The use of evidence based medicine in health technology assessment.

Nick Freemantle PhD

Professor of Clinical Epidemiology & Biostatistics, University of Birmingham, UK.

Contents

Introduction

In this discussion paper I consider aspects of the statistical synthesis of evidence intended to support decisions on pharmaceutical reimbursement. Many of the comments are more general in scope and would, for example, be relevant for the appraisal of devices or other health technologies. The paper is not intended as a guide to good practice for novice users, but instead to address issues of importance and interest among those currently conducting work in this field. The author has deliberately taken a specific position on several points, rather than prevaricated and left the reader in doubt regarding his general position. However the strength of such conclusions is indicated in the text, and where these may reasonably be considered controversial, this point will also be acknowledged.

1. The Basis for Appraisal

Health systems throughout the world are increasingly concerned with the evaluation of the effectiveness and cost effectiveness of pharmaceuticals (and health technologies more generally), in order to enhance the extent to which scarce health resources are used efficiently.[Freemantle et al 1995] Such evaluation is an extension from the conventional regulatory process rather than a replacement, and may be considered to be a step undertaken after a product has gained marketing authorisation. This appraisal has been described as a *fourth hurdle* process, where the regulatory hurdles of demonstration of efficacy, quality and safety are added to with the concept of efficiency or value for money. The regulatory process serves to identify and endorse that a product does indeed have a therapeutic effect; that it is manufactured to acceptable standards; and that it does not bring undue hazard to patients with use. The appraisal process subsequently considers if and where society might rationally bring the intervention into usage.

Developing appraisal systems that make optimal use of the best available evidence to make rational health policy decisions is not straight forward. And several different versions exist around the world from NICE and the SMC in the UK, PBS in Australia among others. The IQWIG appears unique in the requirement of an established ‘benefit’ for a new treatment (eg an advantage in health outcome over existing treatments) before that treatment may be eligible for a reimbursement recommendation. This requirement is surprising and can lead to inefficiencies for the health system.

The evaluation of cost effectiveness requires the understanding of the benefits and costs of therapy, and should not be limited to benefits alone in the first instance. Both pieces of information are required for rational health policy making. If a product achieves very similar effectiveness to other available products on all criteria and advances in no way upon existing alternatives, and if it is more costly, there is no rational basis for its use and a process of appraisal should identify and proscribe such use. If alternatively a product achieves the same outcomes as the standard therapy but at a lesser overall cost then the appraisal process should identify this attribute, and the use of the new product should be recommended in place of the old. Currently this possibility appears excluded by the German system. Where the new product achieves less than the existing product, but at a reduced price, or where the product achieves more than the comparator, but at an increased price, the magnitude and value of the incremental benefits must be determined, and the opportunity cost of adopting the new strategy (eg the benefits forgone through using scarce resources in this way rather than in a different way) must be assessed. It may be efficient for a health system to adopt a less effective but less costly treatment where overall limited resources may be allocated in such a way as to achieve greater health benefit. Similarly it may be cost effective to implement a more costly treatment which has incremental attributes the achievement of which represent worthwhile health benefits for patients.

Hence the appraisal process brings many challenges which extend those experienced in the regulatory sector, but also many potential benefits for society. The regulatory system has served society very well, but is necessarily limited and an insufficient basis to inform rational evidence based decision making. The strengths of the regulatory process include the prospective nature of the experimental plan, the standard requirement for two statistically significant confirmatory trials in phase IIIa to achieve a marketing authorisation, the control of type 1 error, and the high level of regulation. None of these strengths is available to an assessor in the post marketing authorisation period. Conceptually the registration process converts a complex scenario (the attributes of a new pharmaceutical product) into a dichotomy, aiding decision making. However such an approach is quite inappropriate for the assessment of the cost effectiveness of a new product since here we need an overall understanding of the incremental costs and benefits of a therapy across all relevant attributes.

Thus the regulatory system may be considered to identify candidate interventions which might prove cost effective in specific circumstances. It is not appropriate to expect health professionals (prescribers) to make the decision on whether such candidate interventions should be used, and when they should be used, on the basis of their clinical judgement alone (although their judgement will form part of a subsequent decision) since prescribers cannot be in possession of a full understanding of all likely consequences of different actions. For example the adoption of fibronylitic agents after myocardial infarction was probably delayed because clinicians saw the adverse consequences of such intervention (bleeds) but did not observe directly the additional life years gained by suitably treated patients. [Lau 1992] Instead, a formal evidence based appraisal is required.

The manner in which different health systems have addressed the question of appraisal has differed to some degree. For example in the Australian Pharmaceutical Benefits Scheme, the sponsor either makes an application on the basis of superiority or of non inferiority, with the scheme pointing towards the use of cost effectiveness analysis for the former, and cost minimization analysis for the latter.[Australian Govt 2007] The NICE in England and Wales has a strong preference to cost effectiveness analysis, and does not invite submissions based on cost minimization for the reference case, although not ruling out different forms of analysis [NICE 2007]. However common to both approaches is a broad and scientifically valid assessment of the strength of evidence on the relative attributes of a new product, and the consideration of these results alongside the costs associated with different treatment options.

Appraisal in this context is a complex and multi attributed process, and one which advances from the necessarily simplified considerations of the regulatory process. This basic requirement to appraise the attributes (both benefits and costs) of new technologies in comparison with standard therapies is what makes the process challenging, but there are no short cuts to achieving the correct outcomes. The onus on us to make good decisions is the responsibility to ensure that patients receive the maximum health benefits possible from the inevitably limited resources available and if we fail in this regard, we disadvantage the people for whom appraisal occurs; those with a need for effective health care intervention.

2. A short introduction to meta analysis

It is highly unlikely that a single trial will provide all the necessary information to inform treatment decisions. As chance alone will lead to different results in different trials we would be faced with the difficulty of deciding what trial to consider if we were only to select one. Meta analysis is a generic name for an array of statistical approaches to combining the results from all suitable trials in order to provide the best scientific summary of treatment effects. Meta analysis provides further opportunities and challenges beyond the consideration of individual trials. Thus meta analysis estimates a weighted average of the effects estimated within individual trials, preserving the structure of the contributing trials and thus the advantages of randomisation. The main potential pitfalls of meta analysis relate to the quality of the studies which are included, and the adage of ‘rubbish in, rubbish out’ works well; where a meta analysis includes poor quality trials, it provides a poor estimate of the true effects of treatment. A further caveat is that meta analysis in almost all circumstances occurs retrospectively, including data which are already available in the public domain, and thus cannot have the protection of prospective method. Indeed, meta analyses are often undertaken because of the availability of interesting results from existing trials, with the clear risk of bias associated with this.

