2 November 2010
EMA/314182/2011
Human Medicines Development and Evaluation

Benefit-risk methodology project

Comments received from Dr William Holden on the Work Package 2 report

7 Westferry Circus • Canary Wharf • London E14 4HB • United Kingdom
Telephone +44 (0)20 7418 8400 **Facsimile** +44 (0)20 7<XXX XXXX>
E-mail info@ema.europa.eu **Website** www.ema.europa.eu

Comments on Benefit-risk methodology project: Work package 2 report.

I would like to commend the EMA for its foresight and proactive engagement with the new and burgeoning field of quantitative benefit-risk analysis (QBR). You have set the standard for regulatory agencies around the world. It is my opinion, however, that the Work Package 2 Report may not meet the standards set or provide adequate information about currently available QBR methods. From the outset, the document has a narrow focus and is self-serving when not self-referencing. For example, on page 6 of the .pdf version, the criteria used to assess benefit risk models are evaluated according to precepts ostensibly from a publication of one of the co-authors (L Phillips) on multi-criteria analysis, which happens to be one of the three models (of 18 models evaluated) found to be 'sufficiently comprehensive' to be of any utility.

My goal is not to defend one method against others, although I will certainly differentiate the truth from the cursory and rather inadequate description found in the report of my methods (and let others do the same, should they be so inclined). Rather, it is to point out some inconsistencies in the document that will serve neither the goals of EMA nor the progress of science.

Criteria for logical soundness

The criteria for logical soundness are as follows (taken from page 6, .pdf version):

- The overall benefit-risk evaluation is decomposed into separate elements that are demonstrated theoretically and/or empirically to be meaningful.
- The elements are recombined according to a theoretically sound rule.
- The approach is coherent, that is, it ensures that related decisions based on the approach do not contradict each other or the objectives that are to be met.
- The approach aids rational thinking about benefits and risks.
- The approach gives results that do not change relative evaluations when alternatives are added or removed.

Many of these criteria are circular in the philosophical sense and not defined. The dilemma with such an approach is that it risks assumptions that can be imposed only in a subjective manner and, ironically, ends up less than logically sound. For example, the first criterion states that the evaluation must be decomposed into 'meaningful' elements. Nowhere is it explained why such decomposition is necessary (indeed, one might argue that such decomposition is antithetical to the spirit of benefit risk analysis, which should imply a synthesis of data) and certainly nowhere is the concept of 'meaningfulness' explained. If meaning, like beauty, is in the eye of the beholder, then the evaluations of models contained in the document portend to be naught but subjective observations.

The second criterion is similarly subjective, as 'theoretically sound' remains an undefined concept. Indeed, the word 'sound' appears often enough in the document to appear as if it has a singular definition, which is hardly the case.

Thus far, elements in a model should be decomposed meaningfully and then recomposed soundly.

I will skip over the next two criteria and focus on the final one which, I admit, I have difficulty understanding. I'm not sure what a 'relative evaluation' is and I certainly don't understand why a model must not yield results that change this 'relative evaluation' when 'alternatives are added or removed'. Isn't adding alternatives what the basis of benefit risk analysis is all about? For example, say a benefit risk (BR) model finds that a new diabetes drug has a positive BR profile based on the data available. Is it not possible that an added alternative of, say, adding this new diabetes drug to baseline metformin might alter that profile? Is it not possible that an unknown, as yet unobserved interaction with another drug might alter that profile? Might there not be subgroups –pregnant women, patients with certain polymorphisms, to name but two – who would have different BR profiles? Again, I'm not sure what is meant by 'alternatives' here, so I may be off target (and if so, I apologize).

Other evaluation criteria

I would like to comment briefly on the other criteria presented.

The criterion for Comprehensiveness seems impractical, as there are very few statistical methods that can comply with these demands.

The criterion Acceptability of Results seems to be requesting a model's ability to check on data consistency and consistency in people's judgments. Evaluating data consistency does not appear to be a *sine qua non* for doing BR analysis but rather is the purview of epidemiology and biostatistics. Obviously the strength of a model will partially depend on the quality of the data that are put into it, but that is true of many [non-BR] models and it is not clear how a BR model, *per se*, would be able to evaluate such consistency. I have been doing epidemiology for almost 20 years and if there is one thing I have learned, it is that there is little consistency in people's judgments – and less point in checking them.

