The need to consider the wider agenda in systematic reviews and meta-analyses

As well as focusing on a precise question, systematic reviewers also need to consider the whole research programme for the interventions under study, argue **John Ioannidis** and **Fotini Karassa**

The problem: a wide research programme

A powerful company develops a promising new blockbuster. The more diseases and conditions the drug can get approved for, the greater the sales. Therefore the company launches trials for many different indications, as its clinical research programme unfolds. Independent committees set interim analysis and appropriate stopping rules for these trials, to avoid harming people on placebo for too long if the drug proves effective. Then, some trials start showing statistically significant benefits, so they are stopped early and the drug gets approved for those indications. Suppose all findings and all results are reported—that is, no reporting bias[1] operates. This sounds like the ideal success of honest drug development and clinical investigation. However, it can be shown that the drug is less effective than these trials suggest—and sometimes not effective at all. Why?

Explanations for the problem

Two reasons explain this paradox. Firstly, the drug is tested for many indications and secondly, the first trials have been stopped early. The first reason refers to the breadth of the evidence. The second refers to the timing and depth of the evidence.

Firstly, multiplicity of analyses[2]: if we test a totally ineffective drug for 20 indications, by chance it is likely to show a significant effect ($P<0.05$) for one of them. If we test 10 different independent outcomes in each indication, then we expect at least one outcome to be statistically significant for almost half of the indications even without any reporting bias.[3]

Secondly, early stopping: trials stopped early because of perceived effectiveness give inflated estimates of the treatment effect.[4-6] An empirical investigation of 91 early stopped trials showed that on average the true effect was only 70% of what these trials suggested and less than half when trials stopped with fewer than 200 or so events.[4] This inflation of the effect is an example of regression to the mean[6]: when we select results because they cross a statistical significance threshold, the effect sizes are expected to be inflated. If we test the drug again it will not be as effective as in the first, early stopped trials. For the same reason, if several trials are launched on the same, similar, or different indications, the ones that stop early and get published first may overestimate the treatment effect. Trials that don't see such a large effect don't hit stopping rules. They continue follow-up and thus completion and publication are delayed.[7]

That does not mean that when a drug is proved effective in an early stopped trial it is completely ineffective. However, if it is only half as effective as we thought its benefit-risk ratio and cost-effectiveness may not justify its use or they may justify it in far fewer patients than we thought.

The solution: examining the whole programme

The solution to this problem is to examine the whole clinical research programme for any new intervention. Systematic reviews and meta-analyses have traditionally focused on one intervention at a time. Moreover, they depend on published data, or, at the most, they try to unearth data from completed but unpublished trials.

Systematic reviewers should be aware of the breadth, timing, and depth of all the evidence on the drug they are reviewing. They should take into account the extent of diversity in the research programme, use caution with early stopped trials, and consider the depth of the total evidence for each indication. The total evidence includes published trials; completed (and analysed) but unpublished trials; and continuing trials. Obviously, analysed outcome data can be obtained only for completed analysed trials. But it is important to know whether the data in the meta-analysis calculations represent all, most, or only a small portion of the full programme and whether the available data suffer from early stopping.

We show here two examples to illustrate our points. The first example shows the importance of considering the breadth and depth of the evidence for anti-tumor necrosis factor (anti-TNF) agents, a class of expensive drugs tested for a large number of indications. The second example shows the importance of considering also the timing of the evidence for one of the most expensive cancer drugs, bevacizumab.