This paper covers the following points: The Role of randomisation; Meta analysis of multiple clinical trials; Subgroup analyses and their particular challenges; Sensitivity analysis, especially for handling missing values and drop outs; Methods for combining classes of evidence / grading of evidence; Analysis of adverse events; Indirect meta analyses; Handling of heterogeneity; Combination of evidence for substances to assess evidence of class effect; Where the primary endpoint of a meta analysis is not the same as the primary endpoints for the original studies; Interpretation of results (Estimation versus testing)

3. The role of randomisation

Using Randomisation to Avoid Bias in Comparisons of Different Treatments

Many new treatments for medical conditions have the potential to achieve important but modest improvements in health outcomes. For example, treatments for moderate to severe heart failure have advanced dramatically over the last 20 years, including the development of a sound evidence base for angiotensin converting enzyme (ACE) inhibitors, and beta blockers. Estimates of the absolute benefits of ACE inhibitors taken from randomised trials suggest that all cause mortality is reduced by around 1.5% per year as a result of their use.[Cleland 1999] Further trials suggest an additional absolute benefit of around 3.5% in reducing mortality from the use of beta blockers.[Cleland 1999]. Thus taken together we see a reduction in mortality of around 5% per year. While highly desirable to achieve, this health gain which is achieved through the use of two different interventions has required trials randomising several thousand subjects to develop evidence to inform clinical practice and ultimately to benefit patients.

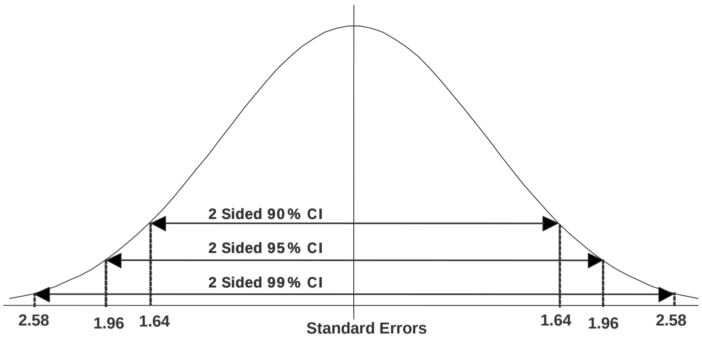
When attempting to estimate relatively modest but important treatment effects, such as the benefits in heart failure from beta blockers and ACE inhibitors described above, differences in patient characteristics, which may be poorly understood, can lead to quite large differences in outcome between different patients. Similarly, clinician judgement or patient preference may be more or less effective in identifying patients likely to do well or badly. These factors confound our abilities to estimate the effects of such treatments from routine clinical practice. Randomisation is the main approach used in attempts to avoid such biases, as described below.

Randomisation is used in clinical trials to avoid bias. It does this through the distribution of both known and unknown biases between treatment groups on the basis of the play of chance. Random allocation indicates that subjects are allocated to the different treatment groups or ‘experimental conditions’ on the basis of the play of chance. A common misconception is that randomisation makes the comparison groups “the same” which is of course not the case unless an infinite number of subjects are randomised, which would be both impractical and an inefficient use of resources. Instead, properly conducted, randomisation ensures that the comparison groups differ only on the basis of the play of chance (rather than by some systematic feature which would lead to potential bias) thus providing a good theoretical basis for comparison. Further, randomisation leads us to the useful circumstance where there are only two potential explanations for observed differences between randomised treatment groups in a trial. Such differences may either be attributable to the treatment allocated, or to the play of chance. This orthogonal situation enables us to draw unbiased conclusions about the effects of treatment. We can use statistical methods to examine how plausible it may be that an observed difference between the treatment groups or experimental conditions may simply be attributable to the play of chance. If it is not plausible that differences observed between the treatment groups are attributable to the play of chance then the only available alternative is that these differences must be attributable to the treatment. Where there is only a single randomised trial, the difference between the groups provides us our best estimate of treatment effect. The situation is complicated somewhat where there are multiple clinical trials, in which case we need to consider a systematic summary or overview of those trials, which should include all trials which have been conducted, or at least an unbiased sample of such trials.

Thus, to return to the interpretation of trials results, we can take the point estimate from a randomised trial (eg the mean difference between the randomised groups) to be the best available estimate of the true treatment effect, the confidence intervals can be taken to describe the plausible bounds around our estimate of the true effect, and finally we can understand that it becomes increasingly implausible that a point on the confidence interval describes the true treatment effect as we move away from the point estimate derived from the trial and towards the limits of the confidence intervals. The latter point is illustrated by the shape of the cumulative normal distribution in Figure 1.

As an aside, a confidence limit produced using a Bayesian approach may arguably be a reflection of the true variability of the result in question, given appropriate priors, rather than taking the assumption that the coverage properties of new randomisations describes uncertainty for the population value. However the Bayesian approach is itself open to some strong assumptions, and will provide similar results given similar assumptions.

Figure 1 Confidence Intervals and the Standardised Normal Distribution



The concept of external validity within appraisal or health technology assessment

It is almost always appropriate only to include randomised trials for estimates of the effect of treatments, although by contrast it is often appropriate to include non randomised studies in the population of other aspects of value for money arguments (in particular for the population of patient characteristics in an economic model). This is because properly conducted randomisation ensures that the groups in a randomised trial differ only on the basis of the allocated intervention and the play of chance. Thus differences between groups may be attributed either to chance or to the intervention. Non randomised studies do not have this protection, and thus differences between groups may be due to chance, the intervention received, or some form of latent selection bias.

A number of authors have derived hierarchical tables of evidence based primarily on the protection of randomisation from bias. See Table 3 which was derived from Guyatt et al 2000.

