Related papers [A-G]

A number of peer-reviewed papers besides the five included were published in the course of my research. In all cases it was felt my contribution warranted first author status. The seven related papers [A-G] all drew on the same MCDA/Annalisa-based approach and software as those submitted in this thesis. They are briefly mentioned here to give an indication of the wider scope of the research undertaken during my thesis period.

[A] **Without a reconceptualisation of 'evidence base' evidence-based person-centred healthcare is an oxymoron** was the third in the sequence of reconceptualisations prompted by the empirical experiences. It argues that the evidence base in person-centered care should be the unsynthesised matrix of performance rates on the person's important criteria, not the pre-synthesised option evaluations (using group average preferences) that constitute the conventional evidence base. The synthesis of performance rates with the person's importance weights should only occur at or near the point of decision.

[B] **Enhancing informatics competency under uncertainty at the point of decision: a knowing about knowing vision** writes up the probability elicitation and evaluation instrument 'PROBER' for the first time since the software was revised. It makes the case for healthcare providers, who routinely make clinical probability judgements, gaining insight into 'how much they know about how much they know' via visual feedback, and being able to distinguish between their calibration and discrimination competencies. This reflected an interest stimulated by the probability elicitation exercises undertaken with expert clinicians in the context of developing the IBD decision aid.

[C] **Can a Discrete Choice Experiment contribute to person-centred care?** was produced out of a concern that the need for elicitation of the individual person's preferences at or near the point of decision - and particularly the development of tools to support this elicitation - was threatened by claims in numerous studies using the Discrete Choice Experiment (DCE) approach that establishing group-level average preference results could somehow facilitate clinical decision making.(DCEs are perhaps the main technique used in health economic evaluation studies of public preferences.)

[D] **Addressing preference heterogeneity in public policy by combining Cluster Analysis and Multi-Criteria Decision Analysis: Proof of Method** took up the challenge of how individual level importance weights, such as those emerging from widespread individual use of MCDA-based decision aids, could contribute to population-level policy making by way of clustering of the preferences of individuals, including 'persons-as-researchers'.

[E] **Bringing feedback in from the outback via a generic and preference-sensitive instrument for course quality assessment** was the result of dissatisfaction with standard forms of student feedback in teaching and the realisation that a dually-personalised course assessment instrument could be developed, providing a Student Reported Outcome Measure (STROM) equivalent to MyDecisionQuality as a Patient-Reported Outcome Measure (PROM), for formative but, possibly, also for summative use.

[F] **Health informatics can avoid committing symbolic violence by recognizing and supporting generic decision-making competencies** argues that failing to recognise and exploit a widespread form of functional decision literacy, leads to the symbolic violence experienced within healthcare consultations by individuals at any and all levels of general literacy. Many highly literate persons resort to the same range of avoidant and other undesirable strategies observed in those of low basic literacy. The alternative response we propose exploits the generic decision literacy which comes in the form of the ability to access the decision-relevant resources provided by comparison websites and magazines. Our MCDA-based approach extends this approach to healthcare options and permits the incorporation of personal criterion weights in furtherance of person-centred care.

[G] **Enhancing both provider feedback and personal health literacy: dual use of a decision quality measure** sets out a protocol for a study to establish the feasibility of using a web-based survey to simultaneously supply healthcare organisations with feedback on a key aspect of the care experience they provide and increase the generic health decision literacy of the individuals responding. The focus is on the person's involvement in decision making, an aspect of care which is seriously under-represented in current surveys from the perspective of person-centred care. By engaging with an instrument to assess decision quality the person can, in the one action, provide a retrospective evaluation of a past decision making experience in a specific provider context and enhance their competency in future decision making in any setting.

[A] Kaltoft, M.K., Nielsen, J.B., Eiring, Ø., Salkeld, G & Dowie, J., 2015. Without a reconceptualisation of "evidence base" evidence-based person-centred healthcare is an oxymoron. *European Journal for Person Centered Healthcare*. (Forthcoming)

[B] Kaltoft, M.K., Nielsen, J.B., Salkeld, G. & Dowie, J. 2014. Enhancing informatics competency under uncertainty at the point of decision: a knowing about knowing vision. In C. Lovis, ed. *e-Health - For Continuity of Care*, 192:879-883.

[C] Kaltoft, M.K., Nielsen, J.B., Salkeld, G. & Dowie, J., 2015. Can a Discrete Choice Experiment contribute to person-centred healthcare? *European Journal for Person Centered Healthcare*. (Forthcoming)

[D] Kaltoft, M.K., Turner, R., Cunich, M, Salkeld, G, Nielsen, J.B & Dowie, J, 2015. Addressing preference heterogeneity in public health policy by combining Cluster Analysis and Multi-Criteria Decision Analysis: Proof of method. *Health Economics Review*, 5:10.

[E] Kaltoft, M.K., Nielsen, J.B., Salkeld, G., Lander, J. & Dowie, J, 2015. Bringing Feedback in From the Outback via a Generic and Preference-Sensitive Instrument for Course Quality Assessment. *JMIR Research Protocols*, 4(1), e15.

[F] Kaltoft, M.K., Nielsen, J.B., Salkeld, G. & Dowie, J. 2015. Health informatics can avoid committing symbolic violence by recognizing and supporting generic decision-making competencies. *Studies in Health Technology and Informatics*, 218: 172-177.

[G] Kaltoft, M.K., Nielsen, J.B., Salkeld, G. & Dowie, J. 2015. Enhancing both provider feedback and personal health literacy: dual use of a decision quality measure. *Studies in Health Technology and Informatics*, 218: 74-79.

e-Health – For Continuity of Care
C. Lovis et al. (Eds.)

doi:10.3233/978-1-61499-432-9-975

Enhancing informatics competency under uncertainty at the point of decision: a knowing about knowing vision

Mette Kjer KALTOFT[a,1] Jesper Bo NIELSEN [a] Glenn SALKELD [b] and Jack DOWIE [c]
[a] *University of Southern Denmark*
[b] *University of Sydney School of Public Health*
[c] *London School of Hygiene & Tropical Medicine*

Abstract. Most informatics activity is aimed at reducing unnecessary errors, mistakes and misjudgements at the point of decision, insofar as these arise from inappropriate accessing and processing of data and information. Healthcare professionals use the results of scientific research, when available, and 'big data', when rigorously analysed, as inputs into the probability judgements that need to be made in decision making under uncertainty. But these judgements are needed irrespective of the state of 'the evidence' and personalised evidence on person/patient-important criteria is very often poor or lacking. This final stage in 'translation to the bedside' has received relatively little attention in the medical, nursing, or health informatics literature, until the recent appearance of 'cognitive informatics'. Positive experience and feed-back from several thousand students who have experienced exercises in assigning probabilities informs our future vision in which better decisions result from healthcare professionals – indeed all of us – having accepted that probability assignment is a skill, with the internal coherence and external correspondence of the probabilities assigned as twin evaluative criteria. As a route to improved correspondence – in the absence of the systematic recording and monitoring of real world judgments that would be the normal pathway to quality improvement - a 'Prober' is a set of statements to which the respondent supplies their personal probabilities that a statement is true. They receive the proper Brier score and its decomposition as analytical feedback, along with graphic representations of their discrimination and calibration, the two key components of good correspondence. Provided with estimates of their sensitivity (mean probability true for true statements) and specificity (1 minus mean probability true for false statements) they can visualise themselves as a 'test' when making diagnostic and prognostic judgements , thereby being given the cognitive foundation for such reflection in their clinical practice, including 'reflection in action'. They acknowledge that an appropriate balance of intuition and analysis is required, as in Hammond's Cognitive Continuum, and are made aware of the cognitive and motivated biases that can prevent us knowing 'how much we know about how much we know', with its deleterious effect on decision quality. Probability exercises, such as 'Probers', are proposed as an enhancement of professional courses and virtual learning environments, such as the TIGER initiative in nursing, through which the competency portfolio of all those seeking to deliver high quality person/patient-centred care can be expanded.

[1] Corresponding Author: Mette Kjer Kaltoft, Health Visitor RN MPH, Research Unit for General Practice, Institute of Public Health, University of Southern Denmark, Odense 5000, Denmark E-mail: mkaltoft@health.sdu.dk

Keywords. Probability, judgement, coherence, correspondence, calibration, discrimination

Introduction

Our vision is of the better decisions that will characterise the coming era of person/patient-centred care as a result of healthcare professionals – indeed all of us - accepting that, in decision making, we are necessarily Bayesians.[1] We accept that the assessments of the future chances which permeate decisions are ontologically personal and subjective, whatever the extent to which they are epistemologically-based on robust frequencies and however widely they are inter-subjectively agreed. All parties have rejected the temptations of right-wrong thinking, reflected in testing by non-probabilistic Multiple Choice Questions, along with the unwarranted confidence, trust and denial it often generates. Healthcare professionals treat the results of scientific research, when available, and 'big data', when rigorously analysed, as relevant inputs into the probability judgements that need to be made irrespective of the state of 'the evidence'. It is accepted that competence in making probability judgements is the key to improved handling of uncertainty at the point of decision so it is part of the training and education of clinicians.

Most informatics activity is ultimately aimed at reducing unnecessary errors, mistakes and misjudgements at the point of decision, insofar as these arise from inappropriate accessing and processing of data and information. For some criteria and some conditions high-quality 'evidence- based' probabilities can be acquired directly or through a nomogram or 'risk calculator' (preferably a *probability* calculator). [2] But in many cases the clinician will need to use their personal belief probability judgements to remedy the absence of, or to better personalise, the available estimates.

This frequently necessary final stage in 'bench to bedside translation' has received relatively little attention in the medical, nursing or health informatics literature. The widespread assumption has been that this is an intuitive competence that can, and can only, be acquired intuitively, through experience. However, this ignores a significant literature on how the quality of probability judgements can be assessed, on the empirical evidence on clinician performance in this respect, [3,4] on the possible sources of limited performance, and on possible routes to improved quality. Since it will take time to overcome the institutional-professional barriers to systematic judgemental recording and monitoring in practice – the normal route to competence improvement - our vision is pessimistic in this respect. However, as part of the increasing interest in 'cognitive informatics', clinicians can be provided with the cognitive basis for reflecting continuously on their judgemental practice and performance, both 'in action' and outside it, [5,6] accepting that an appropriate balance of intuition and analysis is required, as in Hammond's Cognitive Continuum, [7–9] as well as an awareness of the likely cognitive as well as motivated biases that may hinder them knowing 'how much they know about how much they know'. [10] Probability exercises (such as 'Probers') are therefore an integral part of our vision, enhancing professional courses and virtual learning environments, such as The TIGER Initiative in nursing.[11]

In relation to the evaluation of probability assessments - and assessor - Kenneth Hammond and others have emphasised that two distinct criteria are relevant, and drawn attention to the fact that, for a variety of reasons, including different meta-theoretic

paradigms, the two have attracted different sets of adherents. [12] There are those who wish to judge probabilities primarily by their *internal coherence* and those who wish to judge them primarily by their *external correspondence*. The vast majority of those who emphasise coherence are pessimistic about judgemental competence, because clinicians typically perform poorly on coherence tests, such as calculating the predictive value of a test result, given the sensitivity of the test and the prevalence of the target condition. Most optimists emphasise external correspondence, arguing that abstract tests of coherence are not 'ecologically valid', [13] since the items are not representative of those that actually arise. But there are also pessimists among those who favour the correspondence criterion, doubting whether experience will be productive in the absence of quick and unbiased feedback. [14,15] The 'clinical versus actuarial' controversy, associated primarily with the name of Paul Meehl, [16] rumbles on.

1. Methods

In the case of the coherence criterion, teaching of the way in which probabilities should be combined is required. Correspondence can only be taught through probabilistic exercises with relevant feedback. A Prober is a set of statements to which the respondent supplies their personal probability that a statement such as 'The true positive rate indicates the sensitivity of a test' is true. The set used currently consists of 50 statements relating mainly to research methods. A variety of probability response sets are available for use in the software. A compromise between response granularity and item set size is necessary to achieve a reasonable number of observations for an individual at each probability. We currently use seven discrete probabilities: 0, 10, 30, 50, 70, 90 and 100%. Respondents are advised that they should enter their honest probabilities and in order to avoid 'motivated biasing', they will receive full marks for completion of the exercise. In any case, the accompanying teaching makes clear that the assessments are scored by a proper scoring rule (Brier's) which ensures that respondent's expected score will always be maximised by reporting honest beliefs [17].

