Methodological Notes

Cerebrovascular Diseases

Cerebrovasc Dis 2011;31:615–620
DOI: 10.1159/000326068

Received: January 28, 2011
Accepted: February 17, 2011
Published online: April 13, 2011

The Authors Say: 'The Data Are Not So Robust because of Heterogeneity' – So, How Should I Deal with This Systematic Review?

Meta-Analysis and the Clinician

Peter Sandercock

Department of Clinical Neurosciences, University of Edinburgh, Western General Hospital, Edinburgh, UK

Key Words
Systematic review · Meta-analysis · Heterogeneity

Abstract
Systematic reviews can, if done well, provide a convenient and unbiased summary of the evidence on a particular topic. The presence of substantial heterogeneity in a meta-analysis is always of interest. On the one hand, it may indicate that there is excessive clinical diversity in the studies included, and that it is inappropriate to derive an estimate of overall effect from that particular set of studies. On the other hand, appropriate exploration of the source of heterogeneity may either identify poor design of the studies included or perhaps not even identify the cause; in either case, investigating the source of the heterogeneity may be identified as a topic for future research.

Systematic Reviews: The Good, the Bad and the Ugly

Systematic reviews can, if done well, provide a convenient and unbiased summary of the evidence on a particular topic. As a result, they are very popular, and often widely cited. However, the interpretation of the findings of a systematic review must be based on a sound understanding of the method used to acquire the data. Systematic reviews are just as prone to methodological faults and errors as any other type of research study. A badly conducted systematic review could magnify the weaknesses of the studies included and increase the risk of a biased conclusion.

So, how should busy clinicians separate the 'sheep' from the 'goats' and focus their limited reading time on good-quality reviews? A good, systematic review will aim to: pose a clearly defined question; set out a protocol; search extensively for eligible studies; extract the data in an unbiased way, and – only if appropriate – derive an overall estimate of effect (i.e. perform a meta-analysis) and interpret the results with due caution. Interpretation of the overall estimate of effect in a meta-analysis is only possible if the studies included are sufficiently similar and an appropriate statistical method has been used. So, how can one be sure that studies are 'sufficiently similar'?

What Is Heterogeneity?

Inevitably, the studies brought together in a systematic review will differ. This variation is called heterogeneity. The Cochrane Collaboration has published a data-

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1015–9770/11/0316–0615$38.00/0

Accessible online at:
www.karger.com/ced

Prof. Peter Sandercock, MA, DM, FRCPE, FMedSci
Professor of Medical Neurology and Honorary Consultant Neurologist
Director Edinburgh Neuroscience, Department of Clinical Neurosciences
University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU (UK)
Tel. +44 131 537 2928, E-Mail peter.sandercock@ed.ac.uk

base of over 4,000 systematic reviews, which were assembled according to the methods set out in the *Cochrane Handbook for Systematic Reviews of Interventions* (freely available online with invaluable guidance on how to perform reviews) [1]. The *Handbook* defines the different types of heterogeneity:

> Any kind of variability among studies in a systematic review may be termed heterogeneity. It can be helpful to distinguish between different types of heterogeneity. Variability in the participants, interventions and outcomes studied may be described as **clinical diversity** (sometimes called clinical heterogeneity), and variability in study design and risk of bias may be described as **methodological diversity** (sometimes called methodological heterogeneity). Variability in the intervention effects being evaluated in the different studies is known as **statistical heterogeneity**, and is a consequence of clinical or methodological diversity, or both, among the studies. Statistical heterogeneity manifests itself in the observed intervention effects being more different from each other than one would expect due to random error (chance) alone. We will follow convention and refer to **statistical heterogeneity** simply as **heterogeneity**. [1]

Handling Heterogeneity: The Example of Thrombolytic Therapy for Stroke

So, is heterogeneity good or bad? It is easiest to consider this question by working through a sample review, the Cochrane systematic review of thrombolytic therapy for acute ischaemic stroke [2]. The first stage is to define 'the question' in terms of patients, interventions, comparisons and outcomes (PICO). In this example, the question is: in patients with acute ischaemic stroke (P), what is the effect of thrombolytic therapy (I), compared with control (C), on major clinical outcomes (O)?

