EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

6 May 2011
EMA/61755/2011
Human Medicines Development and Evaluation

Benefit-risk methodology project

Reply to the comments received from Dr William Holden on the Work Package 2 report

Before we engage in the specific criticisms of our work raised by Holden, we wish to make clear that the purpose of the Work Package 2 (WP2) report was to review and comment on benefit-risk approaches that could be useful to regulators of medicinal products. The sections headed "Our View" maintain this focus; we limited our discussion to the potential usefulness for regulators concerned with balancing benefits and risks as part of the process of making decisions about medicinal products.

Holden's critique of the entry on NNT/NNH is preceded by the criticism that many of the criteria "are circular in the philosophical sense and not defined". In elaborating the criticism he first asks why an evaluation should be decomposed and suggests decomposition might even be considered antithetical to the spirit of benefit-risk analysis, and that synthesis of data is the point. We agree with the latter, but synthesis requires elements to by synthesised, and those elements are not necessarily in one-to-one correspondence with the data. We distinguish between real-world properties of medicinal products on the one hand, and the quantitative measures of those properties. One of the tasks of scientifically sound benefit-risk analysis is to question the correspondence between the real-world properties and the characteristics of the numbers that are used to represent those properties.

It is that distinction which leads us to the question of meaningfulness: are the statements we make on the basis of the numbers we have assigned also true of the real-world properties? We didn't feel that a tutorial on measurement theory (Krantz, Luce et al. 1971) was appropriate for the WP2 report, but we can assure Holden that we do not see meaningfulness as in the eye of the beholder. Rather, we consider the test of invariance, that the statements we make on the basis of the measures considered relevant to benefits and risks should not be different under a change in the unit or origin of the measure, as crucial to establishing meaningfulness. For example 100°C ÷ 50°C certainly results in a "2", but that doesn't mean that one represents twice the temperature of the other, as can be seen when those temperatures are converted to Fahrenheit: 212°F ÷ 122°F ≠ 2. Ratios of measured temperature on these scales do not correspond to ratios of actual temperature, whereas ratios of differences of measured temperature are true of actual ratios of temperature differences. Thus, Celsius and Fahrenheit scales are considered as interval scales: linear transformations of the measures are interpretable of temperature itself, whereas ratios of temperature measures are not.

The idea that a B/R approach should be theoretically sound is not original with us. Here we admit to being influenced by decision theory, which as we explained in section 3.2.3 of WP2 report, is based on

7 Westferry Circus • Canary Wharf • London E14 4HB • United Kingdom
Telephone +44 (0)20 7418 8400 **Facsimile** +44 (0)20 7418 8613
E-mail info@ema.europa.eu **Website** www.ema.europa.eu

the attractive concept of coherence, that a decision maker would not at a point in time knowingly choose actions that are contradictory. From this idea, first proposed by Ramsey (1926), others developed the idea to provide the theoretical foundation of decision theory, a set of axioms and proofs of the theorems implied by the axioms (von Neumann and Morgenstern 1947; Savage 1962). It is this base that led us to consider decision theory as theoretically sound. It would also be possible for a set of empirical studies to provide the motivation for a new theory, which when confirmed by further studies would then constitute a "theoretically sound" basis.

One of the axioms of decision theory is "independence of irrelevant alternatives", which means that the ordering of preferences between A and B should not change with the introduction of C, a third alternative. Thus, if a drug is overall preferred to a placebo, then introducing a comparator into the choice set should not then cause the placebo to be preferred to the drug. We are aware that human choice behaviour does not always obey this axiom, but in WP2 we were reviewing approaches as normative guides to behaviour, not descriptions of actual behaviour. Of course, it is normatively acceptable for C to be most preferred in the choice set [A, B, C], but the ordinal preference between A and B must remain after the introduction of C. These normative concepts are what we meant by the term "relative evaluation". We define "alternatives" as choices that are within the control of the decision maker. All the examples given by Holden, adding a new drug to metformin, interactions with other drugs, subgroups with different profiles, can be accommodated within this theoretical formulation.

We turn next to the section "Other evaluation criteria". We agree that few approaches are fully comprehensive; that is one of the reasons we introduced the qualitative framework, which is more comprehensive than any of the quantitative methods, though it leaves out how each step in the framework would be realised.

