also impacts on the validity of the findings. Furthermore, many users of the literature believe that randomisation fully protects against bias. However, if randomisation fails to achieve groups that are balanced for factors that determine outcomes, results can still be biased.

Unfortunately, few trials in respiratory research, and chronic obstructive pulmonary disease (COPD) research in particular, fully deal with these important issues that are linked to the questions of whether the study results are internally valid and how they can be applied to clinical practice [1, 2]. For example, only 27.0% and 11.6% out of 344 COPD trials report on the generation of an appropriate randomisation or concealment of randomisation, respectively [3]. Although larger sample sizes reduce some methodological problems of trials in COPD, there are additional and, perhaps more fundamental, issues that increase the validity and enhance the

European Respiratory Journal
Print ISSN 0903-1936
Online ISSN 1399-3003

application of trials [4]. The aim of this article is, therefore, to discuss selected methodological issues in the design, analysis and interpretation of randomised trials that inform readers and inform investigators of future studies. We selected the following four issues based on the limitations encountered in the literature and in teaching at international workshops for advanced users of the clinical research literature: 1) key concepts to reduce residual confounding in randomised trials; 2) methodological aspects regarding the choice and use of primary and secondary outcomes; 3) sample size calculations and the importance of event rates for these calculations; and 4) how to present and use outcome information in randomised trials.

KEY CONCEPTS TO REDUCE RESIDUAL CONFOUNDING IN RANDOMISED TRIALS

Highly valid trials provide estimates about treatment effects that are as close as possible to the truth. These effect estimates may be expressed as relative estimates (relative risks or odds ratios) or as absolute effects (including natural frequencies, risk differences or numbers needed to treat). There are a number of methodological factors that can bias results, such as lack of blinding of investigators who assess outcomes (*e.g.* when assessing 6-min walk distance) or an inappropriate per-protocol instead of an intention-to-treat analysis. These latter factors are well described in other texts [5].

Another fundamental concern that may lead to systematic over or underestimation of treatment effects is the presence of confounding [6]. Confounding refers to an observed effect (*e.g.* on mortality) that is not only due to the exposure of interest (*e.g.* treatment with a long-acting bronchodilator) but also due to some other outcome-determining factor (*e.g.* disease severity). In ''classic'' observational epidemiology, a confounder is a factor that is associated with the exposure (causally or noncausally) and influences the outcome (*e.g.* an association between yellow fingers from nicotine exposure and death is completely confounded by smoking, which causes both yellow fingers and death). An example of partial or incomplete confounding is the influence of disease severity in observational studies on the association between bronchodilators and death. Severely ill patients are more likely to receive a bronchodilator and are also at a greater risk of death. In the context of randomised trials, unbalanced groups (*e.g.* one group suffers, on average, from more severe disease) can cause confounding. Although it may be argued that this is not a typical form of confounding but a form of random error resulting from randomisation, the consequences remain analogous: disease severity influences and, thereby, confounds the observed effects. Unbalanced groups in randomised trials are, therefore, not a form of random error that would reduce precision of effect estimates but a source of residual confounding that needs to be controlled for. It should be noted, however, that a factor such as disease severity may also be of interest because it could alter the effects of a treatment. The latter, called effect modification, is particularly important for clinicians because it may inform research for targeting treatments to patient groups that benefit most. A prerequisite for evaluating effect modification is, however, that groups are evenly balanced for the effect modifier of interest (*e.g.* disease severity) in order to allow for appropriate subgroup analyses and interpretation according to well-defined criteria [7].

While there is no protection against imbalance of unknown confounders, investigators can protect their study results against known factors that may cause residual confounding. We will describe three common approaches to control for known and measurable confounders: restriction, pre-stratification and statistical adjustment.