The Practicality criterion is largely subjective. No rationale for the required linear growth in model extensions is provided (nor what this even means).

I agree with all the points under the Generativeness criterion.

Specific comments on the evaluation of NNT/NNH

To the methods at hand or, more precisely, the evaluation of my methods, which are the relative-value adjusted NNT and the Minimum Clinical Efficacy approaches. I will reproduce the section *in toto* below (copied and pasted from the .doc version) and I have highlighted some of the more interesting comments which are referred to following this section. Suffice it to say that by referring to these methods as 'NNT/NNH' – a misnomer – sets up a straw man argument that mischaracterizes the methods, goals, and output.

The NNT/NNH section from the report

3.3.2 NNT/NNH

NNT, the 'number needed to treat', is the average number of patients that would have to be treated in order for just one of them to receive the expected favourable effect. NNH, the 'number needed to harm', is the average number of patients that would have to be treated in order for just one person to experience a particular unfavourable effect. Both NNT and NNH are calculated as the inverse of the difference in proportions of the effects between the treatment and control groups [65]:

$$NNT(NNH) = \frac{1}{p_t - p_c}.$$

The denominator is often referred to as the absolute risk reduction. For a given disease, a smaller value of the NNT (i.e., a big improvement in the probability of a favourable effect—which might mean a reduction in the chance of a negative outcome) is better as it indicates a drug that is effective for more people, while a larger value of the NNH (a small increase in the chance of an undesirable effect) is preferred because the adverse effect caused by the drug is so rare. Thus, a small NNT means fewer people have to be treated to see one favourable effect, while a large NNH shows that only by treating many people will just one person show the unfavourable effect. Of course, that is true only on average, since the proportions are uncertain.

Our view

NNT and NNH might seem practical because of their simplicity but this simplicity is deceiving. Their main problem is that they cannot be combined to determine if benefits outweigh risks because neither statistic takes account of clinical relevance [66]. For example[1], suppose a drug is found to reduce the incidence of death in vCJD sufferers from 100% to 90%. The NNT is 10. Now imagine a drug that reduces pneumonia deaths from 50% to 40%. The NNT is also 10, but it seems unlikely that anyone would consider these two cases of equal value, and that is just for comparing NNT for two different disease states. A comparison of NNT with NNH for the same disease state requires considering the clinical significance of the favourable event with the unfavourable event: if NNT=NNH, that doesn't mean that the benefit-risk ratio is one. A further problem arises in attempting to

[1] Thanks to Rob Hemmings for this example, and for bringing to our attention other problems about NNT and NNH

apply these statistics to outcomes over time—different values of the statistics would be obtained at different time periods, as for example shown in Kaplan-Meier curves. This difficulty led Hildebrandt et al [67] to conclude "there is much room for improvement in the application of the number needed to treat to present results of randomised controlled trials, especially where the outcome is time to an event."

The underlying problem here is that preference judgements based solely on differences in probabilities violate the criterion of logical soundness, and no amount of 'fixing' the statistics can overcome this problem. In decision theory, probabilities multiply by utilities and it is the probability-weighted utilities, i.e., expected utilities, that are compared, not the probabilities themselves. It is expected utilities that express clinical relevance as well as the probability of realising the effects.

More generally, preferences cannot be well informed by proportions, where both numerator and denominator can take on different ranges of values, with the result that the same proportion could result from very different base conditions. Doubling a survival rate from one month to two months is surely not equivalent in preference to a doubling from one year to two years. Thus, relative risk, p_t/p_c, by itself also violates logical soundness. As Fahey et al have shown empirically [68], different measures yielding the same results can lead to different preferences.

Finally, we should add that Holden's suggestions [65] to incorporate relative utility values (RVs) into the NNH calculations, and to apply minimum clinical efficacy (MCE) analysis for comparing therapeutic options, use probability differences in these models, which can lead to failures of logical soundness.

Note that our critique here is aimed only at the usefulness of NNT/NNH for decisions by drug regulators. We have not formulated a view about applications for other purposes, such as communication by physicians to their patients.