We believe that consistency checking in B/R assessment is an important concept. This is not done by the model, but rather by the modeller interacting with experts providing inputs to the model. For example, if ratio-scaling is used for either a benefit or risk criterion, then consistency checks can be made by comparing ratios of measures. If the scales are interval, then ratios of differences of values provide consistency checks on the measures. We know from our experience that these kinds of checks are made in Decision Analytic modelling, and it is surprising how frequently people find that input data or judgements fail these tests, and require reconsideration of the input data. On this point, our experience mirrors that of Holden who reports that in his work on epidemiology "there is little consistency in people's judgments". In Decision Analytic modelling we try to catch those inconsistencies as they occur, and invite experts to make revisions so that the inputs are consistent.

We agree that the practicality criterion consists of items that have to be judged subjectively. That is why we administer questionnaires before and after our modelling interventions to see in what ways, if any, the intervention was found to be helpful. As for the linear growth item, consider a decision tree, which can grow exponentially as more and more uncertain events are included.

Now we turn to NNT/NNH, each of which is the reciprocal of differences in probabilities. All Holden's criticisms seem to flow from one simple misunderstanding of our view that these two statistics do not take account of clinical relevance. He says they do; we say that clinical relevance is only considered in the mind of the physician, but not in those statistics. In decision theory, a probability represents an individual's degree of belief about the chance that an event or uncertain quantity will occur. As probabilities are unit-less numbers, a difference in probabilities is itself a probability, which represents a difference in degrees of belief about a favourable effect, for NNT, or an unfavourable effect, for NNH. Neither these differences nor their reciprocal, represents the desirability of a favourable effect or the severity of an unfavourable effect, both of which are matters of clinical relevance, usually expressed in decision theory as utilities or preference values. As such, probabilities do not explicitly include any

consideration of clinical relevance, which no doubt is taken into account by a prescribing physician faced with an individual patient.

Another issue is that different NNT values would typically be obtained for different favourable effects, and for different unfavourable effects, which would pose problems for a regulator. Could one meaningfully add those different numbers for the favourable effects and separately for the unfavourable effects? It is not clear what either of those sums would represent, let alone their ratio or difference, so we continue to maintain that these multiple statistics cannot be combined meaningfully across all those effects to give an overall benefit-risk balance. Even if a single NNT for a primary endpoint and one NNH for the most serious unfavourable effect were considered, what would their ratio mean? Furthermore, we do not see how to form preferences for these ratios because both numerator and denominator can vary.

Holden says that "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 positive benefit-risk profile." That is a worrying statement if researchers truly believe this. If a new pain killer completely eliminates pain for everyone, compared to a placebo giving no pain relief at all for anyone, then the NNT is 1. But if the drug has a 50% chance of killing the taker, then the NNH is 2. It seems clear that although NNT<NNH, this is not a positive benefit-risk profile.

We have argued that benefit-risk preferences cannot satisfactorily be formed from simply considering NNT and NNH because they do not include considerations of utilities or values. Holden's attempt to fix this by introducing the notion of relative value violates the scaling limitations of utilities. His equation for relative value, (1 – utility of AE) ÷ (1 – utility of disease of interest) is a ratio of utilities. But utilities are not ratio-scale numbers; they are interval scales, like Celsius and Fahrenheit temperature, so ratios of utilities are not interpretable. So, his 'fix' for relative value violates the criterion of coherence.

In summary, we still maintain that for the purposes of determining the quantitative benefit-risk balance of medicinal products from the perspective of drug regulators, the only appropriate way to combine utilities and probabilities is with the expected utility calculation: the utility (or preference value) of each effect is multiplied by the probability of realising the effect. Sums and differences of those expected utilities can be meaningfully interpreted. If the sum of the expected utilities of all the favourable effects exceeds the sum of the expected utilities of the unfavourable effects, then the benefit-risk balance can be considered favourable. No satisfactory quantitative representation of the balance of favourable and unfavourable effects can be accomplished just by considering only the reciprocal of differences of probabilities, nor can the addition of relative utility value adjustments. Only the expected utility model provides a coherent basis for helping regulators to form their preferences.

References

Krantz, D. H., R. D. Luce, et al. (1971). Foundations of Measurement Volume 1: Additive and Polynomial Representations. New York, Academic Press.

Ramsey, F. P. (1926). Truth and probability. The Foundations of Mathematics and other Logical Essays. R. B. Braithwaite. London, Kegan, Paul, Trench, Trubner & Co., New York: Harcourt, Brace and Company. 1999 electronic edition.

Savage, L. J. (1962). The Foundations of Statistical Inference. London, Methuen.

von Neumann, J. and O. Morgenstern (1947). Theory of Games and Economic Behavior. Princeton, NJ, Princeton University Press.