Table 1: Hierarchy of study designs

Systematic review of randomised trials

Individual randomised trial

Systematic review of observational studies addressing patient important outcomes

Single observational study addressing patient important outcomes

Physiologic studies

Expert opinion

[Derived from Guyatt et al 2000]

A latent selection bias may not be estimated directly nor conditioned for directly in a statistical model of an observational study since it is confounded completely with intervention. The propensity score approach may be used to describe the observed risk factors for a subject efficiently, however this approach will not help with confounding by indication since in that setting, the potential bias is the additional risk identified by the treating clinician above and beyond the risks described by the observed risk factors.   Where there is some form of instrumental variable such as distance from relevant treatment setting or other non patient characteristic derived cause for selection which can be seen as to some degree a chance selection factor, then it may be possible to do a more thorough job of conditioning results for selection bias. For example in a study of the relationship between aprotinin and risk of death in patients undergoing cardiac surgery, the policy of the treating centre changed during the period of the study from one in which aprotinin was used only in high risk cases to one in which aprotinin was used universally apart from in very low risk cases [Pagano et al 2008]. Thus the changing selection bias could be conditioned for by including time period in the analysis (the unconditioned risks of aprotinin appeared ‘U’ shaped over the total time period), although statistical models are limited in the extent to which they can address such biases, and simply including a confounding factor in a statistical model cannot, unfortunately, be taken to mean that the confounder has been fully addressed.

There are, however, some potential advantages of including supportive observational data alongside the results of meta analyses of randomised trials. This is because randomised trials are inevitably limited in scope and address a somewhat selected population. For example in work addressing the effect of beta blockers in myocardial infarction, many of the trials were conducted before the availability of modern pharmaceutical interventions such as ‘statins [Freemantle 1999]. This had led to the perception that beta blockade may no longer be a relevant treatment in acute myocardial infarction, leading to underuse [Owen 1999]. Comparison of the results of the meta analysis with the results from a large US data base study [Gottlieb 1998] proved particularly useful in this case, as the latter was conducted subsequently to the availability of the newer pharmaceutical agents of interest, in an unselected population and in a population with a much broader set of inclusion criteria. In addition, the effect of beta blockers is quite large in absolute terms (and thus less likely to be overwhelmed by a confounding factor than a smaller effect).

Although randomised trials may provide the only convincing evidence of treatment effect, they are limited in terms of the degree of selection and the artificial setting of a trial. Patients are selected, or self selected, on the basis that they are likely to engage with therapy and be more adherent than normal to recommended treatment. The treatment setting in a trial is usually somewhat artificial, particularly in terms of the relatively intensive schedule of visits, the exclusion of patients with co-morbidities and the extent to which the treatment centre ensures that subjects are not lost to follow up. In a double blind trial all investigations and interventions (be they real or placebo) are taken by all randomised patients, and thus any difference in the resource implications of different interventional strategies (such as the requirement for additional investigation in one group) will not be observable directly in the trial. Where an economic model is being developed these factors may severely restrict the extent to which trial data alone may be used to describe patients of interest in a relevant manner. Observational studies may be particularly helpful in this situation, providing an opportunity to describe the characteristics and experience of real patients not receiving the intervention of interest. Treatment estimates from randomised trials may then be used to modify the expectation of outcome in treated patients within an epidemiological disease model. Further, costs may be included in order to estimate the costs and benefits of the candidate treatment.

4. Subgroup Analyses and their problems

A decision maker will be interested in several outcome measures. Further, it is clear that subgroups of patients from clinical trials may be of importance in economic analyses of health technologies since the absolute benefits of treatment is an important driver of cost effectiveness. The challenges presented by the presence of multiple questions of interest, and of differences in treatment effects between subjects with different characteristics, are present both in the interpretation of individual trials and also in the analysis and interpretation of evidence synthesis on multiple trials.

The appropriateness of subgroup analysis has long been controversial in the analysis of individual randomised trials and meta analyses of randomised trials [Freemantle 2001]. Subgroup analysis refers to the analysis of a subset of subjects, usually with the aim of assessing whether the effect of the experimental condition (eg treatment) may be larger or occasionally smaller in that subset. Many of the same considerations affect the interpretation of secondary outcomes in randomised trials [Freemantle 2005].

The major problems with subgroup analyses focus on the issue of multiplicity or multiple testing and on bias derived from data driven analyses. For the former question of multiple testing, it is quite clear that one way to achieve a positive test result in the context of a neutral overall result is to conduct a lot of tests. If the standard statistical test in this context is conceptualised as a 20 sided dice, with one of the side marked with a tick, and the other 19 marked with a cross, the probability (p) that the tick will be upwards is .05 for a single throw, but with multiple throws of the dice (n), the probability of having at least one tick is a random variable following a binomial distribution with an expected value of np, and a variance of np(1-p). In other words, the more times you throw the dice, the higher the expected number of ticks.

Individual clinical trials overcome the problem of multiplicity through the identification of a primary outcome measure, or through some form of more sophisticated alpha spending plan. When a single primary outcome is identified a priori, this serves to reduce the number of relevant statistical tests to a single test, as the primary outcome is taken to be the analysis which defines whether a trial may be considered positive, negative or neutral in finding. Secondary outcomes are then used to describe further the results of the trial, perhaps illustrating the consequences or mechanisms for a treatment effect where the result is statistically significantly positive or negative. They may also be useful to provide additional data to support the neutral results of a trial, or to provide exploratory hypotheses which describe potential consequences from the intervention, findings which would always require some form of confirmation in a subsequent trial.

Alpha spending approaches provide an additional sophistication to the *a priori* identification of the primary outcome; with the alpha spending approach the primary and secondary outcomes (or sub-populations of interest) are placed in a hierarchy or order, again a priori, and the available type 1 error (the error relating to making a false positive conclusion) is ‘spent’ down this list of outcomes until it is exhausted to a pre defined level (usually α = 2.5% (one sided)). The outcomes which are on the list above the line where the type I error is exhausted are considered to be proven. Those below that level (regardless of the nominal level of statistical significance for the individual item) are considered to be unproven. The alpha spending approach is commonly used in submissions to the Food and Drug Administration in the USA, and only those items which are proven are subsequently listed on the marketing authorisation for products which successfully receive a license.

The issue of relevance when considering the results of subgroup analyses is that it would be unusual for the subgroup analysis to be part of the original alpha spending plan for the trial (eg to be defined as the primary outcome, or in some way placed in an hierarchical order reflecting alpha spending). Thus the statistical results for a subgroup should be considered in the context of the circumstances where the subgroup result is not the only statistical test which has been undertaken.