Restriction

Restriction is a simple method that describes the exclusion of patients with characteristics that may introduce confounding (*e.g.* patients with more severe disease), and is a very powerful method of reducing the risk of confounding. However, it comes at the cost of recruiting selected study populations, which, in turn, may limit the applicability of findings. Applicability should be distinguished from internal validity, *i.e.* whether the study results in themselves are valid for the setting, population, intervention and outcomes of the study based on its design and execution. Applicability refers to external validity and whether the results are valid when applied outside of the trial environment. While a number of approaches exist to evaluate applicability [8, 9], the GRADE (grades of recommendation assessment, development and evaluation) working group has provided a general conceptual approach for the distinction between internal validity, or risk of bias for a given body of evidence, and how directly the findings of this evidence can be applied to the question in healthcare [10, 11]. The concept of directness includes making judgments about whether similar results could be expected when transferring evidence from the trial PICO (population, intervention, the comparison intervention (*e.g.* placebo or alternative management) and outcomes) to the PICO of the actual healthcare questions that requires an action. For example, if a trial or a body of trials included only patients with moderate COPD, are the results applicable to patients with severe COPD (P); are results obtained with one inhaled steroid expected to differ across the entire class and for different doses (I); are the comparator interventions or alternatives used in the trial the same in clinical practice (C); and are the outcomes directly relevant to patients or is there uncertainty about how important they are for decision making (O). ROTHWELL [7] provides helpful specific examples that fall into the PICO categories, such as genetic factors and comorbidities, as examples for making judgments about the directness of the population. Restriction will limit directness, but if indirectness is judged to be of little concern or if the focus is efficacy, it is a powerful means to reduce potential spurious findings.

Pre-stratification

Pre-stratification is commonly used and means that randomisation is, for example, performed separately for patients with moderate and severe COPD. In this way, equal number of patients with moderate and severe disease will be allocated to treatment groups, reducing the risk of imbalanced groups.

Statistical adjustment

If groups are not well balanced for potential confounders, analysts can still adjust for imbalances in the analyses. For example, in a randomised trial that assessed the effects of respiratory rehabilitation on mortality, the control group had, on average, more severe disease, as expressed by a lower exercise capacity and poorer lung function, two factors known to be associated with mortality [12–14]. The effects of respiratory rehabilitation effects may be exaggerated because the average prognoses of the groups were not the same. Introducing and adjusting for measures of exercise capacity and lung function in the statistical analysis (for example in multivariate regression analyses) could account for these differences. However, it should be noted that statistical adjustment cannot eliminate the problem of unknown or unmeasurable confounders.

Restriction, pre-stratification and statistical adjustment are rarely used in COPD trials [1, 2]. For their optimal use, it is important to identify potential confounders during the design stage of a trial. Therefore, it is useful to explicitly describe the main relationship (fig. 1a) defined by the treatments and the primary outcome. All potential confounders, often suggested by prior research, should be listed and evaluated according to their potential to bias the primary outcome. Thereby, investigators can prioritise the most important confounders and decide how to control for them in the design phase. Control for confounding during the planning stage using randomisation, restriction and pre-stratification, and statistical adjustment during analysis of randomised trials, will reduce the type of confounding we have described.

METHODOLOGICAL ASPECTS REGARDING THE CHOICE OF PRIMARY AND SECONDARY OUTCOMES

The CONSORT (Consolidated Standards of Reporting Clinical Trials) statement, a blueprint for the reporting of clinical trials, asks authors to identify a single primary outcome measure and a clear distinction from secondary outcomes [5]. The rationale is that investigators should specify which outcome they considered to be of greatest importance to guide interpretation of the trial. Specifying only one primary outcome prevents problems of interpretation associated with multiplicity of analyses. A second common and related reason to define a primary outcome relates to sample size calculation. The expected effect for the primary outcome (*e.g.* a difference of 40 m in 6-min walk distance) is used to estimate the required sample size that will allow detecting this effect with appropriate power [15]. These arguments for a primary outcome are justified and help adequate interpretation of sufficiently powered trials.

However, the perhaps most important reason for defining a primary outcome is its role in the main relationship of the trial between intervention and outcomes. The underlying reason is that confounders may differ by outcome. Consider figure 1a, which shows a trial investigating the effects of long-acting bronchodilators on mortality. Plausible confounders to control for include forced expiratory volume in 1 s (FEV1; or, alternatively, the Global Initiative for Chronic Obstructive Lung Disease stage), previous exacerbations and exercise capacity as risk factors for mortality. In contrast, if health-related quality of life (HRQL) is of primary interest (fig. 1b), the most important confounders one might want to control for may include dyspnoea at baseline, previous exacerbations and comorbidities. Restriction and pre-stratification for the latter trial may differ substantially from the former. This simple example shows that the confounders that need to be controlled for by using the aforementioned methods depend on the primary outcome. As a consequence, trial results for secondary outcomes might be prone to residual confounding, particularly if trials are small.

