Published in final edited form as:

J Clin Epidemiol. 2011 June ; 64(6): 583–593. doi:10.1016/j.jclinepi.2010.09.007.

Optimism bias leads to inconclusive results - an empirical study

Benjamin Djulbegovic^{1,2}, Ambuj Kumar^{1,2}, Anja Magazin², Anneke T. Schroen⁴, Heloisa Soares⁵, Iztok Hozo⁶, Mike Clarke⁷, Daniel Sargent⁸, and Michael J. Schell²

¹ University of South Florida, Center for Evidence-based Medicine & Health Outcomes Research, Tampa, FL

² Moffitt Cancer Center & Research Institute, Tampa, FL

⁴ University of Virginia, Charlottesville, VA

⁵ Mount Sinai Medical Center, Miami, FL

⁶ Indiana University North, Gary, IN

⁷ UK Cochrane Centre

⁸ Division of Biomedical Statistics and Informatics, Mayo Clinic

Abstract

Objective—Optimism bias refers to unwarranted belief in the efficacy of new therapies. We assessed the impact of optimism bias on a proportion of trials that did not answer their research question successfully, and explored whether poor accrual or optimism bias is responsible for inconclusive results.

Study Design—Systematic review

Setting—Retrospective analysis of a consecutive series phase III randomized controlled trials (RCTs) performed under the aegis of National Cancer Institute Cooperative groups.

Results—359 trials (374 comparisons) enrolling 150,232 patients were analyzed. 70% (262/374) of the trials generated conclusive results according to the statistical criteria. Investigators made definitive statements related to the treatment preference in 73% (273/374) of studies. Investigators' judgments and statistical inferences were concordant in 75% (279/374) of trials. Investigators consistently overestimated their expected treatment effects, but to a significantly larger extent for inconclusive trials. The median ratio of expected over observed hazard ratio or odds ratio was 1.34 (range 0.19 – 15.40) in conclusive trials compared to 1.86 (range 1.09 – 12.00) in inconclusive studies ($p < 0.0001$). Only 17% of the trials had treatment effects that matched original researchers' expectations.

Corresponding author: Benjamin Djulbegovic, MD, PhD, Professor of Oncology and Medicine, Clinical Translational Science Institute, Center for Evidence-based Medicine and Health Outcome Research, Department of Medicine, USF Health Clinical Research, 13101 Bruce B. Downs Boulevard, CMS3057, Tampa, FL 33612, bdjulbeg@health.usf.edu.

Conflict of interest: None.

Author contributions: Benjamin Djulbegovic conceived the idea along with Ambuj Kumar and Heloisa Soares. Anja Magazin, Ambuj Kumar, Heloisa Soares and Benjamin Djulbegovic collected the data. Benjamin Djulbegovic, Ambuj Kumar, Anja Magazin, Anneke T. Schroen, Heloisa Soares, Iztok Hozo, Mike Clarke, Daniel Sargent, and Michael J. Schell participated in data analysis, interpretation and drafting of this manuscript.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conclusion—Formal statistical inference is sufficient to answer the research question in 75% of RCTs. The answers to the other 25% depend mostly on subjective judgments, which at times are in conflict with statistical inference. Optimism bias significantly contributes to inconclusive results.

Keywords

optimism-bias; inconclusive trials; randomized controlled trials; bias; study design; systematic review

Key finding

Optimism bias refers to unwarranted belief in the efficacy of new therapies, and significantly contributes to inconclusive results. Formal statistical inference alone is not sufficient to answer the research question. The answers to the research question also depend on subjective judgments, which at times are in conflict with statistical inference.

What this adds to what was known

How often and why results from randomized clinical trials are inconclusive, and whether there is a concordance between statistical inferences and investigators global judgments in phase III randomized controlled trials is not known. This is the first empirical study to show the reasons for inconclusive findings.

What is the implication, what should change now

Trial design should not rely on an intuitive approach but should include a detailed rationale for the chosen effect size, ideally based on systematic review of the existing evidence on the topic.