Limit Clinical Diversity – But Not Too Much

Once the question to be addressed by a systematic review is clear, the protocol for the review aims to limit the clinical diversity of the studies included. Thus, this review included randomised controlled trials which met predefined criteria (e.g. only studies on patients with acute ischaemic stroke that reported key clinical outcomes). Reviews need to strike a balance in their scope. It is often appropriate to take a broader perspective in a meta-analysis than in a single clinical trial, for example by examining the effects of a class of drugs rather than of a single drug. The thrombolysis review included studies of different drugs and different routes of administration. At first sight, we could interpret such diverse interventions as a methodological weakness. However, if the total number of studies is limited, including all drugs with similar pharmacological activity does increase the number of patients with major clinical outcome events, and strengthens any conclusions about the class of drug therapy. The Cochrane review dealt with this diversity by grouping together studies of the same intervention, in this case by the type of drug and its route of administration.

How to Measure Heterogeneity?

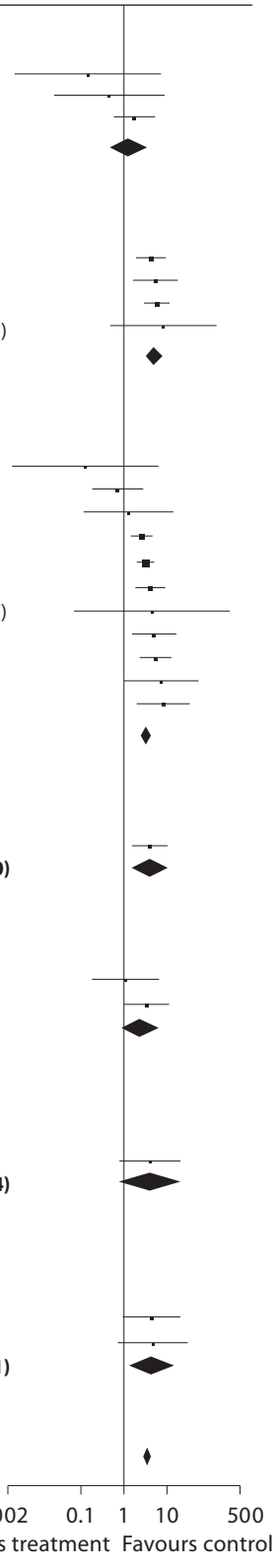
The preferred statistical test to assess the degree of variation between studies beyond that expected by chance (i.e. statistical heterogeneity) is the I^2 statistic, which measures the proportion of inconsistency that cannot be explained by chance alone (I^2 values of over 50% are considered to represent substantial heterogeneity). An analysis of the effects of thrombolysis, compared with control, on symptomatic intracranial haemorrhage (SICH) shows that, within each type of intervention, there is little or no statistical heterogeneity (fig. 1), with values for I^2 of 25% or less, a level which is considered unlikely to be important [1]. Of greater interest, despite the range of thrombolysis interventions, for the overall estimate of effect of thrombolysis on SICH across all groups, I^2 equalled 5%. Thus, the overall estimate of effect on SICH can be considered reliable, and the conclusion that thrombolytic therapy, by whichever drug or route of administration, significantly increases the odds of SICH 3.5-fold ($p < 0.00001$) is not only robust, but clinically useful.

What to Do if There Is Heterogeneity?

The analysis of the effects of thrombolytic therapy on the outcome 'death or dependency (modified Rankin Scale score 3–6)' shows a rather different picture (fig. 2). The only category of thrombolytic therapy to demon-