How to choose the primary outcome

The most important criterion for choosing the primary outcome in trials should be the degree of importance of the outcome to patients. A number of methods exist for determining the importance or value of outcomes from the utility literature, but there is a general lack of agreement on the best method [11, 16]. However, integrative measures such as overall HRQL, exacerbations and mortality are logical primary outcomes in COPD trials. HRQL can be measured with good validity and reliability [17]. Furthermore, many validated HRQL instruments are also responsive to change [18]. However, in some instances investigators might want to gain more insights into pathophysiological processes and decide on primary outcomes that are not important for patients, but help in explaining the occurrence, mechanism or progression of disease.

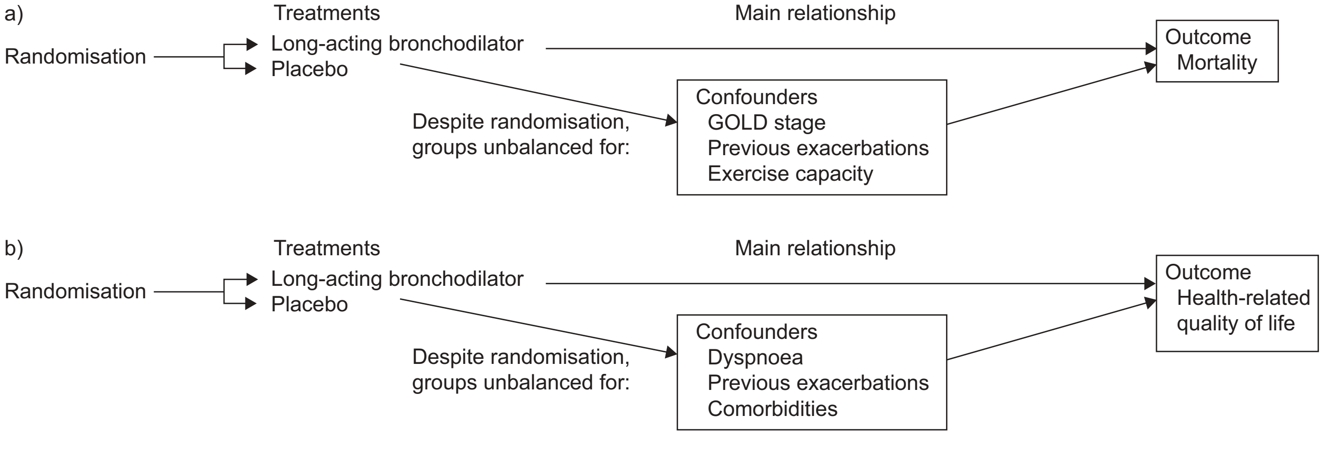

FIGURE 1. Defining the main relationships between treatment and outcome. GOLD: Global Initiative for Chronic Obstructive Lung Disease.

Once investigators have selected a primary outcome, measurement considerations play an important role. Exacerbations, for example, can have a symptom-based [19] or an event-based definition [20]. Adjudication committees should decide, based on complete and unbiased data collection, whether patients experienced an exacerbation or not [21]. For the assessment of HRQL, instruments should be available and validated in the respective language. Finally, it is important to choose instruments for which guidance about interpretation of patient importance exists, such as thresholds for the minimal important difference (MID) [22]. The MID can also guide sample size calculations and trial interpretation [16].

Secondary outcomes

Secondary outcomes remain important because they may offer additional insights and offer hypothesis-generating results. In the TORCH (Towards a Revolution in COPD Health) study [21], for example, the investigators could confirm concerns that inhaled corticosteroids increase the risk of pneumonia. Thus, the use of secondary outcomes is critical for balancing benefits and downsides of treatments. While single trials can often not provide sufficient data on such secondary outcomes, they can then be summarised in well-performed systematic reviews that should be regularly updated to integrate new data. An example of the value of collecting data on such secondary outcomes is supported by the results of the UPLIFT (Understanding Potential Long-term Impacts on Function with Tiotropium) study [23], which inform the concerns raised in a recent systematic review and a larger registry study suggesting an increased risk of cardiovascular adverse effects and death from anticholinergic metered-dose inhalers in COPD [24, 25]. The number of patients and the deaths as secondary outcome in long-term tiotropium treatment trials increased from 3,247 and 97 in the systematic review by SINGH *et al.* [24], to 8,294 and 1,038 after UPLIFT, respectively. The detailed reporting of the secondary outcome data in the UPLIFT study, thus, helped to inform the debate about the increased death rate by providing a large number of additional events.