Anti-TNF agents

Anti-TNF agents have been tested for seven different indications in immune-mediated inflammatory diseases (see appendix 1 on bmj.com) since 1998, and until April 2009 about 1 100 000 patients have been treated with these agents for approved indications.[8] TNF blockers cost about $20 000 (£13 000, €15 500) per patient year.[9] Annual sales exceed $10bn.[10]

We searched MEDLINE and ClinicalTrials.gov (April 2010, search details available from authors) for randomised trials evaluating any selective anti-TNF agent against placebo or another active treatment in patients with immune mediated inflammatory diseases. In ClinicalTrials.gov (see appendix 2 on bmj.com) we identified 54 completed published trials (20 942 patients), 41 (9984 patients) completed unpublished trials, 32 (11 465 patients) ongoing closed trials (9 of these have already published some results), and 43 open recruiting trials (expected enrollment 10 809 patients). There are also 6 registered not yet open trials (expected enrollment 794 patients), 10 terminated trials (847 patient; 1 trial has been published), and 2 suspended trials (1078 patients). Among the 188 trials, only 5 compared head to head different anti-TNF agents (none published yet), and another 32 trials compared an anti-TNF agent with one or more active comparators (only 4 published to-date). All other trials have been comparisons against placebo or no treatment.

Trials stopped early because of perceived effectiveness give inflated estimates of the treatment effect

Overall, the published trials represent only 34% of the total evidence (46% if we exclude the open trials). The percentage of published trials varies for each drug-intervention pair (see appendix 3 on bmj.com). With six agents and seven indications there are 42 drug-indication pairs; trials are available for 28 of them and licensing approval has been granted for 22 drug-indication pairs. For six of these 22 less than half of the registered trials' evidence is published and for several others the data are also limited (table). Inferences about the exact treatment effects of these agents require extra caution.

We found 34 systematic reviews and meta-analyses of trials evaluating the effectiveness of anti-TNF agents in immune mediated diseases (see appendix 4 on bmj.com; PubMed, April 2010). Of those, 12 deal with a single agent and single indication; 22 deal with several agents for a single indication. None considered several indications.

Bevacizumab

Bevacizumab is the first extensively studied anti-angiogenesis monoclonal antibody. It is expensive, with annual sales of $2.7bn (£1.8bn, €2bn).[11] A search in ClinicalTrials.gov returns an amazing number of 1093 registered studies, the vast majority (922) in cancer. Bevacizumab is being tested in nearly all types of malignancy.

Bevacizumab was first approved for use in metastatic colorectal cancer (Food and Drug Administration 2004, European Medicines Agency 2005) based on the results of the E3200 trial, which was stopped very early because of a major survival benefit (hazard ratio 0.66, 95% confidence interval, 0.54 to 0.81). In 2006 it was approved for lung cancer, based on another early stopped trial with significant survival benefit (HR 0.79, 0.67 to 0.92). In 2007 it was approved by the EMA for metastatic renal cancer, based on the results of another trial also stopped early because of benefit in disease progression (prolonged disease remission). Other approvals followed for breast cancer in 2007-8 and glioblastoma in 2009. The FDA approval for breast cancer was given against the recommendations of outside experts, who cautioned that no survival benefit had been documented. The approval for glioblastoma was given without data from phase III randomised trials.

Then, in mid-2009, results from the largest trial of bevacizumab were presented at the American Society of Clinical Oncology meeting. Among 2710 randomised patients with colorectal cancer, bevacizumab improved neither survival nor disease free survival. Actually, a benefit was seen in the first year with a hazard ratio of 0.60 (as in the early stopped trial that had led to the initial enthusiasm), but it was transient and totally reversed with longer follow up.

In Clinicaltrials.gov (March 2010) there are 26 closed phase III cancer trials of bevacizumab *v* placebo (see appendix 5 on bmj.com). Of those, 9 trials (7234 patients) have been completed and published, 3 trials (4669 patients) have had sur-