Introduction

In the conduct of randomized controlled trials (RCT), ethical and scientific principles require a reasonable expectation that the research questions will be answered, thus contributing to general knowledge resulting in societal benefit.[1] The Declaration of Helsinki states that clinical trials must be designed to facilitate successful completion and therefore prohibits “unethical exposure of participants to the risk and burdens of human research”.[2] Trials that fail to answer the questions they were designed to answer – inconclusive trials – are contrary to one of the key rationales for RCTs, namely to resolve disputes about competing treatments alternatives.[3–7] These trials generate little new knowledge about the relative effects of the treatments, since the research question is left unanswered and the evidence remains consistent with the hypothesis before the trial began. The proportion of RCTs in oncology that generates reasonably conclusive statements is unknown. Theoretically, two reasons can be offered as explanations for publications of inconclusive trials: a) inadequate patient accrual [8] b) optimism bias- an unwarranted belief in the efficacy of new treatments[9]. By overestimating the treatment effect of a particular therapy, trials are designed with insufficient power to detect the actual, smaller treatment effects between tested therapies.

In determining whether a trial provides conclusive results, researchers usually use two inferential approaches: (1) hypothesis-driven, formal statistical or mathematical rules aimed to assess the impact of the experimental treatment on the primary outcome of interest in comparison to a control, and (2) global, subjective, assessments of the relative merits of the treatments, which are based on an integration of various factors including data from non—

primary endpoints and external factors such as treatment toxicity, ease of application, resource use, etc. Thus, the overall judgment on the merits of a trial's results takes into account all relevant observations, many of which may have not been anticipated *a priori*. Information on the extent to which these two inferential approaches influences published conclusions in clinical trials is lacking.

We sought to examine how frequently completed phase III oncology trials generate conclusive results, to quantify the impact of optimism bias on trial results, and to assess the nature of inferential processes underlying the conclusions drawn from a trial. Given the important role of the NCI-funded Clinical Trials Cooperative Groups in advancing cancer care, we elected to focus our study on phase III oncology trials performed by these groups.

Methods

We studied all consecutive phase III RCTs conducted, completed and published between 1955 and 2006 by eight National Cancer Institute-sponsored cooperative oncology groups (NCI-COG): Children's Oncology Group, National Surgical Adjuvant Breast and Bowel Project, Radiation Therapy Oncology Group, North Central Cancer Treatment Group, Gynecology Oncology Group, Eastern Cooperative Oncology Group, Cancer and Leukemia Group B and Southwest Oncology Group. Details regarding publication status, quality and overall distribution of outcomes from these trials have been reported elsewhere.[10–13] Trials for which full protocols were not available or trials with missing data on the expected and observed treatment effects or on patient accrual were excluded. We also excluded trials involving multiple comparisons (12%) because our analysis would have required the re-use of data from the multiple treatment groups, thereby violating the independence criterion for statistical analysis. RCTs that were closed early due to poor accrual (arbitrary defined as the trials that accrued less than 40% of the predicted accrual)[10,14] were also excluded.

Determination of trial conclusiveness

Trials were categorized as conclusive or inconclusive based on two criteria: statistical and investigators' judgments.

Statistical criteria—Conclusive trials included trials with a statistically significant result. Conclusive trials also included “true negative” trials in which the observed effect and the 95% CI were entirely within the pre-determined limit of equivalence. Conversely, inconclusive trials were defined as having the treatment effect and the 95% confidence interval (CI) crossing the line of no difference and both limits of pre-determined equivalence. There is debate about whether results which favor one treatment over another but whose 95% confidence intervals (CI) crosses the line of no difference and only one line of indifference should be considered conclusive or inconclusive.[15] We took a conservative approach and categorized these trials as conclusive, labeling them as “true-negatives” regardless of the direction of the effect. Thus, we considered statistically significant (favoring new or standard treatment) and true negative results as conclusive findings and all other results to be inconclusive. Figure 1 illustrates the methods used to determine the designation of conclusive and inconclusive trials.

Since minimally important (clinically meaningful) treatment differences were rarely specified in the protocols, we based the pre-determined limits of equivalence on published estimates related to what can be considered small, moderate, or large treatment effects in oncology.[16] We defined a small clinically important effect as a hazard ratio or odds ratio (HR or OR) of 0.77 to 0.68, medium effect as HR or OR ranging from 0.51 to ≤ 0.67 and large effect as HR or OR ≤ 0.5 . [16] Thus, we defined HR or OR of 0.77 and 1.3 ($=1/0.77$),

respectively as the smallest clinically important effects for our limits of equivalence (Fig 1). [16]