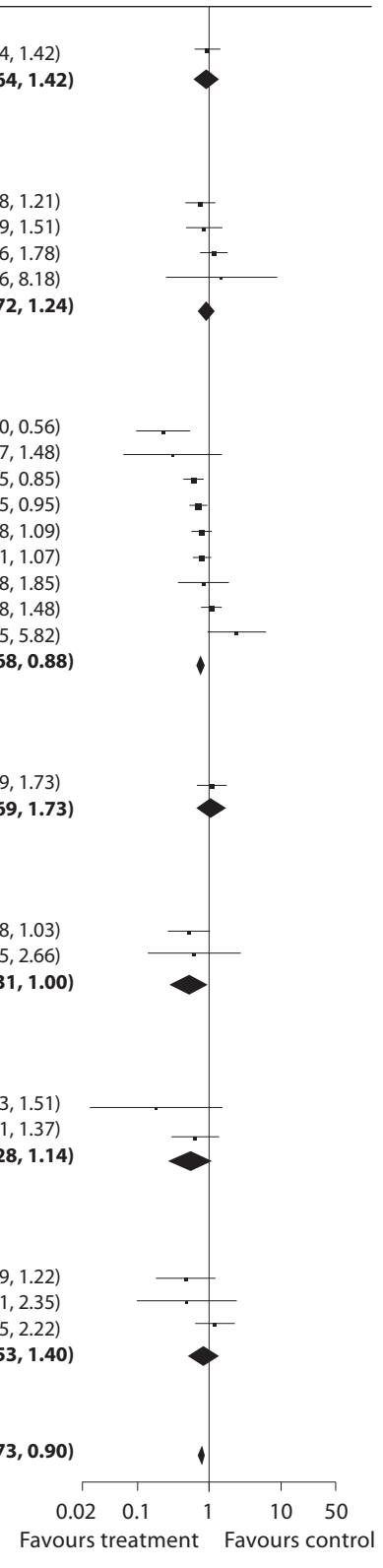
Fig. 1. Forest plot (from the Cochrane review of thrombolytic therapy versus control for acute ischaemic stroke [2]) of the effect on SICH (including fatal). There is 1 line per trial, and the point estimate of the effect within that trial is expressed as an odds ratio (OR) (calculated on a fixed-effects model by the Peto method) displayed as a dot. Horizontal line: 95% CI of the effect (values in parentheses in the Peto OR column). Diamonds: estimate of effect within each category of thrombolytic therapy. Vertical line: OR of unity (i.e. of no effect); OR <1 suggest that thrombolysis reduces the odds of the outcome, OR >1 that it increases the odds.

Study or subgroup	Treatment		Control		Weight, %	Peto odds ratio
	events, n	total, n	events, n	total, n		
1.3.1 Intravenous urokinase versus control						
Abe, 1981	0	54	0	53		not estimable
Ohtomo, 1985	0	169	1	181	0.3	0.14 (0.00, 7.30)
Atarashi, 1985	1	192	1	94	0.5	0.46 (0.02, 8.81)
Chen, 2000	12	317	3	148	3.8	1.75 (0.58, 5.29)
Subtotal		**732**		**476**	**4.6**	**1.28 (0.47, 3.48)**
Total events	13		5			
Heterogeneity: $\chi^2 = 1.96$, d.f. = 2 (p = 0.37), $I^2 = 0\%$, test for overall effect: Z = 0.48 (p = 0.63)						
1.3.2 Intravenous streptokinase versus control						
ASK, 1996	22	174	4	166	7.3	4.24 (1.91, 9.43)
MAST-I, 1995	10	157	1	156	3.2	5.39 (1.62, 17.91)
MAST-E, 1996	33	156	4	154	9.9	5.81 (2.93, 11.53)
Morris, 1995	2	10	0	10	0.6	8.26 (0.48, 142.43)
Subtotal		**497**		**486**	**20.9**	**5.20 (3.25, 8.32)**
Total events	67		9			
Heterogeneity: $\chi^2 = 0.46$, d.f. = 3 (p = 0.93), $I^2 = 0\%$, test for overall effect: Z = 6.86 (p < 0.00001)						
1.3.3 Intravenous tissue plasminogen activator versus control						
Haley, 1993	0	14	1	13	0.3	0.13 (0.00, 6.33)
JTSG, 1993	4	51	5	47	2.5	0.72 (0.18, 2.81)
Mori, 1992	2	19	1	12	0.8	1.27 (0.12, 14.12)
ECASS II, 1998	36	409	13	391	13.9	2.59 (1.45, 4.61)
ECASS, 1995	62	313	20	307	21.5	3.18 (2.00, 5.06)
ATLANTIS B, 1999	21	307	4	306	7.2	4.10 (1.84, 9.13)
Wang, 2003	1	67	0	33	0.3	4.45 (0.07, 287.37)