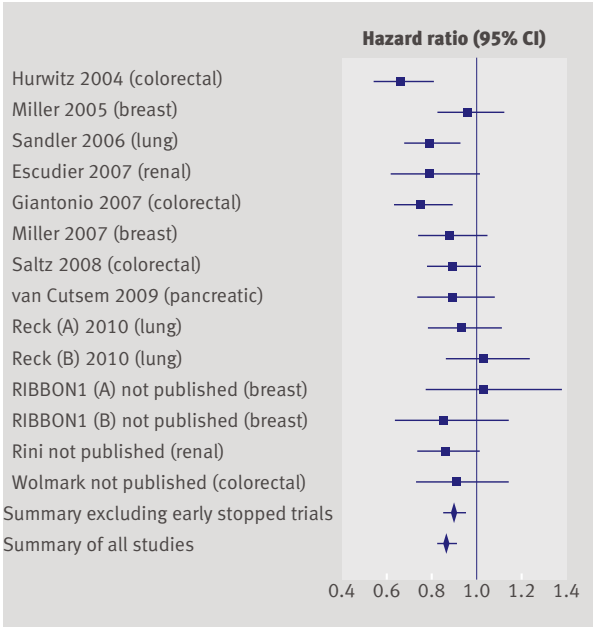
Sample size of published randomised clinical trials/sample size of all randomised clinical trials (percentage) for each anti-TNF agent for each immune-mediated inflammatory disease indication. Data relate only to trials registered in ClinicalTrials.gov

Autoimmune disease indication	Infliximab	Etanercept	Adalimumab	Certolizumab pegol	Golimumab	Onercept
Rheumatoid arthritis	3823/4960 (77.1)*	1763/9127 (19.3)*	2093/9265 (22.5)*	1821/4199 (43.4)*	1720/3729 (46.1)*	NTR
Plaque psoriasis	1462/2796 (52.3)*	1874/2841 (65.9)*	1630/2147 (75.9)*	0/251 (0)	NTR	0/NP (0)
Psoriatic arthritis	200/355 (56.3)*	205/245 (83.6)*	415/415 (100)*	0/390 (0)	405/405 (100)*	NTR
Ankylosing spondylitis	526/640 (82.2)*	124/896 (13.8)*	397/397 (100)*	0/315 (0)	356/356 (100)*	NTR
Juvenile idiopathic arthritis	122/182 (67)	12/173 (6.9)*	171/171 (100)*	NTR	NTR	NTR
Crohn's disease	1411/1514 (93.2)*	NTR	1478/2062 (71.7)*	1090/1779 (61.3)*	NTR	NTR
Ulcerative colitis	728/1438 (50.6)*	NTR	0/1349 (0)	NTR	0/2,702 (0)	NTR

Some anti-TNF agents have been tested also for other, non-autoimmune disease indications, but these are not included here.

NTR: no trials registered; NP: not provided

*Approved indications in the USA (FDA) and Europe (EMA)



Meta-analysis forest plot for survival with bevacizumab *v* control in trials of patients with cancer. Each trial is shown by its year of publication, name of first author, and type of malignancy as well as the hazard ratio for survival and 95% confidence interval. Also shown are summary estimates including all trials and excluding the three trials stopped early, which showed large treatment benefits (Hurwitz 2004, Sandler 2006, Escudier 2007)

vival results reported at meetings, and 14 trials (10 724 patients) have not reported survival results yet. Based on closed phase III trials, the published evidence plus that reported at meetings represents 83%, 57%, 74%, and 62% of the data for renal, breast, lung, and colorectal cancer data respectively, while survival data on ovarian, prostate, gastric, and gastrointestinal stromal tumours are still pending (March 2010). The figure shows a forest plot of the 12 studies with available data on survival (14 comparisons). The summary hazard ratio is 0.87 with 95% confidence interval of 0.82 to 0.91. Testing for between-study heterogeneity gives P=0.14 (Q=18.95, 13 df and I^2=31%), suggesting modest heterogeneity. However, it is obvious that the three first trials that were stopped early had larger effect sizes. The summary hazard ratio for the other 9 trials (11 comparisons) is only 0.90 (95% confidence interval 0.85 to 0.95), and results are very consistent across these trials (Q=8.63, 10 df, p=0.6, I^2=0%). When the early stopped trials are excluded, the summary effects for survival for each type of cancer are also consistent between themselves (Q=3.84, 4 df, p=0.3, I^2=0%).