After completion the respondent can learn whether each statement was actually true or false, along with short elaborations, mainly in the case of false items. The main, analytical feedback comes in the form of the Brier score and its decomposition, [18] (Figure 1a) One key measure is that of *discrimination*, the difference between the average probabilities assigned to true and false items, plotted on the right and left axes respectively. (These represent the sensitivity and 1 minus the specificity of the judge interpreted as a 'test'.) Graphically discrimination is represented by the slope of the line joining them. This can be compared with the 45 degree slope of the diagonal which indicates perfect discrimination. An associated diagram (Figure 1b) provides information relevant to the other key competence, *calibration*. Calibration is measured by the degree to which the 'frequency correct' matches 'probability assigned'. For example, if a respondent assigned 70% to 10 statements, then perfect calibration exists if 7 of these are actually true. Deviations from 7 in *either* direction indicate poorer calibration. Accompanying teaching stresses that calibration should not be improved at the expense of using whatever discrimination ability is possessed

2. Results

The latest in 35 years of Probers use has been in the Translational Health Masters course at the Sydney School of Public Health. In 2012 and 2013, 63 students responded. (Completion rates were high as the exercises were a compulsory assignment). Their Brier scores ranged from .1 to .55 (where 0 is perfect and 1 is worst possible.) The mean score of .25 (SD .08) is actually that which would be achieved by assigning .5 probability to all 50 statements, so that on average the population did no better than chance. The average sensitivity (mean probability true assigned to true statements) was 75% and average specificity (1 minus probability true assigned to false statements was 64%. Only one of the 63 had a specificity exceeding sensitivity and hence a discrimination line with a negative slope.

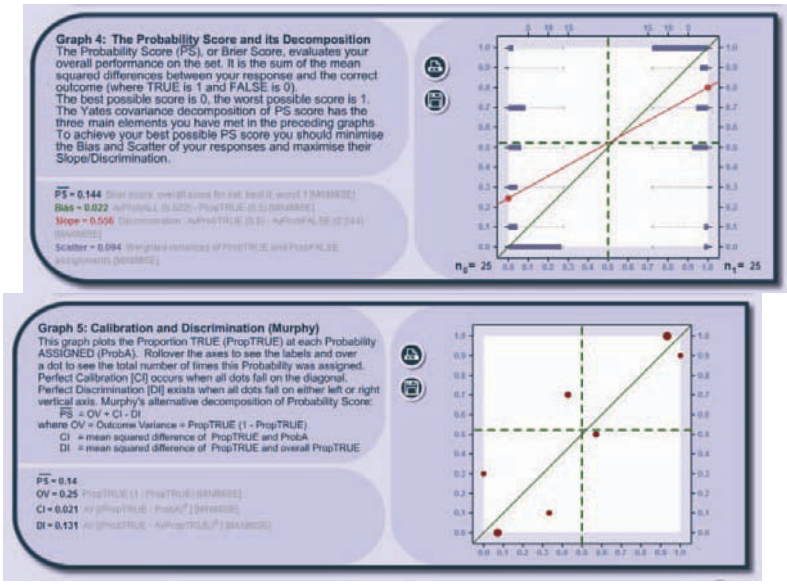


Figure 1 Student showing good discrimination (top) and calibration botton

As in previous settings there was no indication of respondent difficulty in completing the task at a practical level. Feedback comments have been solely about the unfamiliar nature of the task, without questioning of its relevance, and mainly doubts about whether using numerical probabilities and regarding it as a skill would be acceptable 'where I work' because it would be disruptive of organisational routines and/or professional hierarchies

3. Discussion

Having arrived at a numerical estimate of, say, 30%, the Prober-aware health professional will recall that if all their 30%s were monitored and collated the frequency correct should be 30%. They will be able to reflect 'in action' on their *calibration.* In relation to their whole set of judgements and outside any specific case, they can ask themselves whether they assigned a (much) higher average probability to the occasions

when the target outcome occurred, than the average assigned when it did not occur. They will be able to reflect, outside of action, on their sensitivity and specificity and overall *discrimination* competence.

Where is the 'evaluation' of Probers? Real world evaluation requires the systematic recording and monitoring of judgements that seems almost impossible in larger clinical settings. In our vision the 'anatomy of judgment' is taught alongside the anatomy of the human body in clinical curricula. Probers are part of the new cognitive informatics.

References

[1] J. Dowie The Bayesian approach to decision making, in A. Killoran, C. Swann, M. Kelly, Editors. *Public Health Evidence: Tackling Health Inequalities*, Oxford, Oxford University Press; 2006. p. 309–21

[2] J Dowie, Against risk, *Risk Decision and Policy* 4 (1999) 57–73

[3] J.G. Dolan, D.R. Bordley, A.I. Mushlin, An evaluation of clinicians' subjective prior probability estimates. *Medical Decision Making* 6 (1986) 216–23

[4] T.G.Tape, J. Kripal, R.S.Wigton, Comparing methods of learning clinical prediction from case simulations. *Medical Decision Making,* 12 (1992) 213–221

[5] D Schön *The Reflective Practitioner. How professionals think in action*, London, Temple Smith, 1983

[6] P Benner *From Novice to Expert: Excellence and Power in Clinical Nursing Practice*, London, Addison-Wesley, 1984

[7] K.R.Hammond *Human Judgment and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice,* New York, Oxford University Press; 1996

[8] J. Dowie, M.K.Kaltoft Deciding how to decide – and how to support decisions. *Nuffield Trust Webinar*. http://www.slideshare.net/NuffieldTrust/jack-dowie-211111. 2011

[9] J. Dowie, JUDEMAKIA : a personal map of the world of judgement and decision making. https://www.dropbox.com/s/ph800ycdah5no92/Judemakia pdf.pdf.2006

[10] A. Tversky, D Kahneman, Judgment under uncertainty: heuristics and biases, *Science* 185 (1974) 1124–31

[11] M.K.Kaltoft Nursing Informatics AND Nursing Ethics : Addressing their disconnect through an enhanced TIGER-vision, *Studies in Health Technology and Informatics*, *192* (2013), 879–883

[12] K.R. Hammond How convergence of research paradigms can improve research on diagnostic judgment. *Medical Decision Making* 1996 16 (1996) 281–287

[13] Hammond KR. Ecological Validity: Then and Now.1998 http://www.brunswik.org/notes/essay2.html

[14] B Brehmer, In one word: not from experience. *Acta Psychologica* 45 (1980) 223–41

[15] I. Fischer, D.V. Budescu, When do those who know more also know more about how much they know? The development of confidence and performance in categorical decision tasks. *Organisational. Behavior and Human Processes* 98 (2005) 39–53

[16] P.E. Meehl, Causes and effects of my disturbing little book. *Journal of Personality Assessment*, 50 (1986) 370–5

[17] G.W. Fischer, Scoring-rule feedback and the overconfidence syndrome in subjective probability forecasting *Organisational. Behavior and Human Performance* 29 (1982) 352–69

[18] J.F.Yates, External Correspondence: Decompositions of the Mean Probability Score, *Organisational. Behavior and Human Performance* 30 (1982) 132–56.

[19] A.S. Elstein, Beyond Multiple-choice Questions and essays: the need for a new way to assess clinical competence, *Academic Medicine*. 68 (1993) 244–9

Access To obtain access to the Prober set, email jack.dowie@sydney.edu.au **Conflicts of Interest** Prober is ©Maldaba Ltd; Contact: info@maldaba.co.uk . Jack Dowie has a financial interest in Prober, but has not benefited from its use in any of the specified settings. **Authorship** JD and MKK jointly planned the paper's structure and content. JD's first draft was extensively revised with MKK. GS and JBN also provided comments. All authors approved the final version.

Kaltoft et al. Health Economics Review (2015) 5:10
DOI 10.1186/s13561-015-0048-4

Health Economics Review
a SpringerOpen Journal

RESEARCH Open Access

Addressing preference heterogeneity in public health policy by combining Cluster Analysis and Multi-Criteria Decision Analysis: Proof of Method

Mette Kjer Kaltoft[1], Robin Turner[2], Michelle Cunich[3], Glenn Salkeld[4], Jesper Bo Nielsen[1] and Jack Dowie[5*]

Abstract

The use of subgroups based on biological-clinical and socio-demographic variables to deal with population heterogeneity is well-established in public policy. The use of subgroups based on preferences is rare, except when religion based, and controversial. *If* it were decided to treat subgroup preferences as valid determinants of public policy, a transparent analytical procedure is needed. In this proof of method study we show how public preferences *could* be incorporated into policy decisions in a way that respects both the multi-criterial nature of those decisions, and the heterogeneity of the population in relation to the importance assigned to relevant criteria. It involves combining Cluster Analysis (CA), to generate the subgroup sets of preferences, with Multi-Criteria Decision Analysis (MCDA), to provide the policy framework into which the clustered preferences are entered. We employ three techniques of CA to demonstrate that not only do different techniques produce different clusters, but that choosing among techniques (as well as developing the MCDA structure) is an important task to be undertaken in implementing the approach outlined in any specific policy context. Data for the illustrative, not substantive, application are from a Randomized Controlled Trial of online decision aids for Australian men aged 40-69 years considering Prostate-specific Antigen testing for prostate cancer.

We show that such analyses can provide policy-makers with insights into the criterion-specific needs of different subgroups. Implementing CA and MCDA in combination to assist in the development of policies on important health and community issues such as drug coverage, reimbursement, and screening programs, poses major challenges -conceptual, methodological, ethical-political, and practical - but most are exposed by the techniques, not created by them.

Keywords: Cluster analysis; Multi-criteria decision analysis; Preference subgroups; Heterogeneity

Background

Most health care systems are currently under pressure to reconcile the need to deliver services more efficiently and provide more personalised health care. There are a number of reasons for this pressure, including rapid technological advances in medicine and communications, aging populations, and economic crises. A key issue is how population heterogeneity should be respected in policy decisions about health and community issues such as drug coverage, reimbursement and screening. If fully individualised public health care policies are impossible and treating everyone as 'average' is unsatisfactory, then what subgroupings represent the optimal compromise, and how are they to be incorporated into public policy?

The case for using subgroups based on *biological-clinical and socio-demographic variables* to address heterogeneity is well-established in *effectiveness* research, with the main issues being the statistical and clinical/policy significance of such analyses. Subgrouping in *cost-effectiveness* is the focus of ongoing debate, largely concerning the use of particular variables for subgrouping rather than the case for subgrouping in principle. Subgrouping based on age and clinical history is widely employed in analyses for organisations determining cost-effectiveness within specific settings, such as NICE in England and Wales [1]. What remains controversial is the use of subgrouping on the basis of individual *preferences or*

* Correspondence: jack.dowie@lshtm.ac.uk
[5]Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK
Full list of author information is available at the end of the article

values, moving beyond clustering based on such concepts as patient satisfaction [2] or healthcare decision making competencies and motivations [3].

The controversy is subdued in the case of most effectiveness research, where it is accepted that key determinants of effectiveness, especially treatment adherence, may be influenced by individual preferences independent of the person's biological-clinical or socio-demographic characteristics [4]. Little concern has been shown when the suggestion is made that clustered results from individual decision analyses might be useful inputs into group/policy decision making in some indirect and unspecified way [5,6]. The question remains as to whether the preferences of individual citizens, via preference-based subgroups, should have a formal, direct role in cost-effectiveness analysis and policy formation. This is particularly important in relation to resource-consuming decisions in collectively-funded public health services.

The case for acknowledging patient heterogeneity in preferences has been convincingly made by Sculpher in the context of menorrhagia therapy within the National Health Service for England and Wales [7], following the earlier work of Nease and Owens [8]. Sculpher confirmed that the two available interventions maximised the patient-specific QALYs for one subgroup of women; hence a strategy of offering treatment based on individual preferences at the point of care would, at least in principle, be a cost-effective public policy even in the collectively-funded system considered. This stimulated discussion about the possibility of implementing fully individual *patient* preference-based QALYs [9,10], a route subsequently explored by Basu and Meltzer [11-13] when developing their Expected Value of Individualised Care measure, and later by others [14-18].

However, none of these researchers seem enthusiastic about treating subgroup preferences as fundamental phenomena in driving health policy. Their implicit assumption is either that individual or subgroup preferences can be reduced to, and treated as, epiphenomena, i.e. as effectively being 'caused' by the biological-clinical and/or socio-demographic characteristics of the person or subgroup; or that preferences can be given policy relevance only if interpreted and processed through their associations with observable/verifiable objective characteristics of persons. The one exception, which 'proves the rule' - because subgrouping is not involved - is when preferences are elicited at the population level and used to produce a mean tariff applied to all individuals, as in the EQ-5D tariff used in QALY-based analyses. *If* it were decided to treat subgroup preferences as valid and independent determinants of public policy, a transparent analytical procedure will be needed.

The aim of this study is to present a procedure combining two analytical techniques that have not, thus far, featured in the debate: (i) Cluster Analysis (CA) which is used to generate preference subgroups, and (ii) Multi-Criteria Decision Analysis (MCDA) which provides the explicit policy framework for including clustered preferences. Our study has an empirical basis, and the data are from a large RCT about prostate cancer screening. However, the focus is on providing a proof of method for preference subgrouped public policy (via CA and MCDA). Thus the results are presented as a practical background to the discussion we hope to generate on this crucial issue. Our illustration highlights a number of issues that are likely to arise in any substantive implementation.

Methods

The two techniques used in this study, Cluster Analysis (CA) and Multi-Criteria Decision Analysis (MCDA), are separately well-established. However, their combined use in health-related research, as we propose, is innovative. We could only locate one other application of the idea, in production economics, where it was used to evaluate e-commerce enterprises [19]. Before turning to these techniques, we describe the data.