ECASS 3, 2008	10	418	1	403	3.3	5.05 (1.54, 16.60)
NINDS, 1995	20	312	2	312	6.4	5.44 (2.32, 12.73)
EPITHET, 2008	4	52	0	49	1.2	7.41 (1.01, 54.23)
ATLANTIS A, 2000	8	71	0	71	2.3	8.20 (1.98, 33.99)
Subtotal		**2,033**		**1,944**	**59.7**	**3.28 (2.48, 4.33)**
Total events	168		47			
Heterogeneity: $\chi^2 = 13.10$, d.f. = 10 (p = 0.22), $I^2 = 24\%$, test for overall effect: Z = 8.34 (p < 0.00001)						
1.3.4 Intravenous streptokinase + oral aspirin versus oral aspirin						
MAST-I, 1995	15	156	3	153	5.1	4.02 (1.55, 10.40)
Subtotal		**156**		**153**	**5.1**	**4.02 (1.55, 10.40)**
Total events	15		3			
Heterogeneity: not applicable, test for overall effect: Z = 2.87 (p = 0.004)						
1.3.5 Intra-arterial prourokinase + intravenous heparin versus intravenous heparin						
PROACT, 1998	4	26	2	14	1.4	1.09 (0.18, 6.56)
PROACT 2, 1999	12	121	1	59	3.2	3.39 (1.02, 11.24)
Subtotal		**147**		**73**	**4.7**	**2.39 (0.88, 6.47)**
Total events	16		3			
Heterogeneity: $\chi^2 = 1.06$, d.f. = 1 (p = 0.30), $I^2 = 6\%$, test for overall effect: Z = 1.71 (p = 0.09)						
1.3.6 Intra-arterial urokinase versus control						
AUST, 2005	0	8	0	8		not estimable
MELT, 2007	5	57	1	57	1.7	4.04 (0.79, 20.74)
Subtotal		**65**		**65**	**1.7**	**4.04 (0.79, 20.74)**
Total events	5		1			
Heterogeneity: not applicable, test for overall effect: Z = 1.67 (p = 0.09)						
1.3.7 Intravenous desmoteplase versus control						
DEDAS, 2006	0	29	0	8		not estimable
DIAS, 2005	9	75	0	27	1.9	4.38 (0.94, 20.50)
DIAS 2, 2008	5	123	0	63	1.3	4.69 (0.72, 30.51)
Subtotal		**227**		**98**	**3.3**	**4.50 (1.37, 14.81)**
Total events	14		0			
Heterogeneity: $\chi^2 = 0.00$, d.f. = 1 (p = 0.96), $I^2 = 0\%$, test for overall effect: Z = 2.48 (p = 0.01)						
Total		**3,857**		**3,295**	**100.0**	**3.49 (2.81, 4.33)**
Total events	298		68			
Heterogeneity: $\chi^2 = 24.25$, d.f. = 23 (p = 0.39), $I^2 = 5\%$, test for overall effect: Z = 11.37 (p = 0.00001)						
Test for subgroup differences: $\chi^2 = 7.66$, d.f. = 6 (p = 0.26), $I^2 = 21.7\%$						