Even though no heterogeneity is detected when the trials that were stopped early are excluded, some modest heterogeneity may be missed because of the limited power of heterogeneity tests.[12] Even with this caveat, the emerging evidence suggests that bevacizumab confers a survival benefit in patients with cancer that is small on average (10% relative risk reduction) and is unlikely to be large for any specific tumour type. Guideline developers, drug approval agencies, and insurance companies may debate on whether this evidence is sufficient for endorsing this drug.

Such a debate should consider also cost, and other outcomes, both of effectiveness, such as disease free progression, and harms, such as hypertension, perforation, haemorrhage, proteinuria, and wound complications, all of which are increased with bevacizumab. Even if the large benefits from early stopped trials are assumed to be realistic, bevacizumab was barely cost effective in cost utility analyses.[13] With much smaller benefits and better documented harms,[14-16] its prospects are more bleak.

Between June 2009 and March 2010 six meta-analyses of bevacizumab effectiveness for cancer have been published (PubMed search, March 2010; see appendix 6 on bmj.com). All of those deal with a single type of cancer and each considered only one to three phase III trials (up to five trials maximum when they include also smaller phase I/II trials). None of these meta-analyses took into consideration the whole range of bevacizumab trials in cancer or even a modest fraction of it.

Conclusions

In appraising the evidence for new interventions, systematic reviewers should be aware of the breadth, timing, and depth of the wider programme of clinical trials. This includes the diversity of different tested indications, early stopping of trials, and the amount of data in unpublished and ongoing trials. Meta-analyses that ignore this wider agenda could reach narrow, misleading interpretations.

When an intervention shows effectiveness for only a tiny fraction of the tested indications, seemingly promising results may be false positives. Reviewers could consider an overarching meta-analysis unifying data from all indications with similar outcomes, testing whether an overall effect is seen and whether there is compelling evidence for heterogeneity between the trials on different indications, as we have done here for bevacizumab. Such overarching meta-analyses would justify whether evidence should be seen separately for each indication or not.[17]

Evaluation of the wider agenda requires comprehensive trial registries.[18 19] In specialties where trial registration is still rudimentary or non-existent, systematic reviews should be extremely cautious. Sponsors may launch a vast but only partly registered programme of trials. Registration may be eclectic depending on whether national or international agencies require registration. Some drugs may get approved in some countries but not others. For example, antidepressants such as mianserin, milnacipam, and fluoxamine are not approved in the US, but they are approved in other countries where their trials did better.[20] Fragmentation of the programme increases the multiplicity problem. Superimposed selective reporting of analyses and outcomes[1-3] can make anything seem effective.

When a trial is stopped early for effectiveness, this simply means that the treatment effect is not null. However, the effect may be small, much smaller than initially observed. With a small treatment effect the intervention may not have a favourable risk-benefit or cost-utility ratio. Early stopping should be performed sparingly and judiciously, and results from single, early stopped trials are unreliable.[4-6] Furthermore, even when several trials have published their results, it is useful to know how many more have not. Formal approaches can investigate how a meta-analysis of published results would be affected from simulated results of additional trials.[21] Knowing that much of the evidence is still missing should lead to caution.

Finally, most of the clinical research programmes for new interventions rely heavily on placebo controlled comparisons. Ideally, head to head comparisons against other effective agents should also be performed to combine in network meta-analyses.[22-24] However, head to head comparisons are often systematically avoided; companies avoid comparators from other companies.[25]

In summary therefore, although systematic reviewers should ask precise questions, they also need to have a clear view of where each question sits in relation to other questions being asked. A narrow meta-analysis of a single type of comparison for a single indication may miss most of the interesting story about an intervention and its agenda. Decision making can seldom be based on a single systematic review. These issues may have also implications for drug licensing. The licensing process should consider the status of the wider agenda in making appropriate decisions about new agents.