Judgment of trial investigators—The original researchers' interpretations on the results of their RCT were abstracted from publications. These judgments were categorized using a 6 point scale previously shown to have high face and content validity and high reliability.[17,18],[11,13,19,20] This 6 point scale[17,18] is defined as 1=standard treatment highly preferred to experimental treatment; 2=standard treatment preferred to experimental treatment; 3=about equal, but experimental treatment somewhat disappointing; 4=about equal, but experimental treatment somewhat more successful; 5=experimental treatment preferred to standard treatment; and 6=experimental treatment highly preferred to standard treatment. For our study, this scale was collapsed to 3 categories: I= conclusive statements that standard treatment was superior to experimental treatments (score 1, 2); II= conclusive statements that experimental treatment was superior to standard treatments (score 5, 6); III= indeterminate statements (score 3, 4). The latter category could reflect investigators' beliefs that treatments were either truly equal or that the trial results were inconclusive. However, investigators' statements rarely allowed dissecting between these two possibilities. Hence, our final categorization of trials as conclusive or inconclusive was based on concordance between statistical inference and researchers' global judgments (see next section). Categorization of eligible trial comparisons by statistical calculations and investigators' judgments are depicted in Figure 2.

Concordance between statistical calculations and investigators' judgments—

Several definitions were used to establish concordance between statistical calculations and investigator judgment. We defined concordance *a priori* as those in which the investigator's judgment was in category I (score 1 and 2) and II (score 5 and 6) and the trial was "conclusive" as determined by statistical significance. Since a higher burden of proof is required for acceptance of the alternative hypothesis compared to failing to reject the null hypothesis[21], we also recorded concordance if the investigators' judgments in category I (= conclusive statements that standard treatment was superior to experimental treatments; scores 1 and 2)) and category III (= indeterminate statements; scores 3 and 4) matched with the results that were statistically significant "true negative". The results were also deemed concordant if category III judgments were matched with true negative and inconclusive results according to statistical criteria [15]. All other combinations of statistical calculation and investigator judgment were considered discordant. Details of this classification scheme are presented in Figure 3.

Reasons for inconclusive trials and discordance between statistical calculations and investigators' judgments—

To try to explain why some trials failed to answer their research question, we extracted data from research protocols, published reports and supplementary information obtained from the NCI-COG's statistical offices. Extracted data included expected and observed treatment effects, predicted and actual patient accrual, pre-trial power calculations, pre-trial estimates of treatment effects (the expected difference between the treatments) and sample size calculations. The expected (planned) treatment effects for the primary outcome were extracted from protocols and observed treatment effects with 95% CIs were extracted from publications.

Treatment effects were typically expressed either as relative effects (e.g. HR of 0.75 by 3 years) or as absolute (e.g. 10% difference in survival at 3 years) effects. To allow comparisons between different ways of reporting predicted and observed effects, differences expressed in absolute terms were converted into relative effects.

All results were normalized for reduction of bad outcomes (i.e. HR or OR<1 indicates better outcomes). To obtain a more intuitive comparison of expected versus observed treatment effects, given that lower HR or OR indicates a larger treatment effect, we expressed the results as the inverse ratio of HR or OR calculated using the following formula:

$$\text{ratio of HR or OR} = \frac{1/\text{expected HR or OR}}{1/\text{observed HR or OR}} = \frac{\text{Observed HR or OR}}{\text{Expected HR or OR}}$$

For example, an expected HR or OR = 0.4 and observed HR or OR = 0.8 would correspond to a ratio of HR/OR = 2, meaning that the investigator expected the outcomes to be twice as good as what they actually observed. Since all outcomes were normalized as bad outcomes, this means that when observed (HR or OR)/(expected HR or OR) >1, investigators detected smaller treatment effects than they had planned for. Similarly, a ratio of HR or OR=1, means that the expected treatment effect was identical to the observed one, and a ratio of HR or OR<1 would indicate that investigators observed a larger effect than they expected to detect. Although all trials included in our analyses were considered complete, theoretically the use of ratio may underestimate our findings, because the distribution of observed effects in statistically inconclusive studies will likely be closer to one than those for conclusive studies. Nevertheless, we believe the use of ratio represents the most intuitive measure to address our hypothesis. The appendix provides a distribution of the results without employing division of expected with observed results.