The data

For input into a public policy decision framed as a MCDA we required individual preferences from a representative sample of the population, expressed in the form of importance weights for different criteria relating to the decision. We used data from one arm of a Randomised Controlled Trial (RCT) of two online decision aids for Australian men aged 40-69 considering Prostate-specific Antigen (PSA) testing for prostate cancer, which was available and in the required format.[1]

Five criteria were provided in this arm of the trial:

LOSS OF LIFETIME: Avoid losing 5-10% of individual's remaining life expectancy.

NEEDLESS BIOPSY: Avoid having a needless biopsy.

URINARY PROBLEMS: Avoid urinary problems after treatment for prostate cancer.

BOWEL PROBLEMS: Avoid bowel problems after treatment for prostate cancer.

SEXUAL PROBLEMS: Avoid sexual problems (impotence) after treatment for prostate cancer.

These criteria were developed in the context of an individual decision aid, but we believe they are a reasonable set to explore as the effectiveness side of a public policy issue in a proof of method.

The criteria selected were based on the findings of a General Practitioner (GP) pilot study, a full account of which has been presented [20]. GPs provided information on the criteria we had included in the earlier version of the decision aid and other factors they thought were important for patients in making a decision about

PSA testing, thereby supplementing findings from the literature.

The RCT itself was based on a community sample of 1,970 men aged 40-69 years in 2011. Of these, 727 men were allocated to the arm where the interactive decision aid consisted of the five criteria outlined above.

The criterion weightings provided by respondent number 1526 can be seen in Figure 1, which displays the full MCDA decision aid screen. Using this web-based decision aid template, the importance weightings were elicited by respondents dragging the cursor to change the bar lengths, dynamically normalised to add to 100%. (MCDA as a technique does not elicit the inputs into it, but in this case the template was used as the preference-eliciting device.) The bottom Ratings panel contains the evidence base for the analysis in the form of the performance rates for the two options on the five criteria [20]. These ratings were made available to the respondent after their weightings had been elicited. (They were able to change their weightings after seeing this data, but virtually none did this and so it is the original weights which are clustered.) The top panel displays the scores for the two policy options, which result from combining the weightings of respondent number 1526 with the evidence-based ratings by way of a simple expected value calculation.

The only men excluded at survey entry were those with diagnosed prostate cancer. There were no exclusions for men 'at risk', so the 523 men whose preferences were cluster analysed included those reporting a first degree relative with prostate cancer (17%), or being unsure thereof (9%). 204 of the original 727 respondents had been previously excluded on the grounds that they had, at two distinct points in the survey, clicked the same point on a 10 point scale 8 times in a row as likely non-serious responders. (Respondents were recruited by an agency and received points for completion.)

The remaining 523 sets of criterion weights were analysed using CA to produce sets of subgroup means for input into MCDAs of PSA testing.

We supply the above details to give the reader some background to the importance weights being clustered, but emphasise that the methods by which they were elicited are largely irrelevant to our proof of method. Sets of weights may be produced by diverse methods, including Discrete Choice Experiments, and are suitable for clustering so long as they produce a full set of attribute weights for each individual.

Cluster analysis

CA and its various implementations are described in many texts [21-23]. There are several implementation packages, such as the R statistical package which was used in this study [24]. CA has been widely used in subgrouping on the basis of observable characteristics, ranging from types of gut bacteria at the cellular level [25] to the human level, where it is proving useful in the definition, diagnosis, and treatment of complex conditions, such as back pain [26,27] and fibromyalgia [28]. Bass and colleagues [29] used one of the main types of CA (k-means) in pursuit of their aim of nudging Afro-Americans towards colorectal cancer screening, identifying three subgroups which they labelled 'Ready screeners', 'Fearful avoiders' and 'Cautious screeners'.

Clustering

Three different techniques of CA were employed in this study to demonstrate not only that different techniques produce different clusters, but that choosing among clustering techniques is an important task itself in implementing the approach. We used Latent Class Analysis (MCLUST), Partitioning Around Medoids (PAMK) and Hierarchical Agglomeration via Ward's method

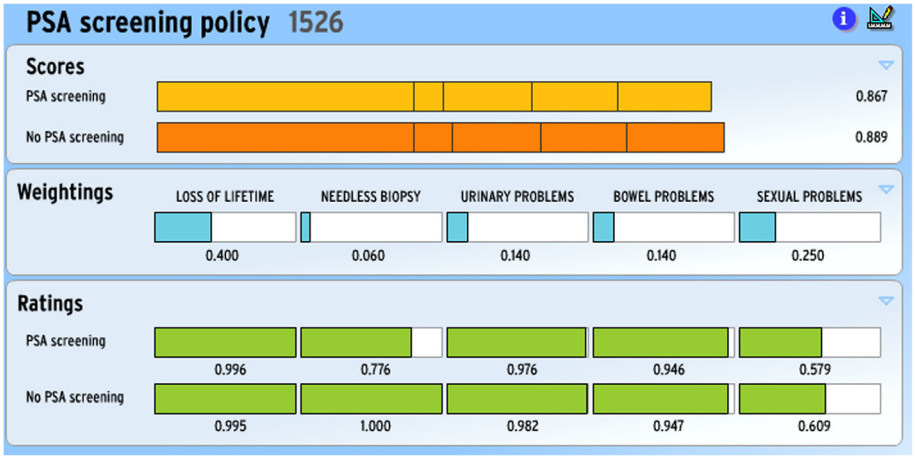


Figure 1 Annalisa MCDA screen with data for respondent 1526 in PSA decision aid trial.

(HCLUST), presenting the solutions generated by requesting 2, 3 and 4 clusters. The silhouette widths cluster quality indicator introduced below was calculated for solutions up to 9 clusters. The 2, 3 and 4 clusters included the maximal widths for all three methods and it was necessary to choose the same set for this comparative analysis. In all cases we used the R statistical package noted in parentheses, which makes our analyses accessible on an open source basis.

Latent Class Analysis (LCA) employs a model-based approach in which probabilities of cluster membership are estimated, and individuals are assigned to the cluster for which their membership probability is highest.

In *Partitioning* methods the cluster membership of an individual and hence the membership of clusters changes throughout the process. The aim is to find a solution that minimises the internal variance within clusters relative to a specified centroid (e.g. the medoid, or mean in the kmeans partitioning) and maximises the distance between cluster centroids.

In *Hierarchical Agglomeration* methods, individuals are progressively grouped in terms of their distance from each other in n-dimensions, where n is the number of criteria for clustering. Once assigned to a cluster they remain in that cluster, while the process of allocating unassigned individuals continues. The Ward's method is a special case, which assigns individuals to minimise the internal distance of each cluster at that point in the process.

Following the clustering analyses, and testing to see if the cluster solutions provided groupings significantly different on the five criteria using ANOVA, we allocated interpretive labels for each solution based on the weight assigned to the highest weighted criterion, and assessed the quality of the clusters produced by the alternative solutions. The evaluation of cluster solutions, which involves establishing the optimal number of clusters, as well as the quality of the grouping, has been the subject of continuing research since the early papers [30,31]. These issues are summarized [32].

It is widely acknowledged that cluster quality assessment is inherently multi-dimensional. Raskutti and Leckie (1999) suggest four criteria, but two of these four - the *compactness* of the cluster (i.e. the mean intra-cluster distance of observation from the centroid) and the *isolation* of the clusters (i.e. the mean inter-cluster distance) - are the ones most commonly used. They are the basis of the silhouette coefficient measure we chose [31]. Summary measures of cluster validity, and numerical differences between clustering solutions on such measures, must be interpreted in the light of the application of the clusters [22]. Considerations of efficiency and equity may lead to selection of a clustering solution which is not highest, or even very highly rated, in terms of purely statistical quality. In marketing, numerous other criteria impact on the selection of a cluster solution. Statistical quality is only one of these. The ten criteria below, collated from the marketing area [33], are all *potentially* relevant in our case. We would omit only criterion two, given our belief that preferences should be elicited directly and separately from 'objective' characteristics, in order not to treat people as a bundle of characteristics. We have translated the marketing terms into ones more appropriate for a health service setting:

1. Substantial: The subgroups are large enough to serve efficiently.
2. Accessible: The subgroups can be effectively reached and served, which requires them to be characterized by means of observable variables.
3. Differentiable: The subgroups can be distinguished conceptually and respond differently to different policy-mix elements and programs.
4. Actionable: Effective programs can be formulated to attract and serve the subgroups.
5. Stable: Only subgroups that are stable over time can provide the necessary grounds for a successful strategy.
6. Parsimonious: To be administratively meaningful, only a small set of substantial clusters should be identified.
7. Familiar: To ensure political acceptance, the subgroups composition should be comprehensible.
8. Relevant: Subgroups should be relevant in respect of the service's competencies and objectives.
9. Compactness: Subgroups exhibit a high degree of within-subgroup homogeneity and between-subgroup heterogeneity.
10. Compatibility: Subgroup results meet other administrative requirements.

Applying such criteria in a substantive application of our method is a task for which we believe MCDA is appropriate since it provides increased transparency in terms of specification of the importance attached to each criterion (the weightings) and the performance ratings of the available options on the criteria, as well as an explicit algorithm for combining the ratings and weightings to produce an overall opinion (the scores). Selecting a set of criteria and assigning importance weightings to them is one part of the task approached in this way. Arriving at ratings for how well each clustering technique/solution performs on each of the selected criteria is the second task. Integrating the weightings and ratings into an overall evaluation of each option is the final requirement, and in MCDA this is normally done using the expected value principle.

We fully accept that whether or not MCDA is the best, or an appropriate, approach to this task is itself a multi-criterial decision, involving both performance ratings and preferences.

Multi-criteria decision analysis

MCDA and its various forms are described and surveyed in numerous texts [34-39] and there are many examples of its use [1,37,40-45]. A large number of software implementations exist, reflecting both varying versions of MCDA and judgements about the extent and type of complexity to be catered for, as well as the time and cognitive resources required [46-49]. In the illustrative analyses reported here we employ Annalisa©, as used in the trial. Annalisa is an implementation of the simple linear additive version of MCDA, in which the scores for each option are produced by multiplying the performance rates for the option on each of the criteria by the respondent's weights for those criteria, and summing across criteria. Its one-screen-fits-all interface was specifically developed to be less complex in both development and delivery than the alternatives [20,49]. However, the selection of a software implementation of MCDA, like the selection of the CA technique (and indeed software for implementing it), is not something we wish to address on the present occasion. It would be a crucial part of the policy-specific development process.

The basic Annalisa screen (Figure 1) shows the expected value Scores which result from combining the evidenced-based Ratings for each policy Option on each criteria with the respondent's relative importance Weightings for the criteria. The data are for respondent number 1526 in the PSA trial from which our data are drawn - see below. (The No PSA score is higher for him, reflecting the importance Weightings he gave.)

Translation into MCDA-based policy analysis

The results for each of the four cluster solutions within the three CA techniques were fed into this MCDA tool, and the subgroup scores for each policy calculated. Subsequently, we conducted sensitivity analysis in relation to the Loss of Lifetime criterion, to see what change in the percentage rating for PSA vs. No PSA screening policy would be needed to bring each subgroup into equipoise, i.e. have equal scores for the two policy options. This seemed the most interesting of the many possible sensitivity analyses to undertake from a policy perspective, given it indicates the subgroup's trade-offs of harms with what is conventionally seen as the main potential benefit (Loss of Lifetime).

Results

Clustering

The clustering solutions from the three cluster techniques are shown in Table 1.[2] The mean subgroup weightings on the five criteria relevant to the PSA test decision (Loss of Lifetime, Needless Biopsy, Urinary Problems, Bowel Problems, and Sexual Problems) are shown for each solution.

Differences in the clusters produced, given the fixed criterion framing of the elicitation, are apparent. However, it is also clear that 3 broad preference patterns are common to all three of the 4 cluster solutions, which are the ones we focus on henceforth:

1. A relatively small subgroup of 10-11% 'Very High Lifers', for whom Loss of Lifetime is almost all-important with this criterion given 86-88% weight;
2. A relatively large subgroup of 'Moderate Lifers', comprising 23-49% of the sample who give this criterion 42-53% weight (and hence include respondent 1526 in Figure 1);
3. The largest group of all ('Equals') at 33-63% of the sample, who gave roughly equal weights to the five criteria (including 14-22% weight to Loss of Lifetime).

Setting these three subgroups apart, leaves a 'Very High Sexers' group at 7% and 11% of the sample who assigned 64% and 59% weights to the Sexual Problems criterion in the PAM and Ward solutions, respectively. They are replaced by 'Moderate Biopsers' at 4% with 53% weight assigned to Needless Biopsy in the LCA solution.

On the basis of roughly averaging this data, a policy based purely on Loss of Lifetime minimisation might just attract majority support.

The statistical quality of the solutions, as approximated by silhouette width, varies from .26 to .44 (see Table 1). A much reproduced scale would attach the label 'The structure is weak and could be artificial' to results in the .26-.5 range, but we can find no validation of this scale. In any case we believe that, as made clear earlier, clustering solutions should be evaluated by their external real-world consequences, as well as their internal qualities.