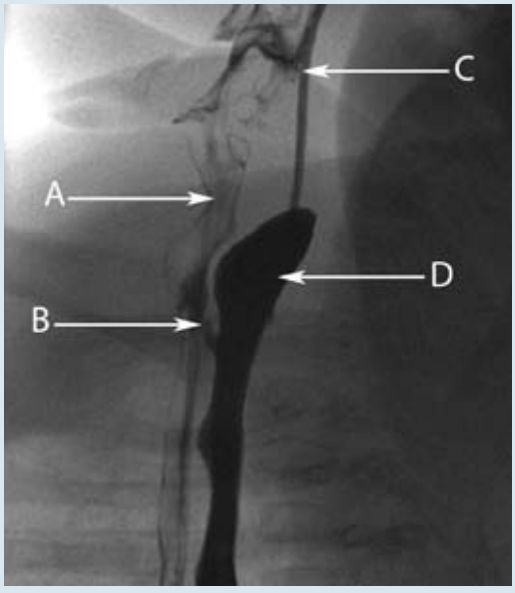
John P A Ioannidis professor and chairman, Department of Hygiene and Epidemiology
Fotini B Karassa lecturer in rheumatology, Division of Rheumatology, Department of Internal Medicine, University of Ioannina School of Medicine, Ioannina, Greece
Correspondence to: J P A Ioannidis, **jioannid@cc.uoi.gr**

Accepted: 21 July 2010
Contributors: JPAI had the original idea and both authors developed the concept; both authors collected and analyzed data; JPAI wrote the paper and FBK critically reviewed it. Both authors approved the final version. JPAI is guarantor.
Funding: None.
Competing Interests: None declared.
Access to data: JPAI had full access to all of the data in the study and can take responsibility for the integrity of the data and the accuracy of the data analysis.
Data sharing statement: Data available from the corresponding author at jioannid@cc.uoi.gr

1 Dwan K, Altman DG, Arnaiz JA, Bloom J, Chan AW, Cronin E, et al. Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One* 2008;3:e3081.
2 Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;2:e124.
3 Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective publication of antidepressant trials and its influence on apparent efficacy. *N Engl J Med* 2008;358:252-60.
4 Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, et al. STOPIT-2 Study Group. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA* 2010;303:1180-7.
5 Hughes MD, Pocock SJ. Stopping rules and estimation problems in clinical trials. *Stat Med* 1988;7:1231-1242.
6 Ioannidis JP. Why most discovered true associations are inflated. *Epidemiology* 2008;19:640-648.
7 Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998;279:281-286.
8 Sfikakis PP. The first decade of biologic TNF antagonists in clinical practice: lessons learned, unresolved issues and future directions. *Curr Dir Autoimmun* 2010;11:180-210.
9 Barra L, Pope JE, Payne M. Real-world anti-tumor necrosis factor treatment in rheumatoid arthritis, psoriatic arthritis, and ankylosing spondylitis: cost-effectiveness based on number needed to treat to improve health assessment questionnaire. *J Rheumatol* 2009;36:1421-8.
10 Abbott. 2009 Annual Report. http://www.abbott.com/annual-reports/2009/downloads/Editorial_section_only.pdf (accessed 6 May, 2010)
11 Genentech Inc., K-10 form. http://www.sec.gov/Archives/edgar/data/318771/000031877109000003/form10-k_2008.htm (accessed 31 March 2010)
12 Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods* 2006;11:193-206.
13 Tappenden P, Jones R, Paisley S, Carroll C. The cost-effectiveness of bevacizumab in the first-line treatment of metastatic colorectal cancer in England and Wales. *Eur J Cancer* 2007;43:2487-94.
14 Ranpura V, Pulipati B, Chu D, Zhu X, Wu S. Increased risk of high-grade hypertension with bevacizumab in cancer patients: a meta-analysis. *Am J Hypertens* 2010; 25 Feb. [Epub ahead of print]
15 Ranpura V, Hapani S, Chuang J, Wu S. Risk of cardiac ischemia and arterial thromboembolic events with the angiogenesis inhibitor bevacizumab in cancer patients: A meta-analysis of randomized controlled trials. *Acta Oncol* 2010; 16 Feb. [Epub ahead of print]
16 Hapani S, Chu D, Wu S. Risk of gastrointestinal perforation in patients with cancer treated with bevacizumab: a meta-analysis. *Lancet Oncol* 2009;10:559-68.
17 Ioannidis JP, Patsopoulos NA, Rothstein HR. Reasons or excuses for avoiding meta-analysis in forest plots. *BMJ* 2008;336:1413-5.
18 Rennie D. Trial registration: a great idea switches from ignored to irresistible. *JAMA* 2004;292:1359-62.
19 Laine C, Horton R, DeAngelis CD, Drazen JM, Frizelle FA, Godlee F, et al. Clinical trial registration: looking back and moving ahead. *Lancet* 2007;369:1909-11.
20 Ioannidis JP. Effectiveness of antidepressants: an evidence myth constructed from a thousand randomized trials? Philos Ethics *Humanit Med* 2008; 27 May:3:14.
21 Sutton AJ, Cooper NJ, Jones DR. Evidence synthesis as the key to more coherent and efficient research. *BMC Med Res Methodol* 2009;9:29.
22 Salanti G, Kavvoura FK, Ioannidis JP. Exploring the geometry of treatment networks. *Ann Intern Med* 2008;148:544-53.
23 Ioannidis JP. Integration of evidence from multiple meta-analyses: a primer on umbrella reviews, treatment networks and multiple treatments meta-analyses. *CMAJ* 2009;181:488-93.
24 Dias S, Welton NJ, Caldwell DM, Ades AE. Checking consistency in mixed treatment comparison meta-analysis. *Stat Med* 2010;29:932-44.
25 Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. *Eur J Clin Invest* 2010;40:172-82.

