Health Economists Study Group

Nottingham

10-12 July 2000

Decision validity is the key to the generic vs. condition-specific issue in HRQOL measurement

Jack Dowie

London School of Hygiene and Tropical Medicine
email: jack.dowie@lshtm.ac.uk

1. HRQOL in health care decision making

Everything significant in this paper is grounded on acceptance of two prescriptive criteria for decision making in health care, where decision making is defined as choosing between alternative actions/options/strategies/allocations. The distinction between clinical and social decision making is not relevant to these criteria and so they apply to decisions at any and every level. The two criteria are:

(a) that one should treat/manage the person/patient (clinical) or persons/patients (social) and not the condition(s);

(b) that one should arrive at a choice from among available alternatives by a process that is requisitely transparent and explicitly coherent, as well as preference-based, evidence-based and, in a resource-constrained publicly-funded health care system, cost-effective.

As argued elsewhere, I find it difficult to accept that any process other than some form of decision analysis (clinical or cost-utility), reflecting the analysis-to-intuition ratio characteristic of 'mode 4' on Hammond's cognitive continuum, can meet the second criteria (1, 2)

Imagine that we have come to the point in such a decision making process where it has been agreed that HRQOL is relevant – and hence (by criteria b above) some HRQOL measure. The key question arises: what sort of measure or measures should be used? Broadly speaking should it be a generic measure (a GEN) intended to cover the entire domain of health, a condition-specific measure (a CSM) restricted to aspects of health associated with the condition concerned, or both? (There may be choices of measure to be made within each group, but these are conditional decisions and not of prime concern here.).

My impression from the literature is that the most popular of these three possibilities with clinicians, wearing either their clinical or policy making hats, is 'CSM only', the second most popular is 'both CSM and GEN' and the least popular 'GEN only'. In fact, the majority of clinicians might give the fourth logical possibility - 'neither' - but I regard this answer as having been excluded from consideration on the basis of the argument in the opening section.

In contrast, I will be arguing that it should be a GEN alone that is used most of the time. That a CSM alone will sometimes be appropriate. But that it will never be appropriate to use both a CSM and a GEN 'alongside each other', 'together', 'in combination', 'in conjunction', or 'in parallel' *for the same decision*.

The argument is based on the distinction between knowledge validity and decision validity elaborated in the following section of the paper. It is also based on a rejection of the widespread assertion or belief that CSMs are more 'sensitive' or 'responsive' than GENs and hence can detect 'small but important changes' that GENs always or often miss. It is this latter belief that underpins most clinicians' - and many health service researchers' - preference for CSMs and their rejection, or scepticism about, GENs. It is a belief that I will argue (in section 3) is usually unjustified and frequently methodologically fallacious, despite being based on otherwise high quality empirical research.

2. Knowledge validity and decision validity

Truth seeking and decision making (defined as choosing between alternative actions) are fundamentally different activities. The criteria appropriate for knowledge evaluation - the conventional 'scientific', 'psychometric' criteria of validity, reliability, etc. (3) , accompanied by an almost exclusive emphasis on avoiding type 1 errors - are of course desirable properties of the inputs into action evaluation, but they cannot be necessary, let alone sufficient ones if decisions are to be made. Decision validity, irrelevant in knowledge evaluation, is, on the other hand, always a necessary condition in action evaluation, and, given that decisions (clinical or social) must and will be taken with or without knowledge validated by such criteria, may often be a sufficient condition.

In knowledge validity we ask of a measure 'Does it measure what it is supposed to measure?' In decision validity we ask 'Does it measure what it is necessary to measure for the decision which is to be made using it?'.

What is necessary for a measure to have decision validity? Answer: it must embrace all the possible outcome states that can result from the choice of any of the alternative actions facing the decision maker. In health care it is therefore the outcomes from alternative treatments (or managements) for the condition that determine the relevant HRQOL measure, not the outcomes from the condition itself. Accordingly, the HRQOL measure that is appropriate for any decision is determined by the structure of that decision - by its action-event-outcome scenarios as mapped out in the decision tree, Markov model or influence diagram. Most CSMs focus, naturally enough given the interests of their compilers as specialist providers, on the set of states associated with the condition, states which will hopefully be changed by the treatment. Unfortunately, it is the set of outcomes states associated with all the management options for the condition that is the relevant set, not just those arising from the condition itself. These will often be very different sets.