1

Study or subgroup	Treatment		Control		Weight, %	Peto odds ratio
	events, n	total, n	events, n	total, n		
1.6.1 Intravenous urokinase versus control						
Chen, 2000	127	317	61	148	6.8	0.95 (0.64, 1.42)
Subtotal		**317**		**148**	**6.8**	**0.95 (0.64, 1.42)**
Total events	127		61			
Heterogeneity: not applicable, test for overall effect: Z = 0.24 (p = 0.81)						
1.6.2 Intravenous streptokinase versus control						
MAST-I, 1995	97	157	106	156	5.0	0.76 (0.48, 1.21)
MAST-E, 1996	124	156	126	154	3.4	0.86 (0.49, 1.51)
ASK, 1996	84	174	74	166	5.9	1.16 (0.76, 1.78)
Morris, 1995	6	10	5	10	0.4	1.47 (0.26, 8.18)
Subtotal		**497**		**486**	**14.6**	**0.94 (0.72, 1.24)**
Total events	311		311			
Heterogeneity: $\chi^2 = 2.06$, d.f. = 3 (p = 0.56), $I^2 = 0\%$, test for overall effect: Z = 0.41 (p = 0.68)						
1.6.3 Intravenous tissue plasminogen activator versus control						
Wang, 2003	29	67	26	33	1.5	0.24 (0.10, 0.56)
Mori, 1992	11	19	10	12	0.5	0.32 (0.07, 1.48)
NINDS, 1995	155	312	192	312	10.7	0.62 (0.45, 0.85)
ECASS II, 1998	187	409	211	391	13.9	0.72 (0.55, 0.95)
ECASS, 1995	171	313	185	307	10.5	0.79 (0.58, 1.09)
ECASS 3, 2008	140	418	155	403	13.1	0.81 (0.61, 1.07)
EPITHET, 2008	28	51	29	49	1.7	0.84 (0.38, 1.85)
ATLANTIS B, 1999	141	307	135	306	10.5	1.08 (0.78, 1.48)
ATLANTIS A, 2000	64	71	56	71	1.3	2.35 (0.95, 5.82)
Subtotal		**1,967**		**1,884**	**63.8**	**0.78 (0.68, 0.88)**
Total events	926		999			
Heterogeneity: $\chi^2 = 20.98$, d.f. = 8 (p = 0.007), $I^2 = 62\%$, test for overall effect: Z = 3.82 (p = 0.0001)						
1.6.4 Intravenous streptokinase + oral aspirin versus oral aspirin						
MAST-I, 1995	99	156	94	153	5.0	1.09 (0.69, 1.73)
Subtotal		**156**		**153**	**5.0**	**1.09 (0.69, 1.73)**
Total events	99		94			
Heterogeneity: not applicable, test for overall effect: Z = 0.37 (p = 0.71)						
1.6.5 Intra-arterial prourokinase + intravenous heparin versus intravenous heparin						
PROACT 2, 1999	73	121	44	59	2.5	0.54 (0.28, 1.03)
PROACT, 1998	18	26	11	14	0.5	0.63 (0.15, 2.66)
Subtotal		**147**		**73**	**3.0**	**0.55 (0.31, 1.00)**
Total events	91		55			
Heterogeneity: $\chi^2 = 0.04$, d.f. = 1 (p = 0.84), $I^2 = 0\%$, test for overall effect: Z = 1.97 (p = 0.05)						
1.6.6 Intra-arterial urokinase versus control						
AUST, 2005	4	8	7	8	0.3	0.19 (0.03, 1.51)
MELT, 2007	29	57	35	57	2.0	0.65 (0.31, 1.37)
Subtotal		**65**		**65**	**2.2**	**0.57 (0.28, 1.14)**
Total events	33		42			
Heterogeneity: $\chi^2 = 1.19$, d.f. = 1 (p = 0.27), $I^2 = 16\%$, test for overall effect: Z = 1.59 (p = 0.11)						
1.6.7 Intravenous desmoteplase versus control						
DIAS, 2005	46	75	21	27	1.3	0.49 (0.19, 1.22)
DEDAS, 2006	13	29	5	8	0.4	0.50 (0.11, 2.35)
DIAS 2, 2008	72	123	34	63	2.8	1.20 (0.65, 2.22)
Subtotal		**227**		**98**	**4.6**	**0.86 (0.53, 1.40)**
Total events	131		60			
Heterogeneity: $\chi^2 = 3.10$, d.f. = 2 (p = 0.21), $I^2 = 36\%$, test for overall effect: Z = 0.61 (p = 0.54)						
Total		**3,376**		**2,907**	**100.0**	**0.81 (0.73, 0.90)**
Total events	1,718		1,622			
Heterogeneity: $\chi^2 = 33.90$, d.f. = 21 (p = 0.04), $I^2 = 38\%$, test for overall effect: Z = 3.90 (p = 0.0001)						
Test for subgroup differences: $\chi^2 = 6.53$, d.f. = 6 (p = 0.37), $I^2 = 8.1\%$						



2

strate substantial within-category statistical heterogeneity was intravenous recombinant tissue plasminogen activator (rtPA; $I^2 = 62\%$). The odds ratio was 0.78 (or rtPA reduced the odds of death or dependency by 22%), the confidence intervals were narrow and the p value for the overall effect was highly significant ($p = 0.0001$), but how should we interpret this result with this degree of heterogeneity? Firstly, the overall estimate should be viewed with caution despite the statistical significance, and secondly, we should undertake a series of analyses to try and understand the source of heterogeneity: is it clinical or methodological?

Exploring the Causes of Heterogeneity: Methodological Diversity

In a systematic review, the design features of each of the clinical trials included may affect the overall estimate of treatment effect. These include: random allocation, allocation concealment, blinding of outcome assessment and sample size [1, 3]. Other aspects of study design may result in heterogeneity, such as the method used to assess outcome, variation in the timing of outcome, and length of follow-up [1]. The first step in dissecting the source of heterogeneity in a meta-analysis is therefore to explore whether variations in these factors might account for the variations between studies with sensitivity analyses (see the *Cochrane Handbook* [1] for details). In the thrombolysis review, whilst some of these factors may have accounted for part of the heterogeneity, they did not account for the majority.

Exploring Statistical Heterogeneity: How Not to Do It

Schulz and Grimes [4] have emphasised the perils of inappropriate subgroup analysis in clinical trials, and the same principles apply – perhaps even more so – to meta-analysis [1]. Counsell et al. [5] have shown that by use of inappropriate methods, a meta-analysis of trials of a truly ineffective treatment (DICE therapy) could be performed to make the treatment appear beneficial and, equally, to suggest that the effects of treatment were materially different in different categories of patients (when in fact they were not) [3–5].