In most cases, researchers did not postulate a single therapeutic effect but, rather, gave a range of treatment effects. Therefore, to determine the lowest expected treatment effect which would meet the criterion of minimally important differences within the range of postulated treatment effects, we extracted the α and β values and their corresponding z-values for each trial. We then calculated the adjusted expected HR or OR by multiplying the originally stated expected HR or OR by the following formula (where Z represents the standard Z-statistic):

$$M = \frac{Z_{(1-\frac{\alpha}{2})}}{Z_{(1-\frac{\alpha}{2})} + Z_{(1-\beta)}}$$

These adjusted expected effects were then compared with the observed effects from the publications. In addition, the ratio of the predicted and actual patient accrual number was used to determine if poor accrual might have contributed to inconclusive results.

Data on predicted and observed treatment effects, power calculations, and accrual statistics were also used to explain discordance between statistical calculations and investigator judgments. This was further supplemented by a qualitative analysis focusing on those trials in which the investigators' judgments and the statistical inference were very different. We looked in particular for any statements to support this divergence.

Mann-Whitney U non-parametric test statistics was used for all comparisons. All analyses were done using the SPSS statistical software package.[22] This research was approved by the USF IRB (No 100449).

Role of the funding source

This study was supported by US National Institute of Health/Office of Research Integrity grants: #: 1 R01 NS044417-01, 1 R01 NS052956-01, 1R01CA133594-01. Sponsors had no role in design, data collection, data analysis, or interpretation of this study, and were not involved in writing this report or in the decision to submit it for publication.

Results

624 phase III trials by the 8 cooperative oncology groups between 1955 and 2006 were identified. This yielded 781 separate comparisons. 45% (353/781) lacked data on expected or observed effect sizes or on patient accrual, and were excluded from analysis. An additional 12% (53/427) were excluded as they involved multiple treatment comparisons i.e. we retain only one, clinically or statistically, most significant comparison in the analysis. Therefore, this report is based on 374 comparisons (359 trials) enrolling 150,232 patients (Figure 2).

Of these 374 comparisons, 27% (102/374) of the results were statistically significant with 10% [10/102] favoring the standard treatment and 90% [92/102] favoring the experimental treatment. The remaining 73% of the comparisons were statistically non-significant (272/374); of these, 59% (160/272) were judged to be true negative and 41% (112/272) were inconclusive (see figure 2). Among 160 “true negative” results, 56% of the point estimates favored experimental and 44% favored standard treatments. Thus, 70% (262/374) of the comparisons in RCTs answered their research questions successfully according to statistical criteria.

Figure 2 also shows the results according to the original researchers’ judgments. New treatments were judged to be superior in 29% (108/374), while standard treatment was favored in 44% (165/374) of trials. The researchers could not decide between standard and new treatments in 27% (101/374) of comparisons. Figure 3 compares statistical criteria and investigators’ judgments. When statistical criteria were combined with investigators’ judgments, the overall number of inconclusive comparisons dropped to 10% (36/374) [comprising, for example, 32% (36/120) of statistically inconclusive results]. However, discordance between statistical criteria and investigators’ judgments was noted in 25% (95/374) of comparisons. In 19 of these 95 instances, investigators concluded that one treatment was superior or equal to another, while the statistical findings supported the opposite conclusion. The most frequent reasons for this discordance were the toxicity with the new treatment, lack of meaningful clinical benefits, and ease of drug administration (see Appendix).

Ideally, in a successfully completed trial, the observed effect should match the expected treatment effects ($HR_{observed}/HR_{expected}$), and the number of accrued patients should match the planned sample size; ($SS_{actual}/SS_{predicted}$) these theoretical ratios are depicted in Figure 4. Theoretically, we would expect to find the majority of conclusive results in quadrant B ($SS_{actual}/SS_{predicted} > 1$, $HR_{observed}/HR_{expected} < 1$). This, however, occurred only in about 12% of completed trials. Similarly, we would expect that quadrant C ($SS_{actual}/SS_{predicted} < 1$, $HR_{observed}/HR_{expected} > 1$) would contain the majority of inconclusive findings. While this was to some extent true (Figure 4), it is actually quadrant D ($SS_{actual}/SS_{predicted} > 1$, $HR_{observed}/HR_{expected} > 1$) that contained by far the largest number of inconclusive results. This finding likely is the result of “optimism bias” – when designing the trial, the investigators overestimated the treatment effect that could be detected. As Figures 4a and b illustrates, researchers consistently overestimated the treatment effect across all trials. In about 5% of completed trials the observed treatment effect was sufficiently large to allow a statistically significant result despite a sample size that did not attain the target (quadrant A). Figure 5 shows the distribution of expected vs. observed results in relation to statistical criteria. The median ratio of observed over expected HR was 1.34 (mean 1.53; range 0.19–15.40) in conclusive trials compared to 1.86 (mean 2.23; range 0.6–12) in inconclusive studies ($p < 0.0001$). Figure 6 shows the distribution of expected vs. observed results in relation to patient accrual. The patient accrual did not significantly affect investigators over-optimistic estimates. Neither treatment effects

(observed vs. expected) nor sample size (planned vs. actual accrual) have changed over time (see Appendix). Additional analyses using adjusted expected HR did not change these results in any important ways.