Comments on the above section, especially regarding the highlighted sentences

Overall, there appears to be little understanding of either method (of note, little mention is made of the Minimum Clinical Efficacy model, which is the more comprehensive of the two models).

Of course, that is true only on average, since the proportions are uncertain.

The first paragraph is an inadequate description of the NNT approach, although it is technically correct. The fact that NNTs are 'true only on average' is, of course, equally true for the results of every clinical trial ever performed (although it could be added 'only on average…for people agreeing to participate and sign an informed consent document'). The qualifier 'since the proportions are uncertain' is odd – all data, at least all sampled data, have some uncertainty to them and it is usually neither possible nor ethical to study an entire population of patients (see Practicality criteria, op cit.). For that reason, the science of statistics (the quantification and measure of uncertainty) was invented.

Their main problem is that they cannot be combined to determine if benefits outweigh risks because neither statistic takes account of clinical relevance [66]

The idea that neither NNT nor NNH 'takes account of clinical relevance' is entirely erroneous. NNTs take account of clinical relevance more than most measures: in addition to being advocated by numerous medical journals, including BMJ, the New England Journal of Medicine, JAMA, and Annals of Internal Medicine – as well as by the Consolidated Standards of Reporting Trials group (CONSORT) – many professional risk communication experts favor them. The very interpretation of NNT is clinical and the Clinical Epidemiology group at McMaster University has used this measure for decades for clinical decision making purposes. And let us not forget that the title of the article in which the concept of NNT was introduced is: An assessment of clinically useful measures of the consequences of treatment (Laupacis A et al, N Engl J Med, 1988;318:1728-33). Oxford University's Centre for Evidence Based Medicine website similarly includes NNTs in its list of EBM tools and describes the derivation of this 'clinically useful measure of therapy' (http://www.cebm.net/index.aspx?o=1044, accessed 1 November 2010). Further, reference 66 at the end of the highlighted sentence is a letter to the editor which is neither an academic nor scientific argument against NNT. Finally, a review by McQuay (Ann Intern Med 1997;126;712-20) stated that NNT 'has that clinical immediacy' [of clinical applicability], which is one reason it is such a popular measure. (As an aside, the whole thrust of both the RV-adjusted approach and MCE is that they take into account clinical relevance by using patient preferences.)

Now, this is not to deny the fact that there are detractors of NNT.

The NNT is also 10, but it seems unlikely that anyone would consider these two cases of equal value, and that is just for comparing NNT for two different disease states.

The next highlighted sentence reflects an apparent lack of understanding of the basic uses of NNT, as it compares NNTs for two different diseases, vCJD and pneumonia (and, one would suppose, therapies). The point is made that therapeutic intervention for both diseases results in a clinical situation in which 10 patients with either disease have to be treated to prevent one adverse outcome, which is the correct interpretation of NNT. The concept of NNT relates to frequency, a function of the specific disease, specific interventions, and specific outcomes. We compare NNTs of different interventions in the *same* disease, not different diseases. It is a straw man argument to suggest, as is done here, that there is something to be gained by considering 'these two cases of equal value' (why anyone would do this is unclear, any more than one would compare two odds ratios from two separate studies, each examining a different disease and risk factor).

A comparison of NNT with NNH for the same disease state requires considering the clinical significance of the favourable event with the unfavourable event: if NNT=NNH, that doesn't mean that the benefit-risk ratio is one.

The very next sentence conflates two issues by stating that there must be some consideration of the 'clinical significance' of the benefit and the risk – and then stating that if the NNT equals the NNH it does not mean that the benefit-risk ratio is one. As mentioned earlier, the thrust of the two models is precisely that the NNT and the NNH must be considered clinically – and weighed appropriately by patient preference data. It is not clear what a benefit-risk ratio of one means to Phillips et al, but if the relative magnitude of two equal quantities expressed as a quotient are the

same, then their ratio is indeed one. Researchers who use NNT generally adhere to the notion that the NNT must be less than the NNH for a drug to be possessed of a positive benefit-risk profile.

A further problem arises in attempting to apply these statistics to outcomes over time—different values of the statistics would be obtained at different time periods, as for example shown in Kaplan-Meier curves.