However, results from secondary outcomes should be interpreted with caution because of several potential caveats, such as less careful ascertainment of the data. For example, mortality was a secondary outcome in the recent INSPIRE (Investigating New Standards for Prophylaxis in Reducing Exacerbations) trial that compared combined treatment (inhaled steroid plus long-acting β-agonist) with a long-acting anticholinergic alone [26]. Although this trial showed very similar results for the primary outcome (exacerbation rates for both treatment regimens around 1.3 exacerbations per person-yr with p=0.66 for between-group difference) mortality was significantly lower in patients with combined treatment (3% *versus* 6% with long-acting anticholinergic; p=0.03). In the context of no difference in exacerbation rates and evidence from earlier trials it does not seem plausible that combined treatment lowers mortality compared to long-acting bronchodilator alone [27]. Indeed, a closer look at how death was ascertained in the INSPIRE trials reveals that death was only counted for patients who did not withdraw from treatment (65.5% with combined treatment and 58.3% with long-acting anticholinergic). Thus, it appears that one-third of patients did not enter the mortality analysis. Following the intention-to-treat principle protects against the bias that this per-protocol analysis may have caused [28, 29]. This example indicates that less stringent measurement of secondary outcomes may cause severe problems with missing data and render intention-to-treat analyses difficult.

In addition, the use of multiple secondary outcomes augments problems with interpretation and reporting. For example, how should we interpret contradicting results from secondary outcomes? Often, the number of measured secondary outcomes is so large that positive results emerge by chance. Methodological and statistical safeguards and solutions are available to address these concerns, but avoiding selective reporting of secondary outcomes will be the most important safeguard. Secondary outcomes should be selected to complement and enhance the plausibility of the main aim of the trial and inform decision making. In the TORCH study [21], for example, exacerbations and FEV1 were secondary outcomes that are associated with death in prognostic studies. The effect of combination treatment was consistent across these outcomes. One might argue that the reduction in mortality compared with placebo, although it just failed to reach conventional statistical significance, can be explained by a reduction in exacerbations and a slower decline in lung function, both of which are associated with lower mortality and suggest disease modification.

Results based on secondary outcomes require careful evaluation before they should inform clinical practice. The reasons include: that data collection might be less stringent and robust than for the primary outcome; that selective reporting of secondary outcomes showing significant results is frequent; and that methodological problems, such as insufficient control for confounding and the use of per-protocol analyses, can limit the validity of results from secondary outcomes. Finally, systematic reviews including those of secondary outcomes should become a standard approach to systematically summarising findings across studies.

SAMPLE SIZE CALCULATIONS AND THE IMPORTANCE OF EVENT RATES

Once investigators have determined the primary outcome and how to control for residual confounding, the sample size should be estimated to ensure adequate precision of the effect estimates. The TORCH study selected mortality as the primary outcome [21]. The investigators determined the required sample size to detect an expected absolute difference in mortality after 3 yrs between the placebo group (17%) and the group with combined treatment (12.7%) with 90% power. The observed mortality rates were 15.2% (placebo group) and 12.6% (combined treatment group), respectively. Thus, the assumed difference of 4.3% (equal to 17% minus 12.7%, for a relative risk reduction of 25.3%) for the absolute reduction in death was much larger than observed. Analogously, the assumed relative risk reduction was much larger than the actually observed relative effect (15.2% minus 12.6% equals 2.6%; 2.6% divided by 15.2% equals an observed effect of 17.1%).

Unfortunately, overestimation of both event rates and relative treatment effects is common for sample size calculation of many trials. Definition of what constitutes an important risk

TABLE 1 The relationship of event rates to relative and absolute risk reductions

	Patients with outcome n	Patients without outcome n	Event rate (risk)
Treatment group	1	9999	0.0001
Placebo group	2	9998	0.0002
Absolute risk reduction or risk difference		0.0001 (95% CI -0.0002–0.0004)	
Relative risk		0.50 (95% CI 0.05–5.51)	
Relative risk reduction		50% (95% CI -95%–450%)	

reduction is challenging for binary outcomes such as mortality. In practice, a steering committee decides what constitutes a relevant treatment effect, but this is often not based on solid data. While there is no simple solution for this dilemma, the estimates should be realistic and relative treatment effects of >20% are uncommon and should, therefore, be applied with caution [30].