The appendix provides additional results of interest.

Discussion

In over 50 years of phase III trials completed and published by 8 cooperative oncology groups, 30% of comparisons were found to be inconclusive by statistical criteria. As depicted in figure 7, our review of research protocols showed that in many trials the design was disproportionally driven by unrealistic expectations. In fact, the trial results matched the expectation of the size of treatment effects in only 12% of trials. In an additional 5% of comparisons, researchers' optimism "paid off" in arriving at conclusive answers about the treatment benefit despite accruing fewer patients than originally planned. So, in total the investigator's expectations were attained in about 17% of completed trials. As disappointing as this figure is, it probably represents an underestimate since, a) the distribution of observed effects in statistically inconclusive studies is likely closer to one that those for conclusive studies, and b) trials that accrued <40% of their sample size were excluded and such trials have virtually no likelihood of meeting investigators efficacy expectations.

This failure to provide relatively convincing answers to questions addressed in trials may represent a waste of precious resources and a possible ethical breach of the contract with patients who expect that answers from clinical trials will help future patients.[23] Although the original trial designs of the comparisons in this study were often based on overestimated effect sizes, this tendency was most exaggerated in inconclusive comparisons. We conclude, therefore, that optimism bias – the unwarranted belief in the efficacy of new therapies – contributed significantly to the inconclusiveness of trials. As a consequence, the inconclusive comparisons were also noted to have smaller average predicted accrual targets. When we examined the differences between the effects that investigators hoped to discover and the effects they detected, we found that the investigators overestimated the treatment effect by 86% on average in inconclusive trials. Alternatively, the investigators may have actually estimated the treatment effect correctly, but to function in logistical constraints of patient recruitment or funding they might have adjusted the power calculations. That is, realistic calculations may show that the required accrual is difficult or impossible to accomplish given an existing fixed budget. The extent to which this occurs is not known but reflects seriously upon the researchers' fiduciary obligations toward patients and the funder.

Trial design cannot rely on an intuitive approach but should include a detailed rationale for the chosen effect size, ideally based on systematic review of the existing evidence on the topic.[24] The latter may need to include previous human trials, but also synthesis of animal data[25,26] as well as the use of registry data to derive a realistic effect size. Furthermore, awareness of the prevalence of optimism bias and its impact on trial design may encourage more realistic, albeit smaller treatment effects estimates.

Our study also identified that investigator's judgments play an important role in interpreting the results of a trial. Consistent with Kuhn's notion that scientific interpretation always represents a mixture between objective and subjective factors[27,28], the original researchers typically supplemented an explicit formal reasoning process through statistical calculations with an implicit, subjective, pragmatic inferential approach provided by their judgment. However, in approximately 25% of cases, we found these two inferential processes to be at odds with each other. Researchers' judgments were typically based on the evaluation of outcomes *after* the trial that also included considerations not anticipated at the

time of the trial design, and therefore could have been biased. Such modifications to the originally proposed hypotheses are considered epistemologically less valuable than hypotheses-driven predictions[29]. However, our previous evaluation of the quality of these trials according to methodological dimensions that have been empirically linked to bias (such as quality of randomization procedure, the magnitude of drop-outs, employment of blinding, use of intention-to-treat analysis etc)[30] showed the NCI COG trials of high quality,[31] indicating no obvious bias.

The seemingly discordant results observed 25% of the comparisons in our study can be explained primarily as a result of statistical reasoning almost exclusively relying on the assessment of a single dimension (i.e. effects of treatment on the primary outcome), while investigators evaluate treatment effects across many dimensions. That is, unlike formal inferential processes, the investigator's global judgment is content and context dependent[32]. Of interest, recently Booth et al [33] analyzed differences between conclusions of the abstracts of RCTs presented at major oncology meetings with the subsequent final analyses published in full papers. They found that in about 10 percent of the trials, there was a reversal in the researchers' conclusions about the efficacy of treatment when the abstracts were compared with the final reports.