Any of the approaches (primary population / outcome measure or alpha-spending approaches) fails to help us in the situation where the test may not have been adequately prespecified and indeed may have been designed with the result in mind; that is that the question asked and the population it is asked in may be driven by the statistical significance of the result. This situation exactly fits the case where a meta analysis is planned. It has been fashionable to prespecify a primary outcome, secondary outcomes and subgroup analyses, but unlike in the case of individual trials, prespecification is not usually helpful as the meta analysis is not usually conducted strictly prospectively (although where a prospective design is used, the approach of prespecification is extremely helpful and analogous in the protection which it provides to that observed in individual trials). This is because candidate trials tend to exist at the stage that the meta analysis is conceived, and researchers tend to know of the results of at least some of them before they develop their protocol or analysis plan, indeed the knowledge of the results of a subset of trials may often provide the impetus to conduct a meta analysis. Some meta analyses are defined a priori, before trials have even completed [Blood Pressure Treatment Trialilsts' Collaborative 2000] although this is the exception rather than the rule. Another approach is sometimes used is to penalise the summary value by applying a higher level of statistical precision (eg using 99% confidence rather than the conventional 95% level, and thus applying a critical α value of 1% rather than 5%). While this addresses (in a somewhat arbitrary way) the bias of multiple testing it does nothing to address the question of data driven analysis.

So how might we decide whether an apparent subgroup analysis result was a true result, particularly in the context of a meta analysis of multiple trials? This is discussed in the next section.

Table 2 describes criteria which may be useful in judging when a subgroup analysis may be believable. These criteria were extended from work originally conducted by Oxman and Guyatt,(1992) the extension aiming to encompass questions of alpha spending.

Table 2: Criteria for considering the persuasiveness of subgroup analyses

Referring to the overall study results:

- Is the primary outcome of the trial or meta analysis statistically significant?

Referring to the subgroup analysis:

- Is the magnitude of the difference clinically important?
- Was the difference statistically significant, including tests for difference and interaction?
- Did the hypothesis precede rather than follow the analysis?
- Was the subgroup one of a small number of hypotheses tested?
- Was the difference suggested by comparisons within rather than between studies?
- Is there indirect evidence that supports the hypothesised difference?

[Freemantle 2001]

It is clear that subgroups of patients from clinical trials may be of importance in economic analyses of health technologies since the absolute benefits of treatment is an important driver of cost effectiveness. Thus there may be particular interest to identify groups where the benefits are greatest. If homogeneity of treatment effect across sub populations is identified, we simply face the task of applying an appropriate outcome measure (e.g. of risk) to a relevant cohort of patients, which is usually conducted in an economic model. However where we are interested in differences in the effects of treatment across sub populations then the issues addressed in Table 1 apply. In practice these analyses tend to be somewhat problematic, as we tend to have only limited numbers of trials and the magnitude of differences in relative effectiveness might normally be expected to be modest in comparison with the overall treatment effect.

In individual trials the estimation of interaction effects using generalised linear models is quite commonplace and methodologically uncontroversial. However in meta analysis the situation is not so well established. Different approaches are possible. For example a *simple test for differences* may be conducted between the estimates of benefits in different subgroups. Such an approach has appealing simplicity. Alternatively a *meta regression* approach can be used, adding a factor to describe the subgroup of trials and examining whether this achieves statistical significance. In the simple situation where there are two discrete subgroups the two approaches are analogous, and the important point is to ensure that the test for interaction is conducted an not just estimation of the main effects.

Where studies contributing to a meta analyses have different patient level characteristics at a sub trial level (different proportions of subjects with a specific confounding condition such as heart failure for example) so called ecological bias (confounding at the study level) can lead to invalid results, and individual patient data for each contributing study may be required in order to address this question in a way that avoids bias. Where individual patient level data are not available and the subgroup effects of interest are not coterminous with trial the exploration of subgroup effects using meta regression may be considered to be largely exploratory.

5. Handling of Heterogeneity

Heterogeneity relates to systematic differences in the results of treatment effects between randomised trials; that is variation in treatment effects which exceed that which might be expected by chance alone. In this circumstance, ignoring heterogeneity would be analogous to ignoring overdispersion in binomial or Poisson models, and would over state the statistical precision of the results. There are several methods to adress and analyse heterogeneity within multiple studies, the most common approaches shall be discussed in the following section.

The *test for heterogeneity* is not a good basis for a decision on whether or not to summarise included studies using an assumption of a single (fixed) treatment effect. Instead it is appropriate to prespecify an analytic approach as the principal analysis on the basis of prior knowledge of the condition and the question, and to explore the consequences of that decision in sensitivity analysis through comparison with different analytic approaches. In a review of beta blockers after myocardial infarction, we felt, a priori, that it made good clinical sense to use a fixed effects approach as our principal analysis approach where trials compared the same pharmacological agent with the control condition [Freemantle 1999]. Thus, for example, all trials of the agent atenolol were pooled using fixed effects approaches. However, although we felt that it make good sense to ask the overall question of whether beta blockade reduced mortality, we did not feel it appropriate to consider that each agent would have the same effect (due to differences in formulation and mode of action) and so this over arching question was approached using random effects methods. In such circumstances it makes much more sense to use an approach which accounts for systematic differences in treatment effects between trials. Additionally there is an argument for parameterising the difference between treatment effects and using a method which does not simply take the observed heterogeneity as the true heterogeneity. In fact a test is of little use here as it is possible to not achieve significant heterogeneity but for the differences between estimates of treatment effect to indicate qualitatively different results, or conversely for significant heterogeneity to be associated with no important concern regarding the overall treatment effect estimate.

The I^2 statistic [Higgins 200]) may be used to describe the extent of heterogeneity, and has some limited advantages beyond the more familiar P value from a heterogeneity test. However the I^2 is simply the difference between the residual deviance and the degrees of freedom divided by the residual deviance, relying on the fact that we would expect by chance alone that the residual deviance to increase by 1 for each additional trial included in a meta analysis. The residual deviance is calculated as the weighted sum of squared differences between individual study effects and the pooled effect across studies, with the weights being those used in the pooling method. Although going some way to quantifying the extent to which heterogeneity may be present, the limitation of the I^2 statistic is that it does not represent the clinical importance of the observed heterogeneity.