Fig. 2. Forest plot (from the Cochrane review of thrombolytic therapy versus control for acute ischaemic stroke [2]) of the effect on the outcome 'dead or dependent in activities of daily living at the end of follow-up'. This outcome corresponds to a modified Rankin Scale score of 3–6. Otherwise, same conventions as in figure 1.

Exploring Statistical Heterogeneity: Metaregression and Other Approaches

There is a variety of statistical approaches that can be taken to explore heterogeneity, and in systematic reviews, the approach to the subgroup and other analyses must be clearly specified in advance in the protocol and interpreted cautiously to avoid undue emphasis on post hoc data-dependent analyses [5]. The ideal approach is to obtain individual patient data from every study included, which then permits a much more detailed analysis of the factors that modify the effects of treatment. However, most systematic reviews are based on summary data extracted from publications, which inevitably limits the range and scope of analyses that can be performed. The Cochrane thrombolysis review took a methodologically rigorous approach and carried out a variety of prespecified analyses to identify sources of heterogeneity in the effects of thrombolysis in general (and rtPA in particular). Although the authors used metaregression to explore some variables, this has relatively limited power if the number of studies is small. The scope of the analyses was also limited by the nature of the data that were reported in the original trial publications, and hence the authors concluded, having examined as wide a range of factors as appropriate, within the limits of the data available:

> Despite the overall net benefit, the available data do not provide sufficient evidence to determine the magnitude of treatment effect, the duration of the therapeutic time window, the optimum agent (or dose or route of administration) or the clinical or radiological features which identify the patients most likely to benefit (or be harmed), the age of the patient, and when and if antithrombotic treatment may be safely used around the time of thrombolysis. [2]

In the case of this review, the exploration of the sources of heterogeneity has identified new research questions and highlighted the implications for future research (which are set out in detail in the full review). In other words, the presence of heterogeneity itself posed a research question that could not be fully answered by the available data.

One pooled individual patient data analysis of the effects of rtPA examined the factors that modify the time

dependence of the effect on benefit from rtPA, but did not examine the effects of any covariates such as age, neurological impairment or study design on the overall estimate of effect [6]. The key determinants of the heterogeneity observed in the Cochrane review will therefore remain uncertain until more randomised trial data and a more extensive individual patient data analysis of all randomised trials has been completed.

Conclusions

The exploration of subgroup effects and of the sources of heterogeneity is potentially beset with problems, with undue emphasis on post hoc subgroup or sensitivity analyses a major source of concern. The planned analysis of data from a collection of similar trials in a meta-analysis of several clinical trials should therefore be specified in advance in a protocol. The analyses should be performed with expert statistical support and interpreted with due caution. Nonetheless, the presence of substantial heterogeneity in a meta-analysis is always of interest. On the one hand, it may indicate that there is excessive clinical diversity in the studies included, and that it is inappropriate to derive an estimate of overall effect from that particular set of studies. On the other hand, appropriate exploration of the source of heterogeneity may either identify poor design of the studies included or perhaps not even identify the cause; in either case, investigating the source of the heterogeneity may be identified as a topic for future research.

References

1 Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2. 2009. www.cochrane-handbook.org.

2 Wardlaw JM, Murray V, Berge E, del Zoppo GJ: Thrombolysis for acute ischaemic stroke. Cochrane Database Syst Rev 2009; 4:CD000213.

3 Collins R, MacMahon S: Reliable assessment of the effects of treatment on mortality and major morbidity. 1. Clinical trials. Lancet 2001;357:373–380.

4 Schulz KF, Grimes DA: Multiplicity in randomised trials. 2. Subgroup and interim analyses. Lancet 2005;365:1657–1661.

5 Counsell CE, Clarke MJ, Slattery J, Sandercock PA: The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? BMJ 1994;309:1677–1681.

6 Lees KR, Bluhmki E, von Kummer R, Brott TG, Toni D, Grotta JC, Albers GW, Kaste M, Marler JR, Hamilton SA, Tilley BC, Davis SM, Donnan GA, Hacke W, ECASS, ATLANTIS, NINDS and EPITHET rt-PA Study Group, Allen K, Mau J, Meier D, del Zoppo G, de Silva DA, Butcher KS, Parsons MW, Barber PA, Levi C, Bladin C, Byrnes G: Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. Lancet 2010;375:1695–1703.