In the vast majority of decisions in health care the range of possible outcomes - unless they are artificially truncated in time or restricted in scope - will imply the use of a GEN. The most obvious reason is that in most decisions there will be differences in life expectancy and/or possibility of significant morbidity among the multiple scenarios. Where the range of outcomes is genuinely limited - and the criteria for this will include the requirement that life expectancy is identical in all scenarios - a CSM is appropriate. And in such cases a GEN will be inappropriate. It all depends on the decision. One cannot therefore properly evaluate a HRQOL measure outside a specific decisional context and the implication that its knowledge validity is determinative of whether or not it is appropriate or ‘usable’ is simply unjustified.

It follows that it is never appropriate to use both a GEN and a CSM, 'alongside each other' or 'together / in conjunction / in combination' (or whatever), because the decision validity of a HRQOL instrument or measure is by definition decision-specific. It can only be established in the context of a specific decision and that decision will determine which measure is appropriate - either a GEN or a CSM. In order to decide which type of HRQOL measure is appropriate, i.e. is decisionally valid, one effectively needs to carry out the structuring stage of a decision analysis.

To illustrate. If one is deciding between TURP and contact laser prostatectomy, or between either and neither, a GEN is required because of the outcomes to which those options may lead. Any CSM - such as the AUA7 or the BPH impact - bothersome' - index - is inappropriate, even as a complement to the GEN. If, however, one is deciding between a

policy of having a drink late at night and not having one, such CSMs are (probably) appropriate - and any GEN (probably) inappropriate. (A little uncertainty exists because I haven't drawn that tree.)

Another way of putting the point would be to say that, if we are interested in health care practice and delivery, we should be evaluating knowledge-in-use not knowledge per se . That is what makes the decisional validity of a HRQOL measure, not its knowledge validity of prime concern. I see this mapping on to Alan Williams' favourite story about the inebriated person, searching for the keys under the light because he could see better there, when he had lost them elsewhere. By analogy, it must be better to use a decisionally valid measure which is defective from a knowledge point of view than one which passes all the knowledge tests but is decisionally invalid.

It may be objected that many CSMs are produced merely in order to 'assess' either a single patient, or a population of patients, in relation to a specific condition and its management. I have a simple question in response. Is this assessment going to be input into some decision, by someone, somewhere, sometime? If the answer is 'yes' - and I think it almost always will be - my argument continues. If the answer is 'no' then I would very much like to know why the assessment is being made and at whose expense.

3. Sensitivity , responsiveness and 'importance'

Given that, by these criteria and this argument, a very large proportion of health care decisions will require a generic, preference-based index measure, what if there are felt to be problems with any or all available GENs, particularly that they purportedly lack 'sensitivity' or 'responsiveness' to the effects that a CSM does/will/would show? I will come back shortly to the problematic aspects of such claims to greater 'sensitivity' (signaled by the 'felt to be' and 'purportedly'), but taking them as unproblematic for a moment, the answer - and the only answer - is to address these problems directly and attempt to produce a better, more sensitive and responsive, generic preference-based index measure. The answer is not to introduce an inappropriate (but purportedly complementary) CSM into the decision making process and suggest that decision makers somehow 'take it into account and bear it in mind', alongside the deficient GEN. This process will certainly be non-transparent, and it will almost certainly be incoherent. Why almost certainly incoherent? Because the impossibility of mapping CSMs of the conventional sort on to GENs has been well-established (4, 5). The fact that most CSMs employ ordinally scaled items and have no theoretical basis for item aggregation are merely two of the manifold reasons why this should cause no surprise. The fact that many clinical and social decision makers (with the encouragement of many health service researchers) think they can accomplish intuitively a task which is analytically impossible proves nothing more than they think they can accomplish intuitively a task which is analytically impossible.