A more appropriate alternative to forms of hypothesis testing is to *examine the extent to which heterogeneity in treatment effect between trials influences the overall estimate* of treatment effect from the trials, and the variance of this measure. Such an analysis is provided by a random effects analysis, which includes heterogeneity in both the calculation of the treatment effect and its variance. If a random effects approach is used, this does lead to the potential of being open to publication bias since in the context of heterogeneity small studies are given higher relative weights due to the inclusion in the weighting of a constant for all trials derived from the between study variance. As smaller studies are particularly open to publication bias giving greater weight to these potentially biased studies may lead to a biased result. Similarly, where there are only a few studies, standard methods for random effects meta analysis may misrepresent the effects of heterogeneity as assume that the observed heterogeneity is the true heterogeneity. More sophisticated 'full random effects' methods based upon numerical simulation techniques enable the parameterisation of the random effects and thus include uncertainty on the estimation of heterogeneity in the pooled results. [Smith 1995]

Probably the best available approach to assessing the potential size of such a bias is to contrast the results of any random effects analysis with a fixed effects approach, having prespecified one of these analyses as the principal analysis, and one as a supportive analysis. Similarly, the Egger test (1997), or where appropriate the related Harbord test (2005) may provide a useful estimate of the asymmetry of a funnel plot, which can be helpfully used to describe the extent to which a meta analysis appears open to publication bias.

A further and related approach is to attempt to *explore the potential causes of heterogeneity*, although the extent that this will be possible is often limited, and the approach is often poorly addressed. One inappropriate practice is to examine the heterogeneity test statistic, removing trials from the analysis in order to minimise observed heterogeneity. Because of the low statistical power of the test this approach cannot really take us forward, and is a form of over fitting which will lead to model optimism. Alternatively a limited number of hypotheses (usually one) may be examined in meta regression analyses where we attempt to explore the causes of heterogeneity. One example where this was conducted involved an examination of the role of comparator dose in trials of atypical antipsychotic drugs versus conventional agents, in which the dosage of the conventional agents varied substantially [Freemantle 2000], providing a potential explanation for the observed differences in treatment effect between trials and agents. However although potentially useful in some circumstances, meta regression approaches are not a panacea, and are based upon observational differences between randomised trials (because subjects are randomised within trials, but not between them), but meta regression approaches may provide a useful method for estimating such indirect treatment effects.

Thus a strategy for addressing heterogeneity should include appropriate random effects analysis, sensitivity analyses using different analytic approaches, and may include some limited exploration of the potential causes of heterogeneity.

There are occasions when the term ‘clinical heterogeneity’ is used. This is, arguably, not a useful concept and may in fact be damaging to the statistical concept of heterogeneity. That is because clinical heterogeneity appears to be used in place of more straight forward language such as ‘differences’ or ‘variability’ which are easily understood, and heterogeneity in the statistical sense does not merely mean differences, but as stated, instead means differences which exceed those which may be explained by chance alone and thus may be considered systematic.

6. Analysis of Adverse Events

RCTs provide the best way to identify convincing evidence of the effect of treatments for both positive effects and adverse events. The case of aprotinin provides an interesting and salutary example of where a single observational study [Mangano 2006; Mangano 2007] which is not in line with a meta analysis of randomised trials [Henry 2007], nor a subsequent observational study [Pagano et al 2008] has led to controversy on the appropriate use of this agent.

Non randomised evidence is confounded and unconvincing both on arguments that treatments are safe and that treatments are associated with untoward events. There are particular challenges where the outcome of interest relates to the level of risk and the potential selection of patients. For example patients with type 2 diabetes in observational studies taking more oral agents are on average experiencing less adequate glycaemic control than those taking fewer oral agents [Calvert 2007]. This is not because additional oral agents lead to poorer control, but because an additional oral agent is preferentially used in patients with poorer control.

It may be considered that the relatively small numbers of subjects recruited to randomised trials leads to low power to detect important safety differences. In the situation when data are spare simulation analysis has indicated that the Peto method may provide least biased estimates of treatment effects, except where there is a substantial imbalance between the numbers of subjects randomised between the treatment of interest, when the Mantel Haenszel approach (with or without continuity correction), logistic regression approach or theoretically exact approach should be used.[Bradburn 2007]

However, low power really suggests that we should be doing more trials, rather than using lesser evidence to attempt to make important decisions about drug safety, although observational data may play a supportive part helping to identify potential safety problems or providing supportive data. This issue is really the preserve of the regulatory agencies rather than reimbursement agencies, and is addressed on a case by case basis. For example, it was regulatory concern regarding heart failure risk which led to the PROactive trial [Dormandy 2005], an outcomes trial of over 5000 subjects followed up for several years.

Thus drug safety per se should not be central to the activities of any appraisal institute. However where a treatment is considered overall to be safe, and associated with an acceptable ratio of untoward events to treatment benefit, these event rates should be considered as part of value for money arguments and should normally be estimated from randomised data. Indirect comparison refers to the situation where inference is drawn between two or more different products which have not been compared directly, but for which data contrasting against a different outcome may be available, in the context of regulatory studies this is often comparison with placebo.

The potential for analysis of multiple comparators in so called mixed treatment effects models has been advanced by several authors [notably: Higgins & Whitehead 1996; Lu & Ades 2003]. These approaches capitalise on the developments in non linear mixed models which enable more sophisticated modelling strategies to be implemented, which can “borrow weight” from indirect comparisons in estimating treatment effects. These methods are an interesting and logical development and have considerable potential in health technology appraisal where the aim is to provide the best possible summary of the data available. Lu and Ades (2003) indicate correctly that the methods for analysis are analogous to those used in incomplete block designs, [Cochran & Cox 1992] where not all treatments are present in each block, and the analysis seeks to recover inter-block information. However one major difference exists which profoundly affects our interpretation of the results, which is that treatments are randomised to different blocks in incomplete block designs (original designed for glass house experiments in agriculture), which is not the case for patients who clearly have not been randomised between different contributing trials. Further in incomplete block designs where we seek to estimate all treatment effects we seek to achieve balance in the strategy for the omission of treatments between blocks, which is conducted using a series of randomised balanced lattices. No such prospective balance may be achieved when we contrast different treatments with mixtures of direct and indirect comparisons.

Thus mixed treatment comparisons, while of considerable interest, suffer from two difficulties. First, the happenstance in trial comparison will mean that treatments present in the trials will be estimated with different levels of precision, and more fundamentally, that trials differ on characteristics which cannot be presumed to be attributable to chance alone. That having been said, the mixed treatment comparison approach maximises the information available from the existing randomised trials and their comparisons. Where large scale randomised trials are available for all relevant treatment comparisons, it is not necessary to perform mixed treatment comparisons. When all such comparisons are not available, mixed treatment comparisons are appropriate and represent the best available estimate of the treatment effect of interest. As Caldwell et al (2006) indicate, the methods for mixed treatment comparisons are currently under development and the exact approach to take in specifying these models remains controversial. Thus we await further methodological development with considerable interest.