Our findings have important implications not only for researchers drawing conclusions about the results of their own trial, but also for the interpretation of evidence of relevance to clinical practice and policy decision-making (e.g. systematic reviews, guidelines recommendations and measurements of quality of care). The results indicate that subjective, intuitive judgments play a large role in drawing inferences, as investigators balance those factors that never anticipated before the trial began (e.g., toxicity, costs, patients' subjective reactions toward treatment, setting of treatment etc) with those factors that went into generation of research hypothesis. This means that the overall value of the same treatment may possibly be differently judged by different research teams working in different social milieus. In turn, this implies that it may be difficult to achieve high levels of interpretative reproducibility of evidential findings when they are transferred to different settings- what is considered "inconclusive trial" by one group of investigators may be deemed otherwise by the other. [27,28]

Our research has several limitations. First, we included only trials that were completed. Therefore, trials that were closed early due to poor accrual were excluded because of the lack of reported results. Thus, our analysis likely overestimated the proportion of all initiated trials that reach a definitive conclusion, and may have underestimated the effect of poor accrual. Similarly, since publication was also an inclusion criteria, the number of inconclusive trials may be underestimated as these trials may be less likely to produce publications. Second, we derived investigators' judgments from statements in their publications and not by interviewing them. The latter was impossible, not least because the cohort of trials we studied span several decades. We did, however, employ procedures for deducing investigators' judgments using well established methods that have been found to be reproducible and valid.[10,11,14,17–19,34–36] Third, our definition of "inconclusive" trials is somewhat arbitrary, being based on what some people consider a minimally important difference[16], which may not be shared by the original researchers. Nevertheless, the criteria we chose are rather conservative, and if anything, biased our analysis toward accepting the null hypothesis. Fourth, we were unable to extract the necessary data from all trials. However, there were no indications of any systematic reasons for missing data. Most missing data relate to older trials and appear to be missing at random, perhaps due to re-organization of the cooperative groups over time. Fifth, our concordance criteria do not address the issue of whether some of the researchers' judgments rendered were correct or not. This could not be answered within the confines of our study. Only by evaluating

findings in subsequent research that confirms or contradicts the conclusions derived in studies we examined can the veracity of both statistical criteria and/or investigators' judgments be addressed. Finally, the issue of inferences and whether the findings should be considered conclusive or not is trial and context-dependent; these judgments are best rendered at the time when the trial was designed and published as they depend on the type(s) and seriousness of outcomes, design of the trial, and whether other existing treatments in the field are believed to be superior/inferior to those tested in a given trial. Although this is undoubtedly true, our very conservative criteria, superiority design and mortality outcomes used in the vast majority of trials indicate that, if anything, our results represent underestimate of the true number of inconclusive trials.

In conclusion, our results indicate either the optimism bias, or overly optimistic power calculation as the primary reason for inconclusive results. We also showed that formal statistical inference was sufficient to answer the research questions in about 70% of clinical trials. This inference is often combined with subjective judgments by the researchers and taken together these two forms of reasoning helped produce answers for the vast majority (90%) of research questions asked in completed, published phase III oncology cooperative group trials in the years 1955–2006. This study highlights the need for tools to allow more accurate prediction of treatment effects and for an emphasis in early trial planning phases to identify and curb optimism bias. Our findings promote the need for clarity in reporting the basis of trial conclusions, and the extent to which these conclusions reflect statistical calculations or researcher judgment. Policy makers should encourage investigators to perform a systematic review of the existing relevant evidence to better inform the chosen treatment effect.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We sincerely thank Dr. Gordon Guyatt for providing critical feedback on the content of this manuscript.

References

1. <http://www.wma.net/e/policy/b3.htm>. [cited; Available from:
2. World Medical Association. World Medical Association Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. 2007. [cited 2008 20th October]; Available from: <http://www.wma.net/e/policy/b3.htm>
3. Chalmers I. Well informed uncertainties about the effects of treatments. *BMJ*. February 28; 2004 328(7438):475–6. [PubMed: 14988164]
4. Edwards SJL, Lilford RJ, Braunholtz DA, Jackson JC, Hewison J, Thornton J. Ethical issues in the design and conduct of randomized controlled trials. *Health Technol Assessment*. 1998; 2(15):1–130.
5. Bradford Hill A. Medical ethics and controlled trials. *BMJ*. 1963; 2:1043–49. [PubMed: 14065059]
6. Djulbegovic B. Acknowledgment of Uncertainty: A Fundamental Means to Ensure Scientific and Ethical Validity in Clinical Research. *Current Oncology Reports*. 2001; 3:389–95. [PubMed: 11489238]
7. Djulbegovic B. Articulating and responding to uncertainties in clinical research. *J Med Philos*. 2007 Mar–Apr; 32(2):79–98. [PubMed: 17454416]
8. Prescott RJ, Counsell CE, Gillespie WJ, Grant AM, Russell IT, Kiauka S, et al. Factors that limit the quality, number and progress of randomised controlled trials. *Health technology assessment (Winchester, England)*. 1999; 3(20):1–143.