The main contention of this section is that the assertions concerning the superior sensitivity and responsiveness of CSMs, and their ability to detect 'small but important' changes that GENs miss, are in fact highly problematic.

These assertions are now almost endemic in the literature. Among the most interesting papers which develop the claims are those by Guyatt et al. (6) and Jenkinson et al.(7). I now give detailed attention to what these papers have to say.

Guyatt et al.

After having noted the advantages of generic measures and advocated their continued use the authors go on to argue

> … unless investigators include responsive and valid disease-specific measures of health-related quality of life in controlled trials in chronic diseases, they risk misleading conclusions about the effects of treatments on health status. (p187)

The basis of this conclusion is restated several times in the paper, often as a generalisation and without particular reference to the chronic airflow limitation (CAL) context that is their specific focus:

> Generic measures may not … perform well in measuring crucial disease-specific aspects of HRQL and in particular may fail to detect important treatment effects if their magnitude is not large. Disease-specific measures provide an alternative to generic health status measures and may be more responsive to small but important changes in HRQL. Indeed many investigators are convinced that specific measures are required to ensure responsiveness. (p187)

> … generic measures do not focus on any area of HRQL in detail. As a result they may fail to detect small but important changes.…Indeed, generic measures have proved less responsive to treatment effects than specific instruments in a number of randomized trials, although in other instances, they have demonstrated comparable responsiveness. (p190)

> In the current analysis a disease-specific measure of HRQL focusing on dyspnea, fatigue emotional function, and mastery demonstrated small to moderate benefit of rehabilitation in patients with CAL. … The three generic instruments … failed to detect treatment effects. (p190-1)

> We view the global instruments as useful instruments for validation, but we would not suggest their use as outcome measures [because] general single item measures tend in general to be less responsive and valid than multi-item measures. (p191)

The empirical support for these conclusions comes (in this particular paper) from the results of applying both types of instrument in a controlled trial of two strategies for CAL: an 8 week inpatient respiratory rehabilitation program and conventional community care. As indicated, greater changes, as measured by effect sizes and responsiveness measures, were found in the CSMs (especially the Chronic Respiratory Questionnaire) than in the GENs (Sickness Impact Profile, Quality of Well-Being and Standard Gamble).

But this raises the question of what comparative sensitivity means and how it is to be established. Well, the answer is not by putting the changes in CSMs and GENs alongside each other and seeing which is bigger. If some action produces an x% effect in a CSM but does not show up on a GEN (or produces, say, only a .1x% effect) nothing at all follows from those two facts, other than that the former number is indeed bigger than the latter. There can be no more warrant for saying the GEN is less sensitive or less responsive than the CSM than for saying that the distances between London and Sheffield and London and Exeter are longer when measured on a 1:25,000 map, than when measured on a 1:250,000 one. If one wants to say that the former map is more sensitive to distance than the latter I suppose one might say

so, but I cannot think of any decision involving choosing between Sheffield and Exeter which would be better taken on the basis of this map rather than the other.

Cynically (see the first quote of the quartet above) one can envisage a CSM being designed to show 'large and important' effects in the same way as changes in the scale of a graph changes the impact of the same data. Consider a hypothetical Lower Limb Length scale, which runs from 0 (both legs completely amputated) to 100 (both legs left completely intact). The unit is 1 inch. It could easily be shown that the LLL scale is much more sensitive and responsive than EuroQol, in that there will often be a movement of several units on it while nothing happens according to EuroQol.

But there is no need to be cynical. All that is necessary is to point out that we cannot know whether any improvement on a CSM is 'important' or 'crucial' - or even 'large' or 'small' in a comparative sense - until we set it into context. And the relevant context will be a GEN! Important and crucial are comparative assessments. To assert that something is important or crucial on the basis of a non-generic scale is to imply that the patient defines their HRQOL solely in relation to a part or sub-domain of the generic scale. Indeed defines them in the precise way implied by the CSM concerned. But most patients are concerned with all the dimensions of HRQOL and the full range of health states, which is of course what makes a GEN generic. While there can be valid criticisms of any particular generic instrument or measure, they can, therefore, only concern its lack of generic validity, not its supposed failure to pick up condition-specific aspects of HRQOL - because there are none which are not part of HRQOL in general.