7. Combination of evidence for substances to assess evidence of class effect

Pharmaceutical products are often members of a related class of drugs, and this can lead us to making inference from trials of one agent to the effects of another agent or assuming a similar effect of all substances within the same class. This is particularly difficult and often the subject of considerable controversy. Under the Australian system, the first drug in class has to establish that it is providing a worthwhile benefit, and further drugs which seek reimbursement later have to either establish that they achieve greater cost effectiveness, or establish that they have similar effectiveness and lower or equal costs.[Department of Health & Ageing 2007]. From a statistical point of view, establishing that a newer pharmaceutical agent has similar, or non inferior, effectiveness to an established agent is not straight forward and requires substantial numbers of subjects randomised in comparative trials. Equally, there are multiple criteria which may be used to establish equivalence, particularly in the context of assessment rather than regulatory approval, and the trial programmes for different agents may bring different information and value. For example, the MERIT trial of metoprolol versus placebo in patients with heart failure (Hjalmarson 2000) demonstrated that a similar effect to that achieved in earlier beta blocker trials in that area, but extended the findings substantially to an older group of subjects. The Australian and Scottish systems take the position that they leave it up to the sponsor of the product (usually the manufacturer) to make the appropriate claim, and then appraise the validity of that claim. That appraisal may include the assimilation of several different sometimes conflicting pieces of information. Thus rather than there being a statistical answer to the question of whether two or more agents share a *class effect*, a more appropriate and workable solution in health policy is to assess each case on its merits. However, the appropriate starting point of an appraisal should not be to simply assume a class effect.

8. Where the primary endpoint of a meta analysis is not the same as the primary endpoints for the original study

Within benefit assessment, all available trial information on a topic should be used for data synthesis. These individual trials were performed for pre-defined questions and often focus on different aspects of the treatment. As a consequence, the assessment of efficacy across several trials is often and quite correctly driven by different criteria from that where we are assessing a single study. As mentioned above, although frequently done, defining a primary outcome measure in a meta analysis does not have the same value as predefining a primary outcome or alpha spending plan in the prospective design of randomised trials and such an approach is open to selection bias since the trials to be included already exist and the results are available in the public domain, with the exception of the special case of the prospective meta analysis. But as the aim of an appraisal is to gain an appropriate understanding of the costs and benefits of a potential health policy decision (eg to reimburse a specific product or not) then the most appropriate way to gain an answer to that question is to include appropriate estimates of the attributes (eg the range of relevant outcome measures) and costs of treatments, and information on their variability, in a fully probabilistic economic analysis. The alpha spending plans of the original investigators may be disregarded in this process, as the meta analysis is a new appraisal and brings new questions.

9. Methods for Combining Classes of Evidence / Grading of evidence in Meta-Analyses

It is well established that randomised trials are the only design with the potential to provide unbiased estimates of treatment effects. As treatment effects are often relatively modest when compared with the potential magnitude of biases, only randomised trials may be relied upon to provide estimates to populate decision models regarding pharmaceutical reimbursement questions. Of course randomised trials may be quite artificial, and basing a decision only on data that comes from randomised trials is probably normally inappropriate. For example, the event rates observed in clinical trials are often rather low in comparison with the event rates from an apparently similar population in general clinical practice. This does not take away the validity and importance of using randomised trials for estimates of effectiveness, but instead means that we should be applying estimated treatment effects from trials, to relevant populations which are defined similarly to the population in the health system of interest. This can be achieved quite readily in economic modelling.

10. Sensitivity analysis

Prespecification is an important aspect of methodological rigour, although as discussed, it has only limited applicability in meta analysis which is normally of retrospective nature. Sensitivity analysis may be used very helpfully to demonstrate that a result is, or is not, robust to the initial assumptions or approach. This technique is used, for example, in the analysis of non inferiority trials where an ‘on treatment' or ‘per protocol' analysis is recommended in addition to the conventional intention to treat analysis in order to address the issue where subjects switch treatment regimens (often the control group switching to active therapy on disease progression for example, occasionally coupled with a proportion of intervention group subjects never actually receiving adequate therapy) and thus the conventional intention to treat analysis risks comparing groups where the treatment regimens differ little between the groups making the likelihood of apparent equivalence between the groups rather likely on the basis of the original randomisation. Of course the per protocol analysis remains biased, but may be useful to indicate whether cross over between treatment groups has indeed diluted treatment effects.

In the situation of meta analysis a range of sensitivity analyses may usefully be undertaken in addition to those prespecified in the protocol. These are listed in Table 3.

Table 3: Range of sensitivity analyses conducted during meta analysis

- Investigating a range of outcomes in addition to that specified as the primary objective of the meta analysis
- Investigating the effects of missing data in individual trials on the outcome of meta analysis
- Investigating the impact of fixed and random effects analysis

- Investigating the impact of using different statistical approaches to fixed and random effects meta analysis
- Investigating potential causes of heterogeneity using subgroup analysis, tests for interaction and meta regression analysis
- Investigating the effects of study quality through selective removal or introduction of methodologically questionable studies to assess the impact of their results

Just like in individual trials, it is normal for those planning meta analyses to identify a range of potentially important outcomes to include in the analysis. As discussed previously, the opportunity to prespecify the analyses conducted prior to the availability of data is not generally available in meta analysis, and this is best overcome through the examination of a limited number of relevant outcome measures, and the need for a high level of statistical precision in order to be convinced of the outcome.

We may generally be more convinced of a treatment being investigated in a meta analysis where an effect is replicated across a range of different outcome measures with a strong level of statistical precision. Thus for example where we saw that a treatment resulted in reduced fatal myocardial infarction, and reduced non fatal myocardial infarction, this finding would be particularly convincing.

missing data

Data will rarely be collected or reported on all outcomes of potential interest in trials contributing to a meta analysis. Similarly some patients may be lost to follow up in trials, and their data not available for synthesis in meta analysis. To the extent that these missing data may be considered *missing at random*, this situation merely affects the width of the confidence intervals on estimates of effect. However, where results may be selectively reported or where missingness may be non random (eg where data are not available on a subset of subjects because they are doing particularly well or badly – a commonly used example is that data may be missing where a subject has died!).