We have confirmed that different techniques and solutions produce different clusters. But also, that the resulting clusters are all capable of meaningful interpretations based on the most prominent criterion (or lack of one). However, to reiterate, we explicitly take no position on the issue of the most appropriate clustering technique, since this should be part of the policy development process and reflect the application of criteria other than statistical quality.

Table 1 Mean cluster weights from 2, 3 and 4 cluster solutions using LCA, PAM and Ward methods

Clustering Method	Cluster Solution	Cluster Number	N (of 523)	%	Quality	MEAN CRITERION WEGHTS					Interpretive Label
						LOSS OF LIFETIME	NEEDLESS BIOPSY	URINARY PROBLEMS	BOWEL PROBLEMS	SEXUAL PROBLEMS	
Latent Class Analysis (MCLUST)											
	4	1	327	62.5	0.24	0.22	0.15	0.20	0.20	0.23	Equals
		2	53	10.1	0.64	0.88	0.02	0.04	0.03	0.03	Very High Lifers
		3	121	23.1	0.31	0.53	0.06	0.14	0.15	0.12	Moderate Lifers
		4	22	4.2	0.39	0.13	0.53	0.11	0.11	0.12	Moderate Biopsers
					0.31						
	3	1	407	77.8	0.29	0.27	0.13	0.19	0.19	0.21	Equals
		2	92	17.6	0.60	0.78	0.03	0.07	0.06	0.06	Very High Lifers
		3	24	4.6	0.36	0.16	0.52	0.10	0.11	0.11	Moderate Biopsers
					0.35						
	2	1	493	94.3	0.25	0.36	0.11	0.17	0.17	0.18	Moderate Lifers
		2	30	5.7	0.36	0.22	0.49	0.10	0.10	0.10	Moderate Biopsers
					0.26						
Partitioning Around Medoids (pamk)											
	4	1	270	51.6	0.33	0.19	0.18	0.22	0.22	0.19	Equals
		2	59	11.3	0.63	0.87	0.03	0.04	0.03	0.03	Very High Lifers
		3	163	31.2	0.26	0.49	0.11	0.14	0.14	0.13	Moderate Lifers
		4	31	5.9	0.36	0.06	0.06	0.11	0.13	0.64	Very High Sexers
					0.35						
	3	1	301	57.6	0.27	0.18	0.17	0.21	0.21	0.24	Equals
		2	59	11.3	0.63	0.87	0.03	0.04	0.03	0.03	Very High Lifers
		3	163	31.2	0.30	0.49	0.11	0.14	0.14	0.13	Moderate Lifers
					0.32						
	2	1	346	66.2	0.40	0.21	0.16	0.20	0.20	0.23	Equals
		2	177	33.8	0.41	0.64	0.08	0.09	0.10	0.09	Very High Lifers
					0.41						
Ward's Hierarchical (HCLUST)											
	4	1	170	32.5	0.34	0.14	0.21	0.23	0.23	0.18	Equals
		2	38	7.3	0.27	0.08	0.07	0.12	0.14	0.59	Very High Sexers
		3	60	11.5	0.68	0.86	0.03	0.04	0.04	0.03	Very High Lifers

Table 1 Mean cluster weights from 2, 3 and 4 cluster solutions using LCA, PAM and Ward methods *(Continued)*

	4	255	48.8	0.17	0.42	0.12	0.15	0.16	0.15	Moderate Lifers
				0.29						
3	1	208	39.8	0.22	0.13	0.19	0.21	0.21	0.26	Equals
	2	60	11.5	0.68	0.86	0.03	0.04	0.04	0.03	Very High Lifers
	3	255	48.8	0.23	0.42	0.12	0.15	0.16	0.15	Moderate Lifers
				0.28						
2	1	463	88.5	0.40	0.29	0.15	0.18	0.18	0.20	Moderate Lifers
	2	60	11.5	0.76	0.86	0.03	0.04	0.04	0.03	Very High Lifers
				0.44						

Also shown are cluster sizes and statistical quality (as measured by average silhouette width). The bold numbers indicate the statistical quality of the cluster solution. N.B. ANOVA showed all clusters to be significant at $p < 0.05$, except LCA 4/4 (Moderate Biopsers).

Entering cluster weights into MCDAs

Pursuing our proof of method, the results from the 4 cluster solutions from the three techniques were now inserted into MCDAs.

None of the preference-based subgroups produced by *any* clustering solution favours a PSA screening policy. There are various ways in which the complex set of results could be displayed, but we feel it most informative to present just one type of sensitivity/threshold analysis. Given the *weight* assigned by a subgroup to the Loss of Lifetime criterion, what proportionate change in the *ratings* for the two policy options on this criterion would result in this subgroup being in policy equipoise (i.e. the option scores being equal in its MCDA)?

The answers for all three of the 4 cluster solutions are presented in Table 2, with Additional file 1: Tables S1, S2 and S3 providing the full calculations, and S4 an illustration of the calculation procedure.

The table confirms that the required changes are a direct reflection of the subgroups' weights, with (in the Ward solution), Very High Lifers (86% weight to Loss of Lifetime) requiring a 1% improvement, and Moderate Lifers (42% weight) an 8% improvement. The high (39%) requirement for Equals reflects their low (14%) weight for Loss of Lifetime, which is not much greater than that of Very High Sexers. The requirement patterns in the LCA and PAM solutions are similar. But the result for Moderate Biopsers in LCA (95%) while it is consistent with the 13% weight assigned to Lifetime Loss, is a useful warning of the need to be cautious in selecting a solution. It is from the one cluster that was not significant in ANOVA (see Table 1 caption).

Table 2 Percentage increase in gap between relative Loss of Lifetime performance ratings for PSA and No PSA screening options needed to produce equipoise for each 4 cluster solution

Cluster	LCA	PAM	Ward's
Equals	19	25	39
Very High Lifers	1	1	1
Moderate Lifers	3	6	8
Very High Sexers	…	56	43
Moderate Biopsers	95	…	…

Age-stratified results

Following the exclusion of those participants 'at risk' of prostate cancer or 'unsure' about their family history, the sample for age-stratified clustering became 388. 156 were in their 40s, 135 in their 50s, and 97 in their 60s.

The same type of interpretable subgroups reappear with different distributions (Additional file 1: Tables S5, S6, S7), but with notably different thresholds on the Loss of Lifetime criterion to produce equipoise. (Table 3) (These were calculated in the same way as illustrated in Additional file 1: Table S4.)

It seems a reasonable inference that age effects exist. The proportions (%N) of both Moderate and Very High Lifers increase progressively from younger to older at the same time, as their equipoise requirement progressively increases. This necessitates that the opposite happens for the proportions of the other subgroups, and we indeed observe that Equals increase from 32% to 44% moving from youngest to oldest groups. Their equipoise requirement also rises dramatically, from near equipoise for the 40s (0.4%) to 21.5% for the 60s. The residual subgroup proportion increases from 8 % to 15%. In the 40s and 50s it is the Very High Sexers, who are in virtual equipoise in the 40s, but significantly divergent from it in the 50s (14.4% requirement). However, in the 60s this subgroup is replaced by Moderate Biopsers, a cluster dominated by concern with needless testing.

Table 3 Percentage increase in gap between relative Loss of Lifetime performance ratings for PSA and No PSA screening options needed to produce equipoise for each 4 cluster solution, by age group

	40-49 years		50-59 years		60-69 years	
	% Change	%N	% Change	%N	% Change	%N
Moderate Lifers	0.1	35	2.7	27	4.1	26
Very High Lifers	0.0	25	0.3	24	0.4	14
Equals	0.4	32	3.5	41	21.5	44
Very High Sexers	0.2	8	14.4	8	…	…
Moderate Biopsers	…	…	…	…	45.4	15

All these variations have modest appeal in terms of face validity, but any inferences need to be drawn with caution, since the three clustering solutions are for different datasets (albeit from same responders), and so are not directly comparable. These age effects are the combined effect of different criterion performance ratings for the age groups as well as different preference patterns.

Discussion

This study presents an example of how public preferences *could* be incorporated into policy decisions respecting both the multi-criterial nature of those decisions and the heterogeneity of the population in relation to their weightings. The various methodological and practical issues to be addressed in implementing such an approach are emphasised. Always to be determined are: the structure of the policy decision (options, criteria in the MCDA); the choice of MCDA version and implementation software; the choice of CA technique; the choice of number of cluster solutions and measure of cluster quality; and the trade-offs between statistical quality and other criteria. It is the primary aim of this paper to ensure that these issues are addressed transparently, rather than dealt with in an exclusively deliberative process.

Objections to cluster analysis as an 'unsupervised' technique only to be used in abductive hypothesis generating – with the resulting clusters requiring 'validation' against some other criterion and insertion into a hypothesis testing framework [27] – are of little relevance to our approach. There is no gold standard against which preference clusters can be compared. We have made clear that regression of preference clusters on biological-clinical or socio-demographic variables is inappropriate, because we are in a policy/decision making practice context, not a hypothesis-testing or scientific research-driven one.

While the decision on which solution to adopt in the presence of clustering differences requires consideration of factors other than statistical quality, one thing should not enter into analysis at the policy level in relation to preference subgrouping regardless of the method used: the characteristics of those individuals who move between clusters depending on the technique and solution. Tracing such individual movements is feasible in all software implementations of cluster analysis, but there seems to be no conceptual justification for doing so. In this sort of analysis an individual is simply a person expressing their preferences in the context of a particular decision. It is vital they are not treated as a 'bundle of variables'. In some practice contexts it will be appropriate to explore the statistical relationship between preference-based subgroups and objective characteristics, typically via regression analysis. Or to look forward and explore the relationship with some future outcome or behavior, probably also via regression analysis. But we argue that neither of these explorations is appropriate when it involves reducing the preferences of a person, or group of persons, to a set of predictive or predictor variables, since this undermines the fundamental personhood of the preference-bearer [50].

A mini-debate provoked by a comment by Robinson and Parkin on their paper [51,52] made clear that one central issue is whether public or patient preferences are appropriate. We are explicitly operating in the extra-welfarist framework where *stated public preferences over outcomes* are the inputs relevant for a subgrouped public policy, not *revealed patient choice of options*. In a collectively-funded health care system we take the view that it is the preferences of members of the public, as citizens which are the appropriate inputs into policy, leaving patient preferences to be applied at the individual/clinical level within the constraints set by community policy. Of course, there is nothing in the techniques themselves which rule out using patient preferences as inputs, but the conflict of personal and public interest at, or near, the point of care, poses major challenges to using those of patients.

We do not address the cost side of policy making here, instead concentrating on how subgroup preferences in relation to effectiveness criteria could be incorporated into Cost-effectiveness Analysis and public policies. As emphasised by Claxton it is important that an MCDA-based policy operating within a budget constraint respects the existence of opportunity costs, ensuring that

any net benefit foregone from the expansion of the criteria on the effectiveness side (beyond QALYs) should be taken into account [53].

In an extended MCDA framework it would be possible to include options that fall within of the South-West quadrant of the cost-effectiveness plane, i.e. are cost-effective by being less effective, but proportionately much cheaper, than the standard one [54]. And one might include an explicit 'Net effect on (generalised) others' criterion for individual respondents to *weight*. In the extreme, this could be split into two on the basis of the 'just deserts' criteria that emerges in most public surveys. We are not advocating this, simply confirming that moving to an MCDA-based public policy will make such issues and their resolution more transparent.

A crucial finding in the Raskutti and Leckie paper, replicating that of Macskassy, is that humans asked to cluster the same data as a CA program, produce equivalent variation in both the optimal number of clusters and their content [32,55]. In other words, individual policy makers engaging in subgrouping are unlikely to outperform a cluster solution, so the same discussion will be needed if policy makers undertake the task.

Conclusions

In attempting to respect the heterogeneity of population preferences in public policy, a subgroup approach of some sort is inevitable. In this paper we illustrate how two types of analysis might, in combination, represent a viable approach. The implementation of Cluster Analysis and Multi-Criteria Decision Analysis, individually and in combination, poses major challenges - conceptual, methodological, ethical-political, and practical. We outline these challenges in the paper, stressing that most are only exposed by these more analytical techniques, not created by them. Alternative analytical or deliberative approaches will face similar challenges, and any proper evaluation must involve comparison of the approaches in empirical practice, not simply against diverse sets of normative principles. This is particularly important because computer technologies quickly expose the 'digital divide', easily obscured in deliberative approaches. Such unbiased comparative evaluation is the next item on the research agenda.

The empirical results from our PSA screening example are consistent with the trend away from advocacy of PSA screening of asymptomatic men without a family history of prostate cancer, based on both worries about the test and preference considerations [56]. But the fact that our results are in line with this observed trend should not be misinterpreted. All we have sought to show as proof of method, is that one can carry out analyses that identify the improvement in criterion performance (e.g. a superior test, less subsequent problems from treatment) needed for a preference-based subgroup to favour a screening policy.

Our finding of age-based preference subgrouping raises the question of whether sub-subgrouping individual preferences on bases such as age, sex, ethnicity, or religion is consistent with truly *person*-centred public policy.