Furthermore, beyond sample size, one of the main factors determining sample size calculations are event rates; that is, the number of patients who experience an outcome. The following calculation exemplifies this problem: consider a large randomised trial of 20,000 patients randomised into two groups, treatment and placebo, of 10,000 patients each. One would believe that trials of this magnitude should be able to detect large effects, such as relative risk reductions of 50%, with great certainty. However, if the event rate is low (*e.g.* one event in the treatment group and two events in the placebo group, as an extreme example shown in table 1), the results would not be statistically significant ($p=0.56$). There would be great uncertainty about the true effect, making a very large reduction in risk of 95% equally likely as a relative risk increase of 450%.

Conversely, a much smaller trial with 2,000 patients (1,000 in the treatment and the placebo group each) and 100 and 200 events, respectively, will yield the same relative risk of 0.5 ($p<0.001$) with sufficiently precise confidence intervals of 0.37–0.60 to indicate that a large effect exists (table 2).

In a more realistic example, investigators might want to observe a relative risk reduction in exacerbation rates of 20% over a treatment period of 1 yr; assuming a baseline event rate of 20% in the control group and a 16% event rate in the treatment group, around 3,000 patients are needed (80% power). If the baseline event rate is 40% only around 1,200 patients are needed to show a statistically significant relative risk reduction of 20%. Exacerbation baseline rates are determined by inclusion and exclusion criteria, the definition of exacerbations and length of follow-up. Investigators should reflect carefully on these three parameters during the planning stage, as they determine sample size, funding requirements and feasibility. For continuous outcomes, such as HRQL data, defining both important treatment effects and sample size is simpler when the MID has been established or other thresholds for interpretability exist [22].

Multi-arm trials such as TORCH offer an additional challenge. The single drug groups (bronchodilator and inhaled corticosteroid) were apparently not considered for sample size calculations. The four-armed TORCH study opens, thus, the debate whether sample size calculations should be based on the most important comparison or on more than one comparison. For clinicians, the relevant question would have been whether to treat patients with a single drug (bronchodilator) or with a combination (bronchodilator plus inhaled steroid) rather than the comparison between combination treatment and placebo. The focus on the latter comparison may have been guided by the believe that trials should always include a ‘‘least possible treatment’’ arm, but investigators and those using the literature should realise that the placebo control may be necessary only for add-on treatments. For example, the TORCH trial could have been powered to show that combination treatment is superior to single bronchodilator treatment if the single treatment arm had also received a placebo. The trial could also have been powered to explore whether single bronchodilator treatment is not inferior to combination treatment.

In superiority trials that focus on the superiority of a drug over placebo, the power to detect expected differences is commonly set at 80%. In order to yield more precise results, the TORCH trial set the power at 90% [21], which increases sample sizes

TABLE 2 Smaller trials with high event rates can provide narrow confidence intervals

	Patients with outcome n	Patients without outcome n	Event rate (risk)
Treatment group	100	900	0.1
Placebo group	200	800	0.2
Absolute risk reduction or risk difference		0.1 (95% CI 0.07–0.13)	
Relative risk		0.50 (95% CI 0.40–0.63)	
Relative risk reduction		50% (95% CI 37%–60%)	

substantially. In noninferiority or equivalence trials, however, investigators are less flexible in their choice of power. The power should be set at ⩾90% in order to minimise the problem of wrongly assuming equivalence if the effects of two interventions actually differ (commonly referred to as type II or β error) [16].

Finally, in the realistic situation of limited funding, sample size calculations will depend on issues of feasibility, and many approaches to sample size calculations exist [30]. We described that fundamental requirements for sample size calculations include: careful selection of the primary outcome; event rates; consideration of the main comparison, in the case of multi-arm trials; and the assumption of effect sizes based on the MID or on realistic estimates of relative or absolute risk reductions.