Different approaches to the inclusion of missing data may prove helpful in addressing the uncertainty caused by patients lost to follow up or otherwise not available to provide outcome information at certain time periods. Thus we might consider undertaking three different analyses: an analysis of the available data in trials; an analysis of the available data in trials, making a pessimistic imputation for missing data (eg presuming a bad outcome for those subjects who are missing); fitting an interaction term to estimate the effect of missingness on the overall effect size for treatment. Assessing the effect of missingness in this way may avoid inappropriate fixed inclusion and exclusion criteria for trials. Trials with strong results but in which a number of patients are lost to follow up may still provide convincing evidence, while trials with modest results and modest loss to follow up may provide less convincing results when drop out is taken into account while otherwise being apparently eligible for inclusion using a fixed rule for drop out rates.

Approaches to effect size

Results of a meta-analysis sometimes vary between analytic approaches. We may therefore be more convinced of a treatment being investigated where the effect is robust to a range of different approaches to meta analysis, although in general terms these may not be considered equal.

Approaches based upon the odds ratio will generally be superior to those based upon other metrics when attempting to assess the statistical precision of a difference between treatments in meta analysis, or a test for interaction. This is because the odds ratio scale enables more straight forward calculation of the standard error, which can include theoretically exact approaches [Martin 2000]. Similarly full random effects approaches based upon numerical simulation techniques have advantages in enabling estimation of the between study treatment effects as a parameter which may be estimated from the data with uncertainty [Smith 1995]. Such an approach has the advantage over a standard random effects approach where the observed heterogeneity is assumed to be the true heterogeneity even though this may only be considered true where a meta analysis includes a large number of large trials (in which circumstances the full random effects approach will asymptote to the standard approximation).

Estimation using the relative risk (or indeed the hazard ratio) has some advantages of interpretation, since the relative risk estimator may be applied directly to an expected rate of events and is frequently of particular use in economic analyses. Thus, if we would expect 100 events in an untreated cohort, and the effect of treatment has been estimated in a meta analysis as providing a relative risk of 0.70, then we might reasonably expect to observe 70 events if the cohort receives treatment. The same may only be considered true of the odds ratio where the rate of events is very low in contrast to the number of subjects at risk, where the odds ratio will approximate the relative risk.

The absolute risk difference is a problematic metric; potentially providing maximal interpretation although being least robust to the confounding effects of trial design or patient severity when compared with the ratio scales described above. In general terms some kind of analysis and presentation which brings together both a robust estimator on a ratio scale and an assessment of the absolute value of treatment would seem optimal. Where the relative risk is used on the ratio scale in order to increase the degree of interpretation, it is important that a more robust secondary analysis using more robust methods has been conducted to ensure that the results on the relative risk scale are not an artefact of inadequate variance estimation.

Random or Fixed Effects Models

The choice of fixed or random effects analysis may prove surprisingly controversial in meta analysis, and may be driven by two non orthogonal questions. First, whether the specified approach to meta analysis fits the observed data well, and second, reasonable expectations on the degree to which treatment effects may differ between studies.

The first question which addresses the issue of the extent to which an analysis fits the data adequately is analogous to the issue of over dispersion in generalised linear modelling. Where there is a degree of clustering on the study level strata then the assumptions of the binomial distribution will be violated. In such cases there is a strong argument for the use of random effects approaches in which both the treatment estimate and its variance (and thus confidence intervals) are modified to account for the between study variability.

If we cannot account for differences in treatment effects between trials, because for example we may believe that a treatment is being delivered differently in different trials, but in a manner which cannot be adequately explained or categorised, then the random effects approach is optimal since the results are described across the range of available trials. Inevitably in the presence of heterogeneity of treatment effects the random effects approach will provide a wider confidence interval than the fixed effects approach.

It may be suggested that potential causes of observed heterogeneity should be explored in meta analysis, although it is of limited practical usage since there are rarely sufficient trials of sufficient size to enable sensible or prespecified analyses to be made (see chapter 6) . A limited number of subgroup analyses may be addressed, which should be biologically plausible and again if possible identified in advance (see chapter 5). However, overall an appropriately specified random effects approach is probably more useful. It is not generally helpful to consider the extent to which excluding studies with specific characteristics reduces the degree of heterogeneity as such tests are low powered, and loose power further when contributing trials are excluded. Similarly the question of whether to analyse using fixed or random effects is not appropriately dealt with through including a decision rule driven by a P value (eg $P < 0.10$) since relatively modest observed heterogeneity may be associated with large differences when measured on a relevant scale, and substantial statistical heterogeneity may have little practical consequence especially where there is a large number of large trials and the test has greater power. In both circumstances analysis using random effects and contrasting the results with appropriate fixed effects approaches will provide an objective measure of the degree of importance of observed and potential heterogeneity and is

A more appropriate and robust rationale for the removal of trials from a meta analysis is provided by concerns about the methodological quality of some contributing trials. It is inappropriate to exclude trials, even where they have some methodological challenge, without also considering a sensitivity analysis including the relevant trials. Thus the Blood Pressure Treatment Trialists (2000) conducted analyses both including and excluding the CAPPP trial due to evidence of bias in the random allocation process in that trial of the primary treatment of modest hypertension. Completely excluding potentially relevant trials which do not meet arbitrary entry criteria such as length of follow up or loss to follow up may harm patients since the results from such analyses may be of lesser statistical power and less able to distinguish between treatments than those which include them. If such trials are included their effects may be examined in meta regression analyses or sensitivity analyses to establish the size and direction of any such bias.

11. Estimation versus testing: how should the results of the assessment process be interpreted?

When appraising the attributes of a treatment, estimation is of particular importance in EBM and health technology assessment. The obsession with statistical significance is probably related to the regulatory need to reduce the assessment of clinical trials to a single yes / no answer. By contrast, once that answer has been achieved, more can be gained from a consideration of the results on a range of different outcomes, and the best estimates (plus uncertainty) of these estimates. Thus rather than providing p values, which have an uncertain interpretation, a more appropriate approach is to provide an estimate of the treatment effect coupled with 95% confidence intervals describing the plausible range of the true treatment effect.

Although controversial, it is apparent that results need not in fact be statistically significantly better than an alternative to present the most worthwhile treatment option. This is clearly the case when a treatment comparison provides an estimate of benefit which does not quite achieve statistical significance, but where the treatment of interest is overall similarly costly to use compared to the alternative treatment. In such a comparison, almost all positions within the 95% confidence interval favour the similarly costly but not quite statistically significantly superior treatment option, and thus on balance that option would represent the best choice, albeit with considerable uncertainty.