9. Chalmers I, Matthews R. What are implications of optimism bias in clinical research? *Lancet*. 2006; 367:449–50. [PubMed: 16473106]
10. Djulbegovic B, Kumar A, Soares HP, Hozo I, Beppler G, Clarke M, et al. Treatment success in cancer: new cancer treatment successes identified in phase 3 randomized controlled trials conducted by the National Cancer Institute-sponsored cooperative oncology groups, 1955 to 2006. *Arch Intern Med*. 2008 Mar 24; 168(6):632–42. [PubMed: 18362256]
11. Kumar A, Soares H, Wells R, Clarke M, Hozo I, Bleyer A, et al. Are experimental treatments for cancer in children superior to established treatments? Observational study of randomised controlled trials by the Children's Oncology Group. *Bmj*. 2005 Dec 3; 331(7528):1295.
12. Soares HP, Daniels S, Kumar A, Clarke M, Scott C, Swann S, et al. Bad reporting does not mean bad methods for randomised trials: observational study of randomised controlled trials performed by the Radiation Therapy Oncology Group. *Bmj*. 2004 Jan 3; 328(7430):22–4.
13. Soares HP, Kumar A, Daniels S, Swann S, Cantor A, Hozo I, et al. Evaluation of new treatments in radiation oncology: are they better than standard treatments? *JAMA*. 2005 Feb 23; 293(8):970–8. [PubMed: 15728168]
14. Soares HP, Kumar A, Daniels S, Swann S, Cantor A, Hozo I, et al. Evaluation of New Treatments in Radiation Oncology: Are They Better Than Standard Treatments? *JAMA*. February 23; 2005 293(8):970–8. [PubMed: 15728168]
15. Korn EL, Mooney MM, Abrams JS. Treatment success in cancer. *Archives of internal medicine*. 2008 Oct 27; 168(19):2173. author reply -4. [PubMed: 18955658]
16. Bedard PL, Krzyzanowska MK, Pintilie M, Tannock IF. Statistical power of negative randomized controlled trials presented at American Society for Clinical Oncology annual meetings. *J Clin Oncol*. 2007 Aug 10; 25(23):3482–7. [PubMed: 17687153]
17. Colditz GA, Miller JN, Mosteller F. Measuring gain in the evaluation of medical technology. The probability of a better outcome. *Int J Technol Assess Health Care*. 1988; 4:637–42.
18. Colditz GA, Miller JN, Mosteller F. How study design affects outcomes in comparisons of therapy. *I:medical. Stat Med*. 1989; 8:441–54. [PubMed: 2727468]
19. Djulbegovic B, Lacevic M, Cantor A, Fields KK, Bennett CL, Adams JR, et al. The uncertainty principle and industry-sponsored research. *Lancet*. 2000 Aug 19; 356(9230):635–8. [PubMed: 10968436]
20. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? *Jama*. 2003 Aug 20; 290(7):921–8. [PubMed: 12928469]
21. Hozo I, Schell MJ, Djulbegovic B. Decision-making when data and inferences are not conclusive: risk-benefit and acceptable regret approach. *Seminars in hematology*. 2008 Jul; 45(3):150–9. [PubMed: 18582621]
22. SPSS Inc. *SPSS Base 16.0 for Windows User's Guide*. 16.0. Chicago, IL: SPSS Inc; 2008.
23. Penslar R. *IRB Guide*. 1993. [cited 2008 20th October]; Available from: http://www.hhs.gov/ohrp/irb/irb_guidebook.htm
24. Clarke M. Doing new research? Don't forget the old. *PLoS medicine*. 2004 Nov.1(2):e35. [PubMed: 15578106]
25. Hackam DG, Redelmeier DA. Translation of research evidence from animals to humans. *Jama*. 2006 Oct 11; 296(14):1731–2. [PubMed: 17032985]
26. Sandercock P, Roberts I. Systematic reviews of animal experiments. *Lancet*. 2003; 360:586. [PubMed: 12241927]
27. Kuhn, T. *The structure of scientific revolutions*. 2. Chicago: The University of Chicago Press; 1970.
28. Kuhn, T. Objectivity, value judgement, and theory choice. In: Klemke, ED.; Hollinger, R.; Rudge, DW., editors. *Introductory readings in the philosophy of science*. New York: Prometheus Books; 1998. p. 435-50.
29. Lipton P. Testing Hypotheses: Prediction and Prejudice. *Science (New York, NY)*. January 14; 2005 307(5707):219–21.
30. Higgins, JPT.; Green, S., editors. *Cochrane Handbook for Systematic Reviews of Interventions*; Section 6. Chichester, UK: John Wiley & Sons, Ltd; 2008. Assessment of study quality.