HOW TO PRESENT AND USE INFORMATION OF BENEFITS AND OUTCOMES IN CLINICAL TRIALS

COPD trials such as TORCH raise, as do other multi-arm studies, several issues for reporting and interpretation. The investigators drew their main conclusions on the basis of comparisons between combined treatment and placebo. Combined treatment appears superior to placebo. However, how should we interpret the comparison between combined and single bronchodilator treatment, which may indeed be the more important clinical question? There were no statistically significant differences between these two treatment options with respect to mortality and severe exacerbations (hospital admissions), but combined treatment significantly reduced any (moderate and severe) exacerbations and moderate exacerbations and reduced the decline of FEV$_1$. Differences in HRQL were statistically significant but these differences were well below the MID; only a comparison of the proportions of patients achieving the MID would inform the discussion. Finally, combined treatment significantly increased the risk of pneumonia (19.6% over 3 yrs) compared with single bronchodilator treatment (13.3%). Therefore, one could argue that single bronchodilator treatment should be favoured because there was no important difference in the primary outcome. Furthermore, there were no differences in other outcomes important for patients, such as hospital admissions or HRQL, while the safety profile was more favourable for single bronchodilator treatment.

This example illustrates that informative but complex trials such as TORCH need to be viewed by balancing all important outcomes, in particular in the absence of a clearly dominating mortality difference. However, how can we ensure that clinicians and patients are informed about all results? The foregoing question raises the more important questions of how should physicians and patients be informed and how should they balance the benefits and downsides of single bronchodilator and combined treatment? We emphasised that a prerequisite for balanced decisions is that reporting is as complete as it is in the case of the TORCH trial. Interpretation should include all comparisons and all outcomes [5]. This would enable investigators to develop decision aids, where, based on the entire body of evidence from all available trials and summarised in high-quality systematic reviews, the benefits and downsides are illustrated comprehensively [31]. Decision aids are particularly valuable for value-sensitive decisions in which the balance of benefits and downsides is not straightforward and patients have different views on outcomes and treatments. Whether patients benefit from such an informed decision making requires, however, testing in additional trials.

CONCLUSION

The design and conduct of COPD trials requires a clear definition of the main relationship between treatment and primary outcomes, identification of confounders and selection of means to optimally control for these confounders during the design phase. Selection of the primary outcome should be guided by the importance to patients. Realistic estimates of treatment effects should inform trials. Secondary outcomes will provide complementary information to explain or describe consequences of the effects observed for the primary outcome, but should be interpreted with caution because they often provide limited data or are incompletely reported. Finally, essential steps to transfer the results from research to practice requires complete reporting of results and developing tools such as decision aids that convey benefits and downsides of treatments in an informative way to patients and clinicians, in order to support their shared decision making.

SUPPORT STATEMENT

M.A. Puhan is supported by a career award of the Swiss National Science Foundation (3233B0/115216/1). H.J. Schünemann is funded by a European Commission grant: the human factor, mobility and Marie Curie actions (Scientist Reintegration Grant; IGR 42192).

STATEMENT OF INTEREST

Statements of interest for M.A. Puhan and H.J. Schünemann can be found at www.erj.ersjournals.com/misc/statements.dtl

ACKNOWLEDGEMENTS

The authors are grateful to L. Fabbri (University of Modena and Reggio Emilia, Modena, Italy) for suggesting this topic and reviewing an earlier version of the document.

REFERENCES

1 Puhan MA, Schunemann HJ, Frey M, *et al.* Value of supplemental interventions to enhance the effectiveness of physical exercise during respiratory rehabilitation in COPD patients. A systematic review. *Respir Res* 2004; 5: 25.

2 Puhan MA, Schunemann HJ, Frey M, *et al.* How should COPD patients exercise during respiratory rehabilitation? Comparison of exercise modalities and intensities to treat skeletal muscle dysfunction. *Thorax* 2005; 60: 367–375.

3 Bausch B, Spaar A, Kleijnen J, *et al.* Quality of randomised trials in COPD research. *Eur Respir J* 2009; Epub ahead of print: PMID 19574328.

4 Calverley PM, Rennard SI. What have we learned from large drug treatment trials in COPD? *Lancet* 2007; 370: 774–785.

5 Altman DG, Schulz KF, Moher D, *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann Intern Med* 2001; 134: 663–694.

6 Rothman KJ. Epidemiologic methods in clinical trials. *Cancer* 1977; 39: Suppl. 4, 1771–1775.

7 Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; 365: 176–186.

8 Rothwell PM. External validity of randomised controlled trials: ‘‘to whom do the results of this trial apply?’’ *Lancet* 2005; 365: 82–93.

9 Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. *Arch Intern Med* 2008; 168: 133–135.

10 Guyatt GH, Oxman AD, Kunz R, *et al.* What is ''quality of evidence'' and why is it important to clinicians? *BMJ* 2008; 336: 995–998.