Endnotes

[1] The trial from which the data come was approved by the University of Sydney HREC (Protocol No.: 05-2011/13712) on May 13 2011 and was included in the Australian New Zealand Clinical Trials Registry (ANZCTR) on 6 July 2012 (ACTRN12612000723886) (https://www.anzctr.org.au/Trial/Registration/TrialReview.aspx?id=343044).

[2] An early version of this paper was presented in a poster at the Lancet Public Health Science conference in November 2013 [57]. This contains links which will enable the reader to engage in interactive exploration of the data in a downloadable spreadsheet and to explore the survey as seen by a respondent.

Additional file

Additional file 1: Table S1. LCA 4 cluster solution subgroup mean weights input into MCDA, policy scores generated and threshold on Loss of Lifetime identified. **Table S2.** PAM 4 cluster solution subgroup mean weights input into MCDA, policy scores generated and threshold on Loss of Lifetime identified. **Table S3.** Ward 4 cluster solution subgroup mean weights input into MCDA, policy scores generated and threshold on Loss of Lifetime identified. **Table S4**. Derivation of proportionate change in Loss of Lifetime ratings for the policy options required by Very High Sexers subgroup (Ward 4 solution) to achieve policy equipoise. **Table S5.** Ward 4 cluster solution subgroup mean weights for 40-49 year olds input into MCDA, policy scores generated and threshold on Loss of Lifetime identified. **Table S6.** Ward 4 cluster solution subgroup mean weights for 50-59 year olds input into MCDA, policy scores generated and threshold on Loss of Lifetime identified. **Table S7.** Ward 4 cluster solution subgroup mean weights for 60-69 year olds input into MCDA, policy scores generated and threshold on Loss of Lifetime identified.

Abbreviations

ANOVA: Analysis of Variance; CA: Cluster Analysis; GP: General Practitioner; LCA: Latent Class Analysis; MCDA: Multi-Criteria Decision Analysis; NICE: National Institute for Health and Care Excellence; PAM: Partitioning Around Medoids; PSA: Prostate-Specific Antigen; QALY: Quality-Adjusted Life Year; RCT: Randomised Controlled Trial.

Competing interests

Jack Dowie has a financial interest in the Annalisa software but did not benefit from its use in the trial from which the data in this paper are drawn. No conflicts were reported by other authors.

Authors' contributions

MKK and JD conceived the study, undertook preliminary cluster analyses in SPSS and STATA, entered the cluster results into MCDAs, and are primarily responsible for the interpretation of the results and the contents of the paper. The final clustering analyses were undertaken in R by RT. JBN, GS, and MC commented extensively on the analysis and paper drafts. GS was Principal Investigator and MC and JD were Co-Investigators in the trial from which the illustrative data come. All authors approved the final paper.

Author details
[1]Research Unit for General Practice, Department of Public Health University of Southern Denmark, J.B. Winsløws Vej 9 B, 5000 Odense C, Denmark. [2]School of Public Health and Community Medicine, University of New South Wales, Sydney NSW 2052, Australia. [3]NHMRC Clinical Trials Centre, Sydney Medical School, Charles Perkins Centre, Johns Hopkins Drive, Camperdown NSW 2050, Australia. [4]Faculty of Medicine, School of Public Health University of Sydney, Edward Ford Building (A27), Sydney NSW 2006, Australia. [5]Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London WC1H 9SH, UK.

Received: 8 October 2014 Accepted: 8 April 2015
Published online: 14 May 2015

References

1. Devlin NJ, Sussex J. Incorporating Multiple Criteria in HTA: Methods and Processes. London: Office of Health Economics; 2012.
2. Bjertnaes O, Skudal KE, Iversen HH. Classification of patients based on their evaluation of hospital outcomes: cluster analysis following a national survey in Norway. BMC Health Serv Res. 2013;13:73. doi: 10.1186/1472-6963-13-73.
3. Williams SS, Heller A. Patient activation among Medicare beneficiaries: Segmentation to promote informed health care decision making. Int J Pharm Healthc Mark. 2007;1:199–213. doi: 10.1108/17506120710818210.
4. Berg AL, Sandahl C, Clinton D. The relationship of treatment preferences and experiences to outcome in generalized anxiety disorder (GAD). Psychol Psychother. 2008;81:247–59. doi: 10.1348/147608308X297113.
5. Dolan JG, Boohaker E, Allison J, Imperiale TF. Patients' preferences and priorities regarding colorectal cancer screening. Med Decis Mak. 2013;53:59–70. doi: 10.1177/0272989X12453502.
6. Deal K. Segmenting patients and physicians using preferences from discrete choice experiments. Patient. 2014;7:5–21. doi: 10.1007/s40271-013-0037-9.
7. Sculpher M. The cost-effectiveness of preference-based treatment allocation: the case of hysterectomy versus endometrial resection in the treatment of menorrhagia. Health Econ. 1998;7:129–42. doi: 10.1002/(SICI)1099-1050(199803)7:2<129::AID-HEC332>3.0.CO;2-9.
8. Nease RF, Owens DK. A Method for Estimating the Cost- Effectiveness of Incorporating Patient Preferences into Practice Guidelines. Med Decis Mak. 1994;14:382–92.
9. Dowie J. Towards the equitably efficient and transparently decidable use of public funds in the deep blue millennium. Health Econ. 1998;7:93–103. doi: 10.1002/(SICI)1099-1050(199803)7:2<93::AID-HEC313>3.0.CO;2-2.
10. Granata A, Hillman A. Competing practice guidelines: using cost-effectiveness analysis to make optimal decisions. Ann Intern Med. 1998;128:56–63.
11. Basu A, Meltzer D. Value of information on preference heterogeneity and individualized care. Med Decis Mak. 2007;27:112–27. doi: 10.1177/0272989X06297393.
12. Basu A. Individualization at the heart of comparative effectiveness research: the time for i-CER has come. Med Decis Mak. 2009;29:NP9–NP11. doi: 10.1177/0272989X09351586.
13. Basu A. Economics of individualization in comparative effectiveness research and a basis for a patient-centered health care. J Health Econ. 2011;30:549–59. doi: 10.1016/j.jhealeco.2011.03.004.
14. Brazier JE, Dixon S, Ratcliffe J. The role of patient preferences in cost-effectiveness analysis: a conflict of values? PharmacoEconomics. 2009;27:705–12. doi: 10.2165/11314840-000000000-00000.
15. Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. Pharmacoeconomics. 2008;26:799–806. doi: 10.2165/00019053-200826090-00009.
16. Sculpher M. Reflecting heterogeneity in patient benefits: the role of subgroup analysis with comparative effectiveness. Value Heal. 2010;13 Suppl 1:S18–21. doi: 10.1111/j.1524-4733.2010.00750.x.
17. Grutters JPC, Sculpher M, Briggs AH, Severens JL, Candel MJ, Stahl JE, et al. Acknowledging patient heterogeneity in economic evaluation: a systematic literature review. Pharmacoeconomics. 2013;31:111–23. doi: 10.1007/s40273-012-0015-4.
18. van Gestel A, Grutters J, Schouten J, Webers C, Beckers H, Joore M, et al. The role of the expected value of individualized care in cost-effectiveness analyses and decision making. Value Heal. 2012;15:13–21. doi: 10.1016/j.jval.2011.07.015.
19. Mistry J, Sarkis J, Dhavale DG. Multi-criteria analysis using latent class cluster ranking: An investigation into corporate resiliency. Int J Prod Econ. 2014;148:1–13. doi: 10.1016/j.ijpe.2013.10.006.
20. Cunich M, Salkeld G, Dowie J, Henderson J, Bayram C, Britt H, et al. Integrating evidence and individual preferences using a web-based Multi-Criteria Decision Analytic tool: An application to Prostate Cancer screening. Patient. 2011;4:1–10. doi: 10.2165/11587070-000000000-00000.
21. Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis. 5th ed. Chichester: Wiley; 2011. p. 346.
22. Tan P-N, Steinbach M, Kumar V. Introduction to data mining. Harlow: Pearson; 2013. p. 568.
23. Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. New York: Wiley; 2005. p. 368.
24. Maechler M., Rousseeuw, P, Struyf A, Hubert M, Hornik K. cluster: Cluster Analysis Basics and Extensions. R package version 1.14.4
25. Manichanh C, Borruel N, Casellas F, Guarner F. The gut microbiota in IBD. Nat Rev Gastroenterol Hepatol. 2012;9:599–608. doi: 10.1038/nrgastro.2012.152.
26. Axén I, Bodin L, Bergström G, Halasz L, Lange F, Lövgren PW, et al. Clustering patients on the basis of their individual course of low back pain over a six month period. BMC Musculoskelet Disord. 2011;12:99. doi: 10.1186/1471-2474-12-99.
27. Kent P, Keating JL, Leboeuf-Yde C. Research methods for subgrouping low back pain. BMC Med Res Methodol. 2010;10:62. doi: 10.1186/1471-2288-10-62.
28. Bennett RM, Russell J, Cappelleri JC, Bushmakin AG, Zlateva G. Identification of symptom and functional domains that fibromyalgia patients would like to see improved: a cluster analysis. BMC Musculoskelet Disord. 2010;11:134. doi: 10.1186/1471-2474-11-134.
29. Bass SB, Gordon TF, Ruzek SB, Wolak C, Ruggieri D, Mora G, et al. Developing a computer touch-screen interactive colorectal screening decision aid for a low-literacy African American population: lessons learned. Health Promot Pract. 2013;14:589–98. doi: 10.1177/1524839912463394.
30. Dubes R, Jain AK. Validity studies in clustering methodologies. Pattern Recognit. 1979;11:235–54. doi: 10.1016/0031-3203(79)90034-7.
31. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65. doi:10.1016/0377-0427(87)90125-7.
32. Raskutti B, Leckie C. An Evaluation of Criteria for Measuring the Quality of Clusters. In: Proc 16th Int Jt Conf Artif Intell, vol. 2. 1999. p. 905–10.
33. Mooi E, Sarstedt M. Cluster Analysis. In: Mooi E, Sarstedt M, editors. A Concise Guide to Market Research. Berlin, Heidelberg: Springer; 2011. p. 237–84. doi: 10.1007/978-3-642-12541-6_9.
34. Department of Communities and Local Government. Multi-criteria analysis: a manual. London: Department of Communities and Local Government; 2009. ISBN: 9781409810230.
35. Belton V, Stewart TJ. Multiple Criteria Decision Analysis: An Integrated Approach. Dordrecht: Kluwer; 2002.
36. Figueria J, Greco S, Ehrgott M. Multiple Criteria Decision Analysis: State of the Art Surveys. Boston: Kluwer; 2005.
37. Diaby V, Campbell K, Goeree R. Multi-criteria decision analysis (MCDA) in health care: A bibliometricanalysis. Oper Res Heal Care. 2013;2:20–4. doi: 10.1016/j.orhc.2013.03.001.
38. Adunlin G, Diaby V, Montero AJ, Xiao H. Multicriteria decision analysis in oncology. Heal Expect. 2014 doi: 10.1111/hex.12178
39. Diaby V, Goeree R. How to use multi-criteria decision analysis methods for reimbursement decision-making in healthcare: a step-by-step guide. Expert Rev Pharmacoecon Outcomes Res. 2014;14:81–99. doi: 10.1586/14737167.2014.859525.
40. Dolan JG. Multi-criteria clinical decision support: A primer on the use of multiple criteria decision making methods to promote evidence-based, patient-centered healthcare. Patient. 2010;3:229–48. doi: 10.2165/11539470-000000000-00000.
41. Thokala P, Duenas A. Multiple criteria decision analysis for health technology assessment. Value Heal. 2012;15:1172–81. doi: 10.1016/j.jval.2012.06.015.
42. Baltussen R, Niessen L. Priority setting of health interventions: the need for multi-criteria decision analysis. Cost Eff Resour Alloc. 2006;4:14. doi: 10.1186/1478-7547-4-14.
43. Tony M, Wagner M, Khoury H, Rindress D, Papastavros T, Oh P, et al. Bridging health technology assessment (HTA) with multicriteria decision analyses (MCDA): field testing of the EVIDEM framework for coverage