This sentence addresses the 'further problem' of the issue of time-varying conditions. The comment is true for any study that doesn't look at the time-varying nature of therapeutic effects (which is the vast majority of studies, of course, since this is very complicated). And this problem was recognized by the inventors of NNT and addressed in the original 1988 article and numerous reviews ever since, some of which have presented solutions for it (eg, Altman et al, BMJ 1999;319;1492-5).

This difficulty led Hildebrandt et al [67] to conclude "there is much room for improvement in the application of the number needed to treat to present results of randomised controlled trials, especially where the outcome is time to an event."

The gratuitous comment from Hildebrandt, which is mistakenly cited as a 2010 publication (it is actually 2009), is from a discussion of different ways of calculating NNT taking time variations, including censoring, into account. It is not a critique of NNT, only of the proper use of NNT.

The underlying problem here is that preference judgements based solely on differences in probabilities violate the criterion of logical soundness, and no amount of 'fixing' the statistics can overcome this problem. In decision theory, probabilities multiply by utilities and it is the probability-weighted utilities, i.e., expected utilities, that are compared, not the probabilities themselves. It is expected utilities that express clinical relevance as well as the probability of realising the effects.

The paragraph above introduces us to the 'underlying problem here', which appears to have nothing to do with NNT and is a comment on preferences, which were not previously discussed. And certainly nothing was mentioned about preferences 'based solely on differences in probabilities', whatever that may mean. The fact is that both my methods rely on expected utilities and the arguments made in the document either confuse, do not understand, or confound this issue.

More generally, preferences cannot be well informed by proportions, where both numerator and denominator can take on different ranges of values, with the result that the same proportion could result from very different base conditions. Doubling a survival rate from one month to two months is surely not equivalent in preference to a doubling from one year to two years.

This paragraph includes the comment that 'preferences cannot be well informed by proportions' again conflates two separate arguments. Following the methods proposed by Guyatt et al (JAMA

1999;281:1836-43) and Djulbegovic et al (Cancer Control 1998 5:394-405), both of which are based on classical decision analytic techniques, I developed my approaches, which are extensions of these methods. The argument about base [rate] conditions is common to many statistical analyses and is not specific to either of my methods. The argument is a continuation of the previous issue about time-varying conditions and has been addressed. The comment that a preference for a doubling of survival time from one month to two not being equivalent to a doubling from one year to two begs the question: if a patient values it such, it is so valued and the preferences can indeed be identical. The point here, and underscored by my models, is that the preferences are dependent on the underlying disease as well as the potential adverse events. Thus, a cancer patient may value an additional month of life as equivalent to a rheumatoid arthritis patient's valuation of an extra year of life.

Thus, relative risk, p_t/p_c, by itself also violates logical soundness.

Incredibly, the next sentence, reproduced above, appears. It is not clear what the context for this comment is, nor what point the authors wanted to make. Relative risk, as pointed out by Rothman et al (Modern Epidemiology, 3rd ed. 2008, Lippincott Williams & Wilkins) 'may be the most common term in epidemiology.' Zhang et al (JAMA 1998;280:1690-1) pointed out that 'Relative risk has become one of the standard measures in biomedical research.' Relative risk is the foundational measure for epidemiologic studies and the principal measure of cause and effect in randomized clinical trials. It is the basis for the approval of drugs and other products – and for their withdrawal. However, it is considered by the report authors to violate logical soundness.

As Fahey et al have shown empirically [68], different measures yielding the same results can lead to different preferences.

This sentence refers to a study by Fahey which purports to demonstrate that 'different measures yielding the same results can lead to different preferences.' The Fahey study, however, was a survey of health authorities' and health commissions' willingness to fund certain programmes given results presented in four different ways - this was a funding study, not a patient preference study. The report authors misstate the conclusions of the study: the correct conclusion is that different measures yielding *different* results can lead to different choices by policy makers. The results were presented as relative risk reductions, absolute risk reductions, percentage of event free patients, and NNT. No one would expect these methods to yield 'the same results' because they are different measures of outcomes and inherently different (eg, a percentage is not a risk difference). The use of the word 'preference' in the report is similarly misleading, as the discussion to this point was of 'preference' as 'utility', not 'preference' as 'policy choice'.