Cite this as: *BMJ* 2010;340:c4875

ANSWERS TO ENDGAMES, p 785. For long answers go to the Education channel on bmj.com



Tube oesophagogram showing an H-type tracheo-oesophageal fistula. A: contrast spilling into the respiratory tract; B: H-type tracheo-oesophageal fistula; C: catheter inserted into oesophagus; D: oesophagus filled with contrast

PICTURE QUIZ
A baby with noisy breathing

1 The radiological investigation is a tube oesophagogram, which showed a dilated oesophagus with contrast spilling into the trachea. The diagnosis is an H-type tracheo-oesophageal fistula with, in this case, a background intercurrent bronchiolitis caused by respiratory syncytial virus and possibly an aspiration pneumonia. In this situation, a thorough respiratory history is important to ensure that such cases are not missed.
2 H-type tracheo-oesophageal fistulas typically present with coughing and choking after feeds, abdominal distension as a result of aerophagia, and recurrent aspiration pneumonia. Because symptoms overlap with those of other respiratory conditions, careful history taking is important when making this diagnosis. When carrying out investigations for a suspected H-type tracheo-oesophageal fistula, no radiological investigation is completely reliable. These abnormalities need to be looked for carefully and specifically.
3 Tracheo-oesophageal fistulas are associated with other abnormalities and may occur as part of the VACTERL association (vertebral anomalies, anal atresia, cardiovascular anomalies, tracheo-oesophageal fistula, oesophageal atresia, renal anomalies, and limb anomalies). Perform a careful examination; chromosomal analysis; cranial, vertebral, and renal ultrasound; and echocardiography to rule out any associations.
4 To avoid further aspiration of feeds into the respiratory tract, the patient should not be fed orally. Feeding can be maintained via a naso-jejunal tube. Perform surgery as soon as possible to close the connection between the trachea and oesophagus.

ON EXAMINATION QUIZ
Sepsis
Answer D is correct.

STATISTICAL QUESTION
Confounding in randomised controlled trials
Answers *a*, *c*, and *d* are all true, whereas *b* is false.