decisions by a public payer in Canada. BMC Health Serv Res. 2011;11:329. doi: 10.1186/1472-6963-11-329.
44. Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Evidence and Value: Impact on DEcisionMaking–the EVIDEM framework and potential applications. BMC Health Serv Res. 2008;8:270. doi: 10.1186/1472-6963-8-270.
45. Goetghebeur MM, Wagner M, Khoury H, Levitt RJ, Erickson LJ, Rindress D. Bridging health technology assessment (HTA) and efficient health care decision making with multicriteria decision analysis (MCDA): applying the EVIDEM framework to medicines appraisal. Med Decis Mak. 2012;32:376–88. doi: 10.1177/0272989X11416870.
46. Riabacke M, Danielson M, Ekenberg L. State-of-the-art prescriptive criteria weight elicitation. Adv Decis Sci. 2012; 1–24. doi: 10.1155/2012/276584
47. de Montis A, deToro P, Droste-franke B, Omann I, Stagl S. Assessing the quality of different MCDA methods. In: Getzner M, Spash CL, Stagl S, editors. Alternatives for environmental evaluation. Abingdon: Routledge; 2004. p. 99–133.
48. Wallenius J, Dyer JS, Fishburn PC, Steuer RE, Zionts S, Deb K. Multiple criteria decision making, Multiattribute Utility Theory: Recent accomplishments and what lies ahead. Manage Sci. 2008;54:1336–49. doi: 10.1287/mnsc.1070.0838.
49. Dowie J, Kjer Kaltoft M, Salkeld G, Cunich M. Towards generic online multicriteria decision support in patient-centred health care. Heal Expect. 2013 doi: 10.1111/hex.12111
50. Entwistle V, Watt IS. A capabilities approach to person-centered care: response to open peer commentaries on "Treating patients as persons: a capabilities approach to support delivery of person-centered care". Am J Bioeth. 2013;13:W1–4. doi: 10.1080/15265161.2013.812487.
51. Robinson A, Parkin D. Recognising diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. A response to Sculpher and Gafni. Health Econ. 2002;11:649–51. doi: 10.1002/hec.735.
52. Sculpher M, Gafni A. Recognising diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. Authors' Reply Health Econ. 2002;11:653–4. doi: 10.1002/hec.736.
53. Claxton K. Three questions to ask when examining MCDA. Value & Outcomes Spotlight. 2015;1:18-20.
54. Dowie J. Why cost-effectiveness should trump (clinical) effectiveness: the ethical economics of the South West quadrant. Health Econ. 2004;13:453–9. doi: 10.1002/hec.861.
55. Macskassy SA, Banerjee A, Davison BD, Hirsh H. Human Performance on Clustering Web Pages: A Preliminary Study. In: Fourth Int Conf Knowl Discov Data Min. 1998. p. 264–8.
56. Ilic D, Neuberger M, Djulbegovic M, Dahm P. Screening for prostate cancer (Review). Cochrane Database Syst Rev. 2013. doi: 10.1002/14651858.CD004720.pub3
57. Kaltoft MK, Dowie J, Turner R, Nielsen JB, Salkeld G, Cunich M. Addressing the disconnect between public health science and personalised health care: the potential role of cluster analysis in combination with multi-criteria decision analysis. Lancet. 2013;383:S52. doi: 10.1016/S0140-6736(13)62477-0.

Original Paper

Bringing Feedback in From the Outback via a Generic and Preference-Sensitive Instrument for Course Quality Assessment

Mette K Kaltoft[1], MPH; Jesper B Nielsen[1], PhD; Glenn Salkeld[2], PhD; Jo Lander[2], PhD; Jack Dowie[3], PhD.

[1]Research Unit for General Practice, Department of Public Health, University of Southern Denmark, Odense, Denmark

[2]School of Public Health, University of Sydney, Sydney, Australia

[3]Faculty of Public Health and Policy, London School of Hygiene and Tropical Medicine, London, United Kingdom

Corresponding Author:
Jack Dowie, PhD.
Faculty of Public Health and Policy
London School of Hygiene and Tropical Medicine
15-17 Tavistock Place
London, WC1H 9SH
United Kingdom
Phone: 44 2079272034
Fax: 44 2079272034
Email: jack.dowie@lshtm.ac.uk

Abstract

Background: Much effort and many resources have been put into developing ways of eliciting valid and informative student feedback on courses in medical, nursing, and other health professional schools. Whatever their motivation, items, and setting, the response rates have usually been disappointingly low, and there seems to be an acceptance that the results are potentially biased.

Objective: The objective of the study was to look at an innovative approach to course assessment by students in the health professions. This approach was designed to make it an integral part of their educational experience, rather than a marginal, terminal, and optional add-on as "feedback". It becomes a weighted, but ungraded, part of the course assignment requirements.

Methods: A ten-item, two-part Internet instrument, MyCourseQuality (MCQ-10D), was developed following a purposive review of previous instruments. Shorthand labels for the criteria are: Content, Organization, Perspective, Presentations, Materials, Relevance, Workload, Support, Interactivity, and Assessment. The assessment is unique in being dually personalized. In part 1, at the beginning of the course, the student enters their importance weights for the ten criteria. In part 2, at its completion, they rate the course on the same criteria. Their ratings and weightings are combined in a simple expected-value calculation to produce their dually personalized and decomposable MCQ score. Satisfactory (technical) completion of both parts contributes 10% of the marks available in the course. Providers are required to make the relevant characteristics of the course fully transparent at enrollment, and the course is to be rated as offered. A separate item appended to the survey allows students to suggest changes to what is offered. Students also complete (anonymously) the standard feedback form in the setting concerned.

Results: Piloting in a medical school and health professional school will establish the organizational feasibility and acceptability of the approach (a version of which has been employed in one medical school previously), as well as its impact on provider behavior and intentions, and on student engagement and responsiveness. The priorities for future improvements in terms of the specified criteria are identified at both individual and group level. The group results from MCQ will be compared with those from the standard feedback questionnaire, which will also be completed anonymously by the same students (or some percentage of them).

Conclusions: We present a protocol for the piloting of a student-centered, dually personalized course quality instrument that forms part of the assignment requirements and is therefore an integral part of the course. If, and how, such an essentially formative Student-Reported Outcome or Experience Measure can be used summatively, at unit or program level, remains to be determined, and is not our concern here.

(JMIR Res Protoc 2015;4(1):e15) doi:10.2196/resprot.4012

KEYWORDS

medical education; nursing education; course assessment; course evaluation; student feedback; Internet; personalization

Introduction

Over several decades great efforts have been put into developing ways of eliciting valid and informative student feedback on courses they have taken in medical, nursing, and other health professional schools, and in continuing education and professional development. An important motivation has been "formative", to help providers—teachers and related services—to improve what is offered. Their use in "summative" ways for administrative purposes, such as institutional promotion or staff evaluation, has increased greatly in recent years. However, whatever their motivation, items, and setting, the response rates have usually been low—only rarely above, or even approaching 50%—and potentially biased as a result. Many responses are produced cursorily, with little sense of engagement with a serious task. We see one of the main reasons for this as being its marginalized and optional status as "feedback" at the termination of the course, whether it is a day, or a year, long. Our goal is a new, generic, *course quality* assessment instrument and process, aimed at not only generating insights for the course provider into potential sources of improvement, but also, through the personalized and structured reflection it involves and encourages, enhancing the educational experience of the student. (In some countries and educational settings the term "evaluation" would be used instead of "assessment" in our context. We use the latter to embrace the former, for reasons that will become apparent.)

Why is a new instrument of this sort needed? A recent systematic review covers the vast literature on student evaluation and the instruments relating to it comprehensively and in depth [1]. While some of the numerous instruments are generic, applicable to all courses whatever the subject or focus, none produces a preference-sensitive index score, for example, an overall quantitative assessment that combines the individual student's weightings for a set of quality criteria (dimensions) with their performance ratings for each of those criteria. Often course assessments are left as an unsynthesized profile of responses, but even where an index score is produced by some weighting procedure (including implicit equal weighting), the weights are not personalized. There is, therefore, a need for a generic and "dually personalized" measure of course quality, paralleling that in decision quality [2].

Beyond these two meta-criteria of genericness and preference-sensitivity, a third fundamental requirement is operational practicality. The instrument must be compatible with the time and other resources of students, on the one hand, and, if it were to be used summatively, capable of providing simple and actionable analyses by providers, on the other. But we see this practicality being established in the context of a substantially enhanced role for course assessment, which is now to be seen as a key source of the student's benefit from the course. Without going so far as to suggest that, paraphrasing Socrates, "the unassessed course is not worth pursuing", we believe that student assessment of the quality of the course they are taking should be a formal part of it, not an optional, terminal add-on conceptualized merely as feedback. The idea is novel, but simply seeks to take advantage of, and gives direction to, the informal and unstructured judgements about, and reactions to, the course, that are occurring every moment the student is engaged with it.

Methods

Sources for the Course Assessment Instrument

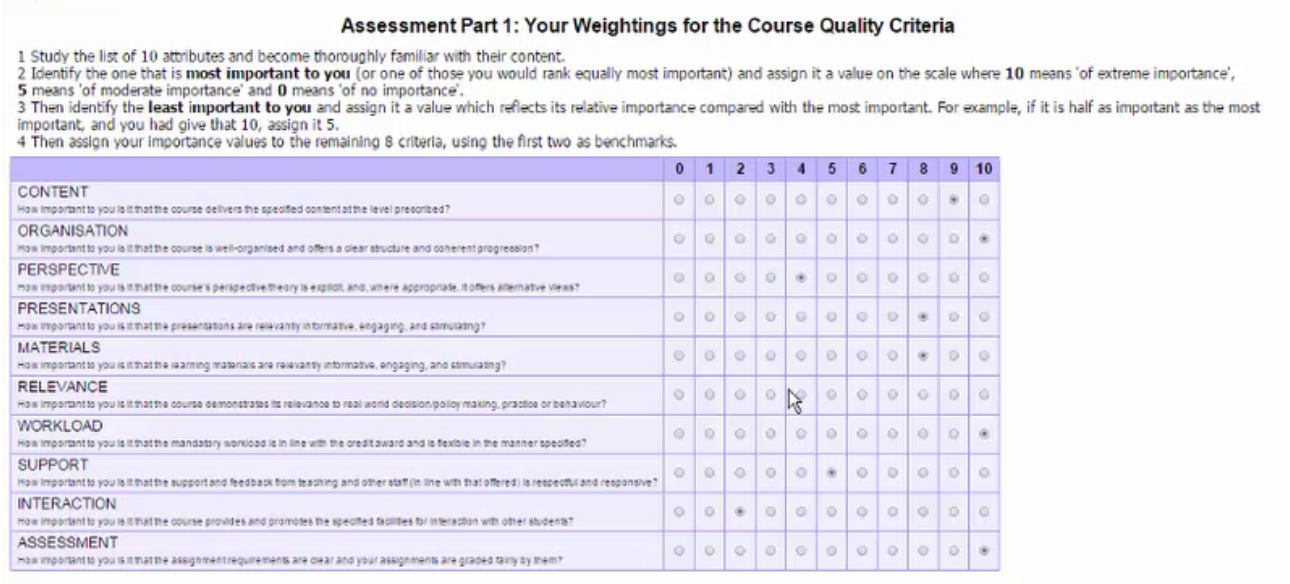
A purposive survey of key references was sufficient to establish a comprehensive list of the attributes/criteria/dimensions that have been used in course assessment, evaluation, and feedback by students. Apart from the tabulation in Spooren [1], we consulted ten other sources: (1) Alderman et al [3], (2) Chalmers [4], (3) Coates [5], (4) Davies et al [6], (5) Fontaine et al [7], (6) Kember and Leung [8], (7) Marsh and Roche [9], (8) Ramsden [10], (9) Richardson [11], and (10) Palmer [12].

Since the instruments reported in these studies were the result of extensive research and validation, the task in constructing a new instrument was not to add to the resulting list of criteria, but to reduce it to ten, the absolute maximum practical for routine use, especially in relation to criterion weighting. Both sets of responses are elicited on a 0 to 10 scale. The ten criteria would need definitions that were meaningful, in the sense that a single value on a 0 to 10 ratio scale could be provided as a response at both the weighting and rating stages. For weighting responses 0, 5, and 10 are labelled as "of no importance", "of moderate importance", and "of extreme importance", respectively, and those values are labelled as performing "extremely poorly", "moderately well", and "extremely well" for course rating. It is made explicit in the instructions (Figure 1 shows this, later) that the scales are to be interpreted as ratio ones, as is necessary for the expected value calculation that produces the MyCourseQuality-10 Dimensions (MCQ-10D) index score (eg, 8 is to be twice as important as 4 on the weighting scale). (Some of the 10 criteria necessarily embrace the subcriteria and subsubcriteria included in more complex assessment instruments, and in these cases, the respondent's holistic high-level response will imply subweighting of these. For example, course materials may include different types of material, such as journal articles; videos; and applications for mobile devices.)

The final set of criteria for MCQ-10D was arrived at by considering the reported construct and content validity of the previous instruments, and maximizing comprehensiveness of coverage and conceptual independence within the constraint of 10 criteria. This necessarily involved making trade-offs based on value judgements, rather than purely statistical procedures.

The protocol for the piloting of the MCQ-10D enhanced course structure is organized using the Population, Intervention, Comparators, Outcomes framework [13].

Figure 1. Screenshot from video on hypothetical student completing MCQ-10D.



Population

Students in health professional education courses, for example, medical schools, subject to approval by the relevant bodies. (There are two approved pilot sites that are left unnamed in this publication).

Intervention

Textbox 1 presents the full details of the MCQ-10D instrument. The Web-based survey in which it is embedded is live [14]. A video of a hypothetical student completing the survey is included as an appendix in this article (see Multimedia Appendix 1) (Figure 1) (Some of the questions supplementary to the instrument would be modified to suit the particular institution and course.).

MCQ-10D is completed in two stages, reflecting the aim to impact on the educational student experience from its beginning and throughout. Immediately prior to, or at the very start of the course, the student completes part 1, where they indicate the importance they personally assign to the 10 course quality criteria, on the 0 to 10 scale. (At both this point of time, and again in part 2 at the end of the course, they can indicate whether they had serious difficulty understanding any of the criteria and can leave comments on them.)