Finally, we should add that Holden's suggestions [65] to incorporate relative utility values (RVs) into the NNH calculations, and to apply minimum clinical efficacy (MCE) analysis for comparing therapeutic options, use probability differences in these models, which can lead to failures of logical soundness.

This paragraph suggests that weighing NNHs with relative values can lead to failures of logical soundness. Since the criteria for 'logical soundness' may themselves be logically unsound, and because the authors do not discuss how such weighing can lead to this condition, and because it has not been demonstrated how 'probability differences' lead to logical unsoundness, and because other researchers have, in fact, weighed NNHs with relative values (eg, Djulbegovic and Guyatt), this criticism seems particularly *ad hominem*.

Correct description of the models

I will present a brief description of the two models.

Both models are direct extensions of - and are rooted in - decision analysis. The work of Djulbegovic et al has repeatedly stressed and demonstrated this fact. The first method, the relative value adjusted NNT approach, weighs one or more adverse events by the relative value – defined as the patient preference for preventing the underlying disease as opposed to incurring one or more adverse events and whatever inconvenience may be associated with the treatment.

The output of such an analysis is an NNT and the adjusted NNH. The interpretation, following McAlister et al (JAMA 2000;283:2829-36) is the likelihood of being helped or harmed, ie, the reciprocal of the NNT divided by the reciprocal of the NNH. Thus, for example, if the NNT was 6 and the NNH (adjusted or otherwise) was 58, a patient would be almost 10 times more likely to derive benefit from the drug than harm.

The minimum clinical efficacy analysis is a more powerful and comprehensive approach that includes in its model the adverse event profile of two compared drugs, efficacy or effectiveness data of the drugs, and the morbidity and mortality of the underlying disease. The output is a graph on which is located the MCE point – the relative efficacy of a test drug – a point which must be met or exceeded by the test drug for it to be considered having a superior benefit-risk profile. Because the model includes the natural history of the underlying disease, it easily allows for changing baseline conditions and therefore determinations of benefit-risk profiles for different populations.

I have used these models extensively and for practical purposes. Although I am currently a consultant, specializing in benefit risk management and pharmacoepidemiology, I was head of epidemiology for three pharmaceutical/vaccine companies and in that capacity presented and taught the models, which were used repeatedly to evaluate different products' benefit-risk profiles. In addition, I have presented these models to FDA and in a variety of public meetings.

Report conclusion

The overall conclusion of the report authors, that the NNT/NNH approach has no utility for regulators, is not based on a scientific understanding or evaluation of the methods. The methods are rooted in decision analysis, which is evaluated by the report authors as having 'high' usefulness, which is contradictory and suspect. The authors own method, the MCDA, which consistently

receives the highest ratings, is not without its own critics. I myself have no direct experience with this approach, but I have read some criticisms, which I share below:

Because the inputs into an MCDA model are the subjective preferences of different decision makers, the outcomes are highly dependent on the ethics and goals of the participating members (Saaty T. Theory and Applications of the Analytic Network Process: Decision Making with Benefits, Opportunities, Costs, and Risks. 2005).

There are many MCDA models in use and this is problematic because different methods may yield different results for exactly the same data – and there is no apparent method available to compare the different MCDA methods – without relying on an MCDA approach (Triantaphyllou et al. An Examination of the Effectiveness of Multi-Dimensional Decision-Making Methods: A Decision-Making Paradox. International Journal of Decision Support Systems 1989;5:303-12).

A review by Kujawski (Multi-criteria decision analysis: limitations, pitfalls, and practical difficulties, 2003: http://escholarship.org/uc/item/0cp6j7sj#page-1) goes into more detail about the method.

I suppose that one of the points here is that it is easy to criticize [other people's] methods; the central theme, however, should be to find common ground and learn how to build and synthesize the best paradigms from which we can all benefit.

My conclusion

Again, it is not my intention to disparage the MCDA approach, which may indeed have a helpful role in benefit-risk analyses of drugs and medical interventions. It is my intention, on the other hand, to demonstrate that my methods, which were denigrated, do in fact have utility beyond the authors' description and recognition. It is also my purpose to point out that the report, if so dismissive of a misnomered NNT/NNH approach, may be similarly disserving to other methods.