The actual empirical results of any CSM-GEN comparison are irrelevant. Even if the results were to suggest a GEN measure was as sensitive as a CSM, or even more sensitive, in a particular context or study (see quote two of the quartet), the fundamental objection to this sort of misinterpretation remains.

It is important to remind oneself of three things at this point.

Firstly, the fact that Guyatt et al. and many others making or endorsing this argument are in favour of a dual strategy - of CSM and GEN being used 'in combination' - is beside the point. There is no merit in a cosy compromise position which threatens to undermine any merit each measure possesses individually.

Secondly, we are not now addressing the question of whether it is a GEN or a CSM that is decisionally valid. That will depend (in the Guyatt study for example) on the entire range of outcomes contained in the scenarios flowing from the two rehabilitation options, not just the outcomes in relation to CAL. We are talking here about the claim that a CSM captures something important that a GEN misses.

Thirdly, the impossibility of mapping CSMs on to GENs is based on multiple incompatibilities, not least among which is the fact that generic measures yielding utilities, such as EuroQol, possess only interval properties. Many analyses that make assertions regarding the comparative (in)sensitivity of such GENs, including the Guyatt et al. paper, implicitly treat them as possessing ratio properties, and are hence formally invalid. On the other hand, CSMs typically involve treating the result of aggregating ordinal scales as possessing ratio properties, so the whole comparative exercise is a methodological nightmare, statistically speaking.

Jenkinson et al.

Jenkinson et al. compared two generic measures of health status (SF36 and EuroQol 5D) with two disease-specific measures (AUA 7 symptom and the Bothersome Index) in a trial of transurethral resection of the prostate (TURP) versus laser vaporization prostatectomy for benign prostatic hypertrophy (BPH). They conclude that 'the disease-specific measures are more sensitive to change than the generic measures of outcome'. (p1109)

Many of the comments already made in relation to the Guyatt paper apply here also, so I will move straight to the authors' extended discussion of the possible explanations of why the two CS instruments suggest 'substantial improvement' (p1115) during the period of the study, whereas the EQ-5D showed no change (and the SF36 very little).

> First, it may have been that the … EQ-5D simply [was] failing to detect clinically significant changes in health that were being detected by the AUA-7 and Bothersome instruments. This may have been because of the fact that patients perceive assessments of overall health as independent from condition-specific assessments, which for the most part refer to symptoms of the condition. This interpretation would be in line with the widely held view that generic measures are generally less relevant and less responsive measures of patient outcome than are condition-specific measures. (p1115)

There are what should be by now familiar problems with this line of reasoning. A GEN is not trying to pick up 'clinically significant' changes (whatever they are) but to pick up changes in the HRQOL of the patient. The final sentence follows from the preceding ones only if patients see overall health as less 'relevant' than condition-specific symptoms (which I very much doubt), as well as distinguishing the generic from the condition-specific (which they may very well do).

> Second, it is possible that the condition-specific instruments were focused too narrowly on specific symptoms to capture important broader outcome measures, such as social functioning or depression, which are being captured by the generic instruments. From this perspective, the changes detected by the condition-specific instruments could be seen as occurring within a small subdomain of the range of outcomes measured by the generic instruments and are not significant within this wider context. … The condition-specific measures, however, largely were concerned with symptoms that, for the most part, were what patients present with and that they wish to have cured. Consequently, although the removal of these symptoms may not have led to dramatic improvements in overall health, their health state could be seen to have improved. (p1115-1116)

Here we can express almost complete agreement, the almost reflecting the fact that the final conclusion about 'health state' strictly requires the mapping of CSM onto GEN.