Students will be automatically reminded of the criteria at appropriate intervals (by email or announcements on their learning platform), for example, monthly, in courses lasting 8 weeks or more. In long courses, interim ratings may be appropriate, but these are not currently envisaged.

At the conclusion of the course, the student completes the lengthier part 2 of the assignment. In this, they provide their overall holistic assessments of course quality and satisfaction with it, followed by their ratings of the course on the MCQ-10D criteria, rephrased in the past tense.

Immediately after entering their ratings, students are presented with their MCQ-10D score in the Annalisa screen, which also displays the component ratings and weightings [15]. The score is the result of multiplying their ratings by their original weightings (normalized to add to 100%) and summing across all ten criteria. The student then has the opportunity to revise their weights, if they feel they are now different from the original ones they supplied (now visible to them), and thereby obtain a revised MCQ score. Next, they are able to see the partial contribution each criterion makes to the overall MCQ score, which will indicate to the providers the student's views as to the possible sources of improved course quality. Note that, for each individual student, these will reflect his or her personalized weightings, as well as ratings. Finally, students are asked to reflect on whether explicit attention to course quality criteria via MCQ-10D has had an effect on their experience of the course, and to respond to other questions of a comparative nature. These questions are not part of the instrument and will necessarily vary with the course and its institutional setting. Those included on the Internet version represent one possibility.

It should be stressed that MCQ-10D can be implemented in many software programs, including macro-enhanced spreadsheets (eg, Excel or open source equivalents). Annalisa is an implementation of Multi-Criteria Decision Analysis, or, as in this use, Multi-Attribute Value Theory, and is simply one piece of software that facilitates the dynamic, interactive reweighting we regard as a key feature of the instrument.

From the outset, students are aware that MCQ-10D is a part of the assignment work for the course, with 10% of the course marks awarded for completion of both parts, the second of which is completed after they are aware of the marks they have received for the other 90% of the assignment work. They can therefore predict their grade with certainty before completing, or not completing, part 2 of the MCQ-10D assignment.

Textbox 1. MCQ-10D, with Internet heading and popup text (line 1) and Weighting and Rating questions (lines 2 and 3) for each dimension.

CONTENT: scope of coverage and level of treatment
How important to you is it that the course delivers the specified content at the level prescribed?
To what extent do you think the course delivered the specified content at the level prescribed?
ORGANIZATION: clear structure and coherent progression
How important to you is it that the course is well organized and offers a clear structure and coherent progression?
To what extent did you find the course well organized and offered a clear structure and coherent progression?
PERSPECTIVE: explicit and offering alternative views where appropriate
How important to you is it that the course's perspective/theory is explicit, and, where appropriate, it offers alternative views?
To what extent did you find the course’s perspective/theory was explicit, and, where appropriate, it offered alternative views?
PRESENTATIONS: relevantly informative, engaging, and stimulating
How important to you is it that the presentations are relevantly informative, engaging, and stimulating?
To what extent did you find the presentations relevantly informative, engaging, and stimulating?
MATERIALS: relevantly informative, engaging, and stimulating
How important to you is it that the learning materials are relevantly informative, engaging, and stimulating?
To what extent did you find the learning materials relevantly informative, engaging, and stimulating?
RELEVANCE: to real world decision/policy making, practice, or behavior
How important to you is it that the course demonstrates its relevance to real world decision/policy making, practice, or behavior?
To what extent did you find the course demonstrated its relevance to real world decision/policy making, practice, or behavior?
WORKLOAD: appropriate to credit level and flexible
How important to you is it that the mandatory workload is in line with the credit award and is flexible as specified?
To what extent do you think the mandatory workload was in line with the credit award and exhibited the specified flexibility?
SUPPORT: from teaching and other relevant staff
How important to you is it that the support and feedback from teaching and other staff (in line with that offered) is respectful and responsive?
To what extent did you find the support and feedback from teachers and other staff (in line with that specified) was respectful and responsive?
INTERACTION: with other students
How important to you is it that the course provides and promotes the specified facilities for interaction with other students?
To what extent did you find the course provided and promoted the possibilities for interaction with other students that were offered?
ASSESSMENT: assignment requirements clear and mine graded fairly
How important to you is it that the assignment requirements are clear and your assignments are graded fairly by them?
To what extent did you find the assignment requirements were clear and your assignments were graded fairly by them?

Comparators

Student reaction to the intervention will be gauged by responses to questions asking for their comparisons with the feedback system they conventionally experience. Also elicited will be their perceptions regarding the comparative effect of the intervention on their own educational experience, including the comparative quality and clarity of the opening course description.

No control group is envisaged, as it would be impractical, unethical, and possibly illegal. However, the group level results from MCQ-10D will be compared with the results from the standard feedback form that students are asked to complete anonymously in the institutions concerned.

Provider reactions to the intervention will be sought in a separate post course questionnaire, and interview/s which will involve requesting comparisons with their typical preparation of course descriptions, materials and presentations, their delivery of courses, and their perceptions of student performance and engagement.

Outcomes

Student reactions to the experience are as specified under the subsection Comparators, immediately above. The MCQ score could be interpreted as a Student-Reported Outcome Measure or Student-Reported Experience Measure, analogous to a Patient-Reported Outcome Measure or Patient-Reported Experience Measure [16,17].

Provider/faculty reactions to intervention are as specified under the subsection Comparators, immediately above.

Results

Initial piloting will occur in two courses during 2015, one in Australia and one in Denmark, with outcome results available by end of the year. However, other courses may be added on request.

Discussion

Student Course Assessment as Graded Assignment

In certification settings, such as medical schools, experience shows that a task will rarely be undertaken if it is optional and does not count substantively to the course award. In many cases, simple (weighted, but ungraded) task completion will be an appropriate and sufficient requirement, as it will be in the case of MCQ-10D. It will effectively be a mandatory part of the assigned work, given a small, but finite weight (10%) in the final grade. Its satisfactory completion, defined purely technically, will add 10% to the student's final mark. The actual course grade the student will receive is therefore predictable with certainty *before* the rating part of MCQ-10D is completed, or not.

If it is to be taken seriously, it is important that a course assessment instrument relates to the course as described in the rubric available to the student before enrollment (if it is an optional elective) or, at latest, at its commencement (if it is mandatory). The MCQ-10D instrument takes it for granted that the course has been designed to increase the person's degree of competency in relation to "knowing that", or "knowing why", or "knowing how", or some combination of these. The content in terms of facts, principles, ideas, concepts, theories and techniques to be covered, the levels and depths at which they are to be (or can be) studied, the broad ways they will be presented and can be engaged with, the type/s of individual support and group interaction on offer, and the way/s competency will be assessed for certification purposes, are all to be spelled out explicitly in the course description. Secondary outcomes of the intervention are likely to be an improved quality of course preparation and greater precision and clarity in relation to the course's aims and delivery methods, as well as wider potential benefits in curriculum development.

There is no provision *in the instrument* itself for the student to say they would have preferred the course to have been different from that offered. For example, to have some face-to-face sessions in a course clearly stated to be purely Internet, for basic material to be provided in what is clearly stated to be an advanced course, or for an "unflipped" course instead of the advertised "flipped" one. However, there is space in the survey, within which MCQ-10D is embedded, for this sort of comment, clearly differentiated and separated. We assume that alternative routes are available for forwarding such suggestions of changes to the course curriculum or rubric, some of which may involve increased resources being made available to the unit providers.

Students, like patients, are primarily persons, and should be treated as such. However, there is a central difference from medical or other health professional practice, in that the student is typically seeking certification from the provider for use in subsequent career situations. They are, in fact, purchasing the service which leads to that qualification, be it on a single unit of study or continuing development, or a complete award such as a degree, as well as gaining wider and noninstrumental benefits. "Person-centeredness" remains a key principle, but is necessarily different in the certification situation from that in a pure learning situation, since the awarding body has a duty of care beyond the individual. The resulting power relationship needs to be acknowledged throughout education, and especially in the seeking of feedback. In our proposal, the sequence of events ensures that the content of the student's course assessment can have little influence on the grade awarded. Final submission of course ratings is essential to maximize marks, but can only occur after the student's grade is predictable with certainty, because they know their marks for all their graded assignments.

As with all other aspects of the course, the student is made aware of this assignment requirement and consents to it by enrolling.

MCQ is explicitly designed for formative use at the course level. Appropriately interpreted, it could serve as one component of a multi-criterial summative assessment for other purposes, but introducing a dually personalized measure of quality as an integral part of the course will pose major challenges for those who seek aggregated "feedback" at unit, program, or higher levels.

What Makes this Approach Different?

The key, almost paradigmatic, difference from previous instruments cited at the beginning of the paper is the use of the student's importance weightings for the criteria. A second key difference is that the criteria presented are limited to ten as a matter of practicality, because of the need to make, or confirm, the explicit trade-offs among the criteria necessary in order to arrive at an overall index, and, hence, opinion as to the overall student-assessed quality of this course.

The individual student receives an immediate and personalized response to their course assessment as soon as their ratings are entered. This makes it somewhat rare among feedback instruments, which in most cases provide only delayed and aggregated information, if any.

Ideally, the instrument will also be completed by the course provider/s in the spirit of self-reflection and professional development. This would provide the basis of exploring dyadic concordances and discordances in an open manner at both overall and criterion-specific levels, and, hence, in relation to both course processes and course outcomes. Ultimately, only transparent discourse, taking place on a sound empirical basis and in a way that reflects student and staff heterogeneity, has the potential to deliver—as well as document digitally—person-centered education. However difficult it may be to implement an approach such as that represented by MCQ-10D within current systems, regulations, and resources, it represents the target to be aimed at from a long term and longitudinal perspective.

We have developed an Internet generic and preference-sensitive instrument for assessing course quality from the student perspective. It is intended to be practically useful for all parties

who are willing to treat quality assessment as an integral part of a course, instead of as a marginal, terminal, and optional add-on as feedback, the focus of all previous instruments. Work is needed to test the instrument in a range of settings to establish its own quality and genericness, and how willing students and providers are to treat quality assessment as a process that both represents and creates educational added value.

This paper is a protocol to establish its feasibility and acceptability, and act as proof of method at the technical and organizational levels. It is to be piloted initially in courses in a medical faculty and a school for health professionals. We invite other health education providers to join in this piloting, using our software, and will be pleased to collaborate in proposals to translate the Internet instrument into other languages.

Acknowledgments

The Region of Southern Denmark, the University of Southern Denmark, and The Health Foundation (Helsefonden) are funding MKK's PhD study. The contribution of GS was supported by the Screening and diagnostic Test Evaluation Program funded by the National Health and Medical Research Council of Australia under program grant number 633003.

Authors' Contributions

JD and MKK developed the concept of a dually personalized student-reported course quality assessment and produced the draft 10 item MCQ from a purposive survey of existing instruments for student feedback. JBN, JL, and GS commented on the instrument and proposed ways to implement it, drawing on the basis of their lengthy experience in developing, administering, and interpreting their institutions' conventional feedback systems. JD drafted the initial text, which was extensively revised in collaboration with MKK.

Conflicts of Interest

JD has a financial interest in the Annalisa software used in the current implementation of MCQ-10D on the University of Sydney server, but does not benefit from its use.

Multimedia Appendix 1

Video of hypothetical student completing MCQ-10D.