11 Schunemann HJ, Jaeschke R, Cook DJ, *et al.* An official ATS statement: grading the quality of evidence and strength of recommendations in ATS guidelines and recommendations. *Am J Respir Crit Care Med* 2006; 174: 605–614.

12 Marquis K, Debigare R, Lacasse Y, *et al.* Midthigh muscle cross-sectional area is a better predictor of mortality than body mass index in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 2002; 166: 809–813.

13 Pinto-Plata VM, Cote C, Cabral H, *et al.* The 6-min walk distance: change over time and value as a predictor of survival in severe COPD. *Eur Respir J* 2004; 23: 28–33.

14 Ramirez-Venegas A, Sansores RH, Perez-Padilla R, *et al.* Survival of patients with chronic obstructive pulmonary disease due to biomass smoke and tobacco. *Am J Respir Crit Care Med* 2006; 173: 393–397.

15 Puhan MA, Mador MJ, Held U, *et al.* Interpretation of treatment changes in 6-minute walk distance in patients with COPD. *Eur Respir J* 2008; 32: 637–643.

16 Puhan MA, Busching G, Schunemann HJ, *et al.* Interval *versus* continuous high-intensity exercise in chronic obstructive pulmonary disease: a randomized trial. *Ann Intern Med* 2006; 145: 816–825.

17 Schunemann HJ, Goldstein R, Mador MJ, *et al.* A randomised trial to evaluate the self-administered standardised chronic respiratory questionnaire. *Eur Respir J* 2005; 25: 31–40.

18 Puhan MA, Guyatt GH, Goldstein R, *et al.* Relative responsiveness of the Chronic Respiratory Questionnaire, St. Georges Respiratory Questionnaire and four other health-related quality of life instruments for patients with chronic lung disease. *Respir Med* 2007; 101: 308–316.

19 Anthonisen NR, Manfreda J, Warren CP, *et al.* Antibiotic therapy in exacerbations of chronic obstructive pulmonary disease. *Ann Intern Med* 1987; 106: 196–204.

20 Celli BR, MacNee W. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *Eur Respir J* 2004; 23: 932–946.

21 Calverley PM, Anderson JA, Celli B, *et al.* Salmeterol and fluticasone propionate and survival in chronic obstructive pulmonary disease. *N Engl J Med* 2007; 356: 775–789.

22 Schunemann HJ, Griffith L, Jaeschke R, *et al.* Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *J Clin Epidemiol* 2003; 56: 1170–1176.

23 Tashkin DP, Celli B, Senn S, *et al.* A 4-year trial of tiotropium in chronic obstructive pulmonary disease. *N Engl J Med* 2008; 359: 1543–1554.

24 Singh S, Loke YK, Furberg CD. Inhaled anticholinergics and risk of major adverse cardiovascular events in patients with chronic obstructive pulmonary disease: a systematic review and meta-analysis. *JAMA* 2008; 300: 1439–1450.

25 Lee TA, Pickard AS, Au DH, *et al.* Risk for death associated with medications for recently diagnosed chronic obstructive pulmonary disease. *Ann Intern Med* 2008; 149: 380–390.

26 Wedzicha JA, Calverley PM, Seemungal TA, *et al.* The prevention of chronic obstructive pulmonary disease exacerbations by salmeterol/fluticasone propionate or tiotropium bromide. *Am J Respir Crit Care Med* 2008; 177: 19–26.

27 Wilt TJ, Niewoehner D, MacDonald R, *et al.* Management of stable chronic obstructive pulmonary disease: a systematic review for a clinical practice guideline. *Ann Intern Med* 2007; 147: 639–653.

28 Lundh A, Gotzsche PC. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC Med Res Methodol* 2008; 8: 22.

29 Melander H, Ahlqvist-Rastad J, Meijer G, *et al.* Evidence b(i)ased medicine – selective reporting from studies sponsored by pharmaceutical industry: review of studies in new drug applications. *BMJ* 2003; 326: 1171–1173.

30 Schulz KF, Grimes DA. Sample size calculations in randomised trials: mandatory and mystical. *Lancet* 2005; 365: 1348–1353.

31 Akl EA, Grant BJ, Guyatt GH, *et al.* A decision aid for COPD patients considering inhaled steroid therapy: development and before and after pilot testing. *BMC Med Inform Decis Mak* 2007; 7: 12.