31. Djulbegovic B, Kumar A, Soares HP, Hozo I, Beppler G, Clarke M, et al. Treatment success in cancer: reply. *Archives of internal medicine*. 2008 Mar 24;168:2173–4. [PubMed: 18955658]
32. Evans, JSTBT. *Hypothetical thinking. Dual processes in reasoning and judgement*. New York: Psychology Press: Taylor and Francis Group; 2007.
33. Booth CM, Le Maitre A, Ding K, Farn K, Fralick M, Phillips C, et al. Presentation of Nonfinal Results of Randomized Controlled Trials at Major Oncology Meetings. *J Clin Oncol*. 2009 Jul 20.
34. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of Funding and Conclusions in Randomized Drug Trials: A Reflection of Treatment Effect or Adverse Events? *JAMA*. August 20; 2003 290(7):921–8. [PubMed: 12928469]
35. Gilbert, JP.; McPeck, B.; Mosteller, F. Progress in surgery and anesthesia: benefits and risks of innovative therapy. In: Bubker, JP.; Barnes, BA.; Mosteller, F., editors. *Costs, risks and benefits of surgery*. New York: Oxford University Press; 1977. p. 124-69.
36. Gilbert JP, McPeck B, Mosteller F. Statistics and ethics in surgery and anesthesia. *Science*. 1977; 198:684–9. [PubMed: 333585]

Figure 1. Classifying results from RCTs as true negative, positive or inconclusive. (Adapted from Alderson, P. BMJ 2004; 328: 476–477)

Figure 2. Flow chart describing the selection of studies. The figure also shows the results according to statistical criteria and treatment success according to investigators' judgments.

Figure 3. Treatment success according to statistical criteria and investigators' judgments (concordance versus non-concordance) (see text for details). (S= standard treatment; E=experimental treatment)

Figure 4.

a) 4b). A ratio of actual/predicted patient accrual vs. observed/expected treatment effects. HR-hazard ratio, OR-observed ratio. Note that in Fig 4b) HR represents HR or OR. Results are shown as the ratio of the observed HR or OR divided by the expected HR or OR. All outcomes are normalized as “bad” (e.g. deaths instead of survival). Thus, a lower HR or OR indicates a larger treatment effect on a bad outcome. For example, an expected HR of 0.4 and an observed HR of 0.8 would correspond to a ratio of 2, meaning that the investigator expected the treatment effect to be twice as good as they actually observed (or, that investigators observed twice as smaller effect than they had planned for). A ratio of 1 means that the expected treatment effect was the same as the observed one, and a ratio below 1 would indicate that the investigators observed a larger effect than expected (the log transformation is used for better display).

Figure 5.

The distribution of treatment effects (expressed as hazard or odds ratio) and shown as the percentage difference between expected and observed results. The line at zero indicates that observed results perfectly matched expected treatment effects. The negative numbers indicate the extent (percentage) above which expectation exceeded the observed effects. The positive findings indicate the percentage above which actually observed results exceeded expectations.

Figure 6.

The distribution of expected vs. observed results in relation to patient accrual. The line at zero indicates that observed results perfectly matched expected treatment effects. The negative numbers indicate the extent (percentage) above which expectation exceeded the observed effects. The positive findings indicate the percentage above which actually observed results exceeded expectations.

Figure 7. Distribution of categories of expected versus observed treatment effect across trials. As it can be seen, there is a tremendous discrepancy over the entire range of treatment effects between what investigators expected and what they actually observed.