Such questions should normally be addressed in rigorously conducted economic models. A correctly specified fully probabilistic economic analysis will incorporate the uncertainty of estimates of treatment effect alongside appropriate distributional information for other estimates used to populate the statistical model, and this uncertainty will be reflected in the width of credibility intervals derived around the cost-effectiveness estimate of interest.

12. References

Australian Govt Department of Health and Ageing. Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (version 4.2). Commonwealth of Australia, Canberra 2007.

Blood Pressure Treatment Trialists' Collaborative. Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials. Lancet 2000; 355: 1955-64

Bradburn MJ, Deeks JJ , Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events. 2007; 26: 53 – 77.

Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ 2005; 331: 897-900

Calvert MJ, McManus R, Freemantle N. The management of people with type 2 diabetes with multiple oral hypoglycaemic agents or insulin in primary care: retrospective cohort study. British Journal of General Practice, 2007; 57: 455-60.

Cleland JGF, Freemantle N, McGowan J, Clark A. The Evidence for Beta-Blockers equals that for ACE Inhibitors in Heart Failure. British Medical Journal 1999;318 824-5

Cochran WG, Cox GM. Experimental Designs. New York, John Wiley & Sons, 1992

Dormandy JA, Charbonnel B, Eckland DJA, Erdmann E, Massi-Benedetti M, Moules IK et al. Secondary prevention of macrovascular events in patients with type 2 diabetes in the PROactive study (PROspective pioglitAzone Clinical Trial In macroVascular Events) a randomised controlled trial. Lancet 2005; 366: 1279-89

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. British Medical Journal 1997; 315: 629-34.

Freemantle N, Cleland JGF, Young P, Mason J, Harrison J. What is the current place of β-blockade in secondary prevention after myocardial infarction? A systematic overview and meta regression analysis. British Medical Journal 1999; 318: 1730-7.

Freemantle N, Henry D, Maynard A, Torrance G. Promoting cost-effective prescribing: Great Britain lags behind. British Medical Journal 1995; 310: 955-6.

Freemantle N. How well does the evidence on pioglitazone back up researchers' claims for a reduction in macrovascular events? British Medical Journal, 2005; 331: 836-8.

Freemantle N. Interpreting the results of secondary endpoints and subgroup analyses in clinical trials: should we lock the crazy aunt in the attic? British Medical Journal, 2001; 322: 989-91.

Geddes JR, Freemantle N, Harrison P, Bebbington PE for the National Schizophrenia Guideline Development Group. Atypical antipsychotics in the treatment of schizophrenia: systematic review and meta-regression British Medical Journal 2000; 321: 1371-6

Geddes JR, Freemantle N, Harrison P, Bebbington PE for the National Schizophrenia Guideline Development Group. Atypical antipsychotics in the treatment of schizophrenia: systematic review and meta-regression British Medical Journal 2000; 321: 1371-6

Gottlieb SS, McCarter RJ, Vogel RA. Effect of β-blockade on mortality among high risk and low risk patients after myocardial infarction. New England Journal of Medicine 1998; 339: 489-97

Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, Wilson MC, Richardson WS. Users' guides to the medical literature: xxv. Evidence-based medicine: principles for applying the Users' Guides to patient care. JAMA 2000; 284: 1290-6.

Harbord RM, Egger M, Sterne JAC. A modified test for small-study effects in meta-analyses of controlled trials with binary endpoints. Statistics in Medicine 2006; 25: 3443-57

Henry DA, Carless PA, Moxey AJ, O'Connell D, Stokes BJ, McClelland B, Laupacis A, Fergusson D. Anti-fibrinolytic use for minimising perioperative allogeneic blood transfusion Cochrane Database of Systematic Reviews: Reviews 2007 Issue 4 John Wiley & Sons, Ltd Chichester, UK

Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003; 327: 557-60.

Higgins JPT, Whitehead A. Borrowing strength from external trials in a meta-analysis. Statistics in Medicine 1996; 15: 2733-49

Hjalmarson Å, Goldstein S, Fagerberg B, Wedel H, Waggstein F, Kjekshus J et al. Effects of controlled-release metoprolol on total mortality, hospitalizations, and well-being in patients with heart failure. The Metoprolol CR/XL Randomized Intervention Trial in Congestive Heart Failure (MERIT- HF). JAMA 2000; 283: 1295-302

Lau J, Antman EM, Jimenez-Silva J, Kupelnick B, Mosteller F, Chalmers TC. Cumulative meta-analysis of therapeutic trials for myocardial infarction. N Engl J Med. 1992; 327: 248-54.

Lu G, Ades AE. Combination of direct and indirect evidence in mixed treatment comparisons. Statistics in Medicine 2004; 23: 3105-24]

Mangano DT, Miao Y, Vuylsteke A, Tudor IC, Juneja R, Filipescu D, Hoeft A, Fontes ML, Hillel Z, Ott E, Titov T, Dietzel C, Levin J. Mortality associated with aprotinin during 5 years following coronary artery bypass graft surgery. Jama. Feb 7 2007;297(5):471-479.

Mangano DT, Tudor IC, Dietzel C. The risk associated with aprotinin in cardiac surgery. N Engl J Med. Jan 26 2006;354(4):353-365.

Martin DO, Austin H. An exact method for meta-analysis of case-control and follow-up studies. Epidemiology 2000;11(3):255-260

National Institute for Clinical Excellence. Guide to the methods of technology appraisal. http://www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf accessed 21/3/08

Owen A. Intravenous blockade in acute myocardial infarction should be used in combination with thrombolysis. BMJ 1998; 317: 226-7

Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116: 78-84

Pagano D, Howell NJ, Freemantle N, Cunningham D, Bonser RS, Graham TR, Mascaro J, Rooney SJ, Wilson IC, Keogh BE. Bleeding in Cardiac Surgery: The use of Aprotinin does not affect survival. Journal of Thoracic and Cardiovascular Surgery, 2008; 135: 495-502.

Perneger TV. What's wrong with Bonferonni adjustments? BMJ 1998; 316: 1236-8

Poole-Wilson PA, Swedberg K, Cleland JGF, Di Lenarda A, Hanrath P, Komajda M et al. Comparison of carvedilol and metoprolol on clinical outcomes in patients with chronic heart failure in the Carvedilol Or Metoprolol European Trial (COMET): randomised controlled trial. Lancet 2003; 362: 7-13

Smith TC, Spiegelhalter DJ, Thomas A. Bayesian approaches to random effects meta analysis: a comparative study. Statistics in Medicine 1995; 14: 2685-99