> Third, it is possible that … the EQ-5D failed to show any changes in health outcomes because the changes measured by the condition-specific instruments were too small to be valued: that is, they had an insignificant impact on overall utility in the valuation procedure underlying the EQ-5D. This, in turn, could occur either because of some crudeness of the EQ-5D as a descriptive instrument, because of deficiencies in the valuation procedure, or simply because the change in outcome was intrinsically too minor to affect overall utility.

Pursuing the first of these possibilities Jenkinson et al. go on to emphasise the lack of sensitivity built into the descriptive system of EuroQol as a result of the use of only three levels within each dimension. They note that its developers

> … have explicitly acknowledged a trade-off between the instrument's detail and its ease of use. A wider range of possible responses rapidly would increase the existing 243 possible health states generated by the instrument, which, in turn, would make it extremely difficult to attach a set of valuations. By forcing respondents into a fairly limited range of descriptive responses, however, one consequence appears to have been that the smallest possible change in a health state described by the EQ-5D - that is, a change of one level in one dimension - equated with a large change in the valuations.

Agreed, but nothing here, or in the arithmetic example which follows (pointing out that the average one-level, one-dimension deterioration from EuroQol state 11111 is a sizeable 16%) leads, as is implied, to the conclusion that we need sensitive CSMs to be used 'alongside' insufficiently sensitive GENs. (Incidentally while 16% on tariff A1, the equivalent figure is only 8% on tariff A6). In the long term the efforts going into CSMs need to be redirected towards making generic measures more sensitive. But if Jenkinson et al. try to do this they will run into exactly the same practical problems as current GEN developers did. It is only because most CSMs are not developed on a comprehensive preference and trade-off basis - symbolic of the fact that they are mostly generated by health care providers and health knowledge producers rather than by decision owners - that they are flourishing. All the fundamental difficulties are thereby exported to an opaque 'taking into account and bearing in mind' phase of decision making.

Jenkinson et al. conclude their paper with the familiar call for compromise, indeed reconciliation

> … although the preceding discussion has focused on the comparative performance of condition-specific and generic outcome measures, it should be emphasised that these are not in competition; the arguments for using a generic core alongside condition-specific outcomes measures have been well-rehearsed, as have the arguments for using utility measures. The resulting explicit comparisons of results and analyses of differences perhaps would encourage some reconciliation of methodologic approaches as well as empirical findings

Regrettably while the arguments may have been well-rehearsed, the rehearsing seems to amount to little more than repetition of a mantra. If we return to the main study cited at this point we find John Ware writing

> There has been considerable debate over the choice between disease specific and generic measures of outcomes of clinical trials and studies of cost effectiveness, and numerous research teams funded by he American government are administering both. In North America the SF 36 is currently being used in more than 200 clinical trials, where investigators are interested in the impact of treatment on the quality of life as well as on more traditional medical outcomes. Measures of generic and specific outcomes usually prove to be more useful than either alone. {Ware, 1993 #4670, 1430}

Ware indeed confirms that a lot of people are adopting the dual strategy, but this provides no help in the quest for the reason why they are doing so. The suggestion that both types

together are 'more useful' than either alone hardly constitutes an argument, rather an indication that one is believed to exist. The desire to complement 'traditional medical outcomes' with 'quality of life impact' is another pointer to a reason, but assumes away the problem of how the two sorts of outcome measures are to be integrated. And of course it begs the question of why 'traditional medical outcomes' that are not concerned with QOL are the traditional medical ones!

4. Conclusions

The increasing body of check-lists for HRQOL instruments (and other analytical techniques) are to be welcomed, but only so long as they are accompanied by a 'PONCE' disclaimer, i.e. make clear that it does not follow from a low score achieved in relation to knowledge evaluation, even a score below some 'acceptable' threshold, that the instrument or technique is not better than the actual alternative that will replace it in decision making (action evaluation) if it is not used. If you are not familiar with my argument on Partial Or Non Comparative Evaluation - PONCE (9)- here is Alan Williams' version of it in relation to QALYs :

> A typical stance is to point out all the difficulties involved with some particular approach and then to sit on the fence waiting for the next candidate to come by and then do the same. This would be fine if the implied ideal method were available to us, or if we could suspend all health-care decision making until it were. But there is no perfect system on offer and we cannot wait. As with a well-conducted clinical trial, the new has to be compared systematically, according to preselected criteria, with what already exists. If the same criteria as are used to criticise the QALY approach were used in an even-handed way to criticise current practice, or any feasible alternative to it, how would these alternatives make out? It is irresponsible to do less. (10, p315-6)

Adopting a decision analytic/cost-utility perspective enables us to see why the current stress on knowledge validity, and associated enthusiasm for condition-specific measures of HRQOL, is dangerous for our health. The undue stress on them leads to a lack of balancing concern with decision validity and generic measures and accordingly has the potential to hinder progress both in the quality of individual patient care and towards the equitable and efficient allocation of resources among patients in publicly-funded health care systems.

Decision analysis from the patient's or patients' perspective - presumably these are the relevant ones - has no place for distinctions between 'primary' endpoints and 'secondary' endpoints or between 'main' effects and 'side' effects. These are essentially specialist clinical-provider and knowledge-researcher artifacts. For the consumer-patient there are only endpoints and effects, in relation to which chances must be assessed and preferences established. A decision tree makes this graphically very clear. And it makes clear why we should work backwards from an analysis of the structure of the decisions we must make to determine the knowledge we need to collect, not the other way round. (11). This point applies throughout the spectrum from setting up population-based trials to setting up individual patient monitoring. The latter should not escape the need to justify the use of condition-specific measures -even though they are more likely to be decisionally valid in the monitoring situation this validity needs to be established, not assumed.

John Ware concluded his 1995 survey by suggesting that

> two dilemmas must be resolved: (a) minimum standards of content for generic measures, and (b) how to handle measures reflecting specific medical conditions or treatments. (12, p339)

I suggest that the dilemmas are actually dissolved if we always take a decisional perspective, rather than a knowledge perspective, on HRQOL. When we know what our options are, we can determine what we (ideally) need to know about HRQOL in order to establish the best option . Which, we should remember, is all we need to do.

References

1. Dowie J. 'Evidence based', 'cost-effective' and 'preference-driven' medicine: decision analysis based medical decision making is the pre-requisite. Journal of Health Services Research and Policy 1996;1(2):104-113.
2. Dowie J. What decision analysis can offer the clinical decision maker. Hormone Research 1999;51:supp 1:73-82.
3. Brazier J, Deverill M. A checklist for judging preference-based measures of Health-Related Quality of Life: learning from psychometrics. Health Economics 1999;8:41-51.
4. Cairns JA, Johnston KM, McKenzie L. Developing QALYs from condition-specific outcome measures. HERU Discussion Paper 1991;14/91.
5. Chancellor J, Coyle D, Drummond M. Constructing health state preference values from descriptive quality of life outcomes: mission impossible? Quality of Life Research 1997;6:159-168.
6. Guyatt GH, King DR, Feeny DH, Stubbing D, Goldstein RS. Generic and specific measurement of Health-Related Quality of Life in a clinical trial of respiratory rehabilitation. Journal of Clinical Epidemiology 1999;52(1):187-192.
7. Jenkinson C, Gray A, Doll H, Lawrence K, Keoghane S, Layte R. Evaluation of index and profile measures of health states in a randomized controlled trial: comparison of the Medical Outcomes Study 36-item Short Form health survey, EuroQol, and disease specific measures. Medical Care 1997;35:1109-1118.
8. Ware JE. Measuring patients' views: the optimum outcome measure. SF36: a valid, reliable assessment of health from the patient's point of view. BMJ 1993;306:1429-1430.
9. Dowie J. The danger of partial evaluation. Health Care Analysis 1995;3:232-234.
10. Williams A. QALYs and ethics: a health economists' perspective. In: Culyer A, Maynard A, editors. Being reasonable about the economics of health: selected essays by Alan Williams. Cheltenham: Edward Elgar; 1997. p. 305-321.
11. Claxton K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. Journal of Health Economics 1998;18(3):341-364.
12. Ware JE. The status of health assessment 1994. Annual Review of Public Health 1995;16:327-354.