[MP4 File (MP4 Video), 4MB - resprot_v4i1e15_app1.mp4]

References

1. Spooren P, Brockx B, Mortelmans D. On the validity of student evaluation of teaching: The state of the art. Review of Educational Research 2013 Aug 20;83(4):598-642. [doi: 10.3102/0034654313496870]
2. Kaltoft M, Cunich M, Salkeld G, Dowie J. Assessing decision quality in patient-centred care requires a preference-sensitive measure. J Health Serv Res Policy 2014 Apr;19(2):110-117 [FREE Full text] [doi: 10.1177/1355819613511076] [Medline: 24335587]
3. Alderman L, Towers S, Bannah S. Student feedback systems in higher education: A focused literature review and environmental scan. Quality in Higher Education 2012 Oct 16;18(3):261-280. [doi: 10.1080/13538322.2012.730714]
4. Chalmers D. Student feedback in the Australian national and university context. In: Nair C, Mertova P, editors. Student Feedback: The Cornerstone to an Effective Quality Assurance System in Higher Education. Oxford: Chandos; 2011:81-97.
5. Coates H. Tools for effective student feedback. In: Nair C, Mertova P, editors. Student Feedback: The Cornerstone to an Effective Quality Assurance System in Higher Education. Oxford: Chandos; 2011:110-118.
6. Davies M, Hirschberg J, Lye J, Johnston C. A systematic analysis of quality of teaching surveys. Assessment & Evaluation in Higher Education 2010 Jan;35(1):83-96. [doi: 10.1080/02602930802565362]
7. Fontaine S, Wilkinson T, Frampton C. Focus Heal Prof Educ. 2006. The medical course experience questionnaire: Development and piloting of questions relevant to evaluation of medical programs URL: http://search.informit.com.au/documentSummary;res=IELHEA;dn=038135413188566 [accessed 2015-01-19] [WebCite Cache ID 6VhVYm0qy]
8. Kember D, Leung DY. Establishing the validity and reliability of course evaluation questionnaires. Assessment & Evaluation in Higher Education 2008 Aug;33(4):341-353. [doi: 10.1080/02602930701563070]
9. Marsh HW, Roche LA. Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. American Psychologist 1997;52(11):1187-1197. [doi: 10.1037/0003-066X.52.11.1187]
10. Ramsden P. A performance indicator of teaching quality in higher education: The course experience questionnaire. Studies in Higher Education 1991 Jan;16(2):129-150. [doi: 10.1080/03075079112331382944]
11. Richardson JTE. Instruments for obtaining student feedback: A review of the literature. Assessment & Evaluation in Higher Education 2005 Aug;30(4):387-415. [doi: 10.1080/02602930500099193]
12. Palmer L. Influence of students' global constructs of teaching effectiveness on summative evaluation. Educational Assessment 1998 Apr;5(2):111-125. [doi: 10.1207/s15326977ea0502_3]

13. Guyatt GH, Oxman AD, Kunz R, Atkins D, Brozek J, Vist G, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011 Apr;64(4):395-400. [doi: 10.1016/j.jclinepi.2010.09.012] [Medline: 21194891]
14. Elicia Home: Fill-In Survey: MyCourseQuality 2.1. URL: http://healthbook.health.usyd.edu.au/index.php?PageID=survey_respond&SurveyID=924 [WebCite Cache ID 6Vhk4vJpD]
15. Dowie J, Kjer Kaltoft M, Salkeld G, Cunich M. Towards generic online multicriteria decision support in patient-centred health care. Health Expect 2013 Aug 2 [FREE Full text] [doi: 10.1111/hex.12111] [Medline: 23910715]
16. Black N, Jenkinson C. Measuring patients' experiences and outcomes. BMJ 2009;339:b2495. [Medline: 19574317]
17. Hodson M, Andrew S, Michael Roberts C. Towards an understanding of PREMS and PROMS in COPD. Breathe 2013 Sep 01;9(5):358-364. [doi: 10.1183/20734735.006813]

Abbreviations

MCQ-10D: MyCourseQuality-10 Dimensions

Edited by G Eysenbach; submitted 18.11.14; peer-reviewed by J Richardson; comments to author 25.11.14; accepted 05.12.14; published 13.02.15

Please cite as:
Kaltoft MK, Nielsen JB, Salkeld G, Lander J, Dowie J
Bringing Feedback in From the Outback via a Generic and Preference-Sensitive Instrument for Course Quality Assessment
JMIR Res Protoc 2015;4(1):e15
URL: http://www.researchprotocols.org/2015/1/e15/
doi:10.2196/resprot.4012
PMID:25720558

Context Sensitive Health Informatics: Many Places, Many Users, Many Contexts, Many Uses
E.M. Borycki et al. (Eds.)

doi:10.3233/978-1-61499-574-6-172

Health Informatics Can Avoid Committing Symbolic Violence by Recognizing and Supporting Generic Decision-making Competencies

Mette Kjer KALTOFT[a], Jesper Bo NIELSEN[a], Glenn SALKELD[b], Jack DOWIE[c]
[a] *Department of Public Health, University of Southern Denmark*
[b] *School of Public Health, University of Sydney*
[c] *London School of Hygiene and Tropical Medicine*

Abstract. 'Symbolic violence' is committed, however well-intentionally, by the imposition of particular conceptualizations of what information, in what form and quality, is needed in order to make an 'informed choice' and hence – by questionable segue - a high quality decision. The social and cultural forms of relevant cognitive capital possessed by those who fail, because of their low general literacy, professionally-set knowledge tests of functional health literacy, are being ignored. Failing to recognise and exploit a particular form of *functional* decision literacy, in fact leads to symbolic violence being experienced by individuals at *any and all* levels of general literacy. It leads many to adopt the same range of avoidant and other undesirable strategies within healthcare situations observed in those of low basic literacy. The alternative response we propose exploits the alternative generic decision literacy which comes in the form of the ability to access and use the decision-relevant resources provided for many consumer services and products on comparison websites and magazines. The methodology is the simple form of multi-criteria analysis in which the products' ratings on multiple criteria are combined with criterion weights (supplied by the site) to produce scores and 'best buys' and 'good value for money' verdicts. Our alternative approach extends this approach to healthcare options and permits the incorporation of personal criterion weights in furtherance of person-centred care. Health informaticians, especially those in the decision support field, should build on this widespread generic competence. The fact that it is generic, far from implying context insensitivity, can be seen as a necessary basis for achieving context-sensitivity and sensitivisation at the level of the individual person as they experience a lifelong sequence of healthcare decisions.

Keywords. Informed choice; health literacy; person-centred care; empowerment

Introduction

A recent paper questions the focus on *functional* literacy in attempts to encourage and support the making of 'informed' healthcare choices [1]. Drawing on the work of Bourdieu, Adkins and Corus see 'symbolic violence' being committed, however well-intentionally, by the imposition of particular conceptualizations of what information, in what form and quality, is needed in order to make an 'informed choice' and hence – by questionable segue - a high quality decision. These conceptions are built into the definitions of health literacy by WHO and the EU and have major policy and resourcing

implications[2]. The social and cultural forms of capital possessed by those who fail, because of their low general literacy, to pass professionally-set knowledge tests of functional health literacy, are being ignored, say Adkins and Corus. These individuals are being characterised, however implicitly and politely, as having deficiencies that need eliminating or at least reducing. 'A substantial amount of research concludes low literate individuals are incapable of taking on the tasks associated with healthcare and such disempowering depictions of low literates propagate stereotypes and biases toward the undereducated and perpetuate disparities and gross inequities in healthcare services…Those who fall short of standard expectations experience denigration, leaving them with no command for social respect.' The experiences of symbolic violence create concerns of being ridiculed and these manifest themselves in avoidance and other strategies inimical to optimal healthcare decision making, producing consequences such as non-adherence.

In this paper we accept the validity of this argument, but move away from its concern with low general literacy to argue that failing to recognise and exploit a particular form of functional *decision* literacy, in fact leads to symbolic violence being experienced by individuals at *any and all* levels of general literacy. It leads many to adopt the same range of avoidant and other undesirable strategies within healthcare situations observed in those of low basic literacy. Our alternative response exploits that form of generic decision literacy. It offers support that does not imply that only an 'informed choice' can be a good decision, with 'being informed' defined professionally. It focuses on the vacuum left at the Point of Decision in the formal definitions.

The argument is most effectively made with reference to what we see as the current orthodoxy within the decision-aiding branch of health informatics. This orthodoxy is grounded in the IPDASi guidelines [3], but encompasses the specific interpretations in publications that proclaim their adherence to them. We can also endorse the conclusion of Joseph-Williams, Elwyn and Edwards, reviewing research into the patient experience, that knowledge is not power, and that information is not in itself empowering unless deployed (deployable) within a more equal clinical power relationship [4]. But we disagree with their assumption that knowledge in the conventional form is to be regarded as a necessary condition, albeit now one of two. We argue that supplying the information in a particular 'unconventional' form and integrating it with the best available estimates, will enable the patient to arrive at an informed decision, even if they know nothing about its content in the sense the orthodoxy seeks. Some patients will wish to engage in the orthodox way. We are concerned with those who will experience this requirement as symbolic violence, as a result of which they will adopt attitudes and behaviours not conducive to optimal health, self-defined. The relative numbers are not known, but may be large.

Our case for a generic approach may appear to endorse or encourage context-insensitivity. Almost the opposite. The argument is that a generic and widely available 'decision language' is essential if context-sensitivity is to be successfully achieved by the individual patient/person in their lifelong sequence of healthcare decisions. To seek to achieve context-sensitivity without such a generic grounding can lead to the detrimental consequences of the 'symbolic violence' inflicted when it is implied that every decision has to be treated on a one-off basis; that (e.g.) a prostate cancer screening decision has no connection with an atrial fibrillation treatment one; and that general decisional empowerment is not possible.

1. The Orthodox Approach to Decision Aiding and Evaluation of Decision Quality

We can make this point in a specific way by referring to the evaluation of the aids being produced by Karen Sepucha and colleagues. While these aids contain both knowledge and goals/values components, only the knowledge score is available at an individual level, since the values component of quality is addressed only *ex post*, at a group level, and in terms of the relationship between goals and eventual actions (group level concordance). The recent herniated disk decision aid study provides a good example of what is advanced as a decision quality instrument, but at the individual level reduces to a measure of the knowledge possessed by the patient - after administration of the aid [5]. This is naturally the knowledge in the aid necessary for the choice to be regarded as 'informed'. The mean knowledge score from the patients who viewed the decision aid was used to set a 55% threshold for 'informed'.

The argument is essentially circular, but the issue for us is not whether a patient's information is incorrect, while being perceived to be correct. The issue is whether showing that it is incorrect and attempting to correct the misperception by providing the correct information will constitute symbolic violence, without leading to a *better* decision, as opposed to (possibly) an 'informed decision' according to the orthodoxy.

It is important to make clear immediately that we are not arguing against this sort of condition-specific information being made available in a decision aid and making it available in the form it is usually provided. Indeed we are in favour of making it available on an opt-in basis, probably via links, and possibly even with some weak nudging towards consulting it. We embed our decision aid, based on Multi-Criteria Decision Analysis, (MCDA) in a wider program, MyDecisionSuite, which offers many opt-in customisation possibilities as well as the personalisation for the aid itself [6,7]. We are arguing against any implication that consulting information, retaining it, and attempting to synthesise it with personal preferences, are *necessary* conditions of a good decision, let alone the sufficient conditions implied by prominent decision quality measures.

In our alternative, information essential to a good decision *is* present in the aid, but it is present in a matrix of option performance rates on multiple criteria. This matrix format is familiar to all those possessing the generic decision literacy that enables them to engage with product and service comparison websites. Even then the information matrix is made available only on an opt-in basis, because we do not want to imply that consulting it, and processing it in a way usually referred to as 'making up one's mind', will lead to a better decision. We remain largely agnostic on that, in the same way we remain agnostic whether a decision informed in the orthodox way will produce a better decision – unless it is assessed by a tautologous outcome measure, that is, one using an individual's score on a knowledge/information test as the measure of decision quality. In order to avoid abdicating from the challenge of measuring decision quality within person-centred care we have offered MyDecisionQuality as a self-reported dually-personalised measure [8].

2. Recognising and Supporting Generic Decision Literacy

This generic decision literacy comes in the form of the ability to access and use the decision-relevant resources provided for many consumer services and products on comparison websites and magazines. The methodology on these sites is almost always the simple form of multi criteria/attribute analysis in which the product's ratings on

multiple criteria are combined with criterion weights (supplied by the site) to produce scores and 'best buys' and 'good value for money' verdicts. A large proportion of the population is familiar with this framework and language, its widespread commercial use and popularity of associated sites (e.g. comparethemeerkat.com) providing the most convincing evidence of this. Over 80% of consumers are reported to have consulted a comparison website in 2010, so the number is likely to be even higher now [9].

In Figure 1 (bottom panel) we enter the ratings for three anonymised free standing washer-dryers that appeared in a recent Which (UK) consumer magazine report on 16 such appliances. Five criteria were rated and weighted to arrive at the overall score. Price was listed separately and not weighted, leaving that trade-off to the consumer.

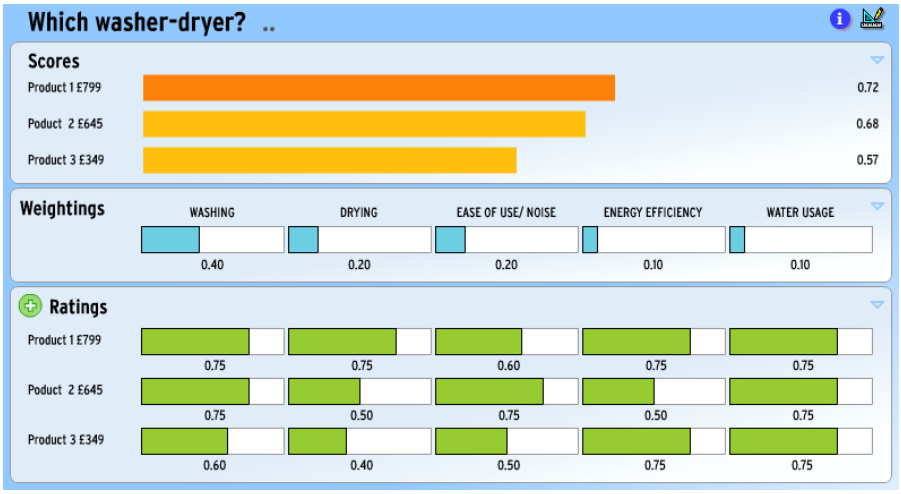


Figure 1. Ratings, Weightings, and Scores for three anonymised Washer-Dryers from a consumer magazine report re-presented in MCDA format

We do not endorse the particular framing (criterion selection and weightings) and use it only as an example of the sort of content presented in such comparative reports.

The Scores are the expected value of the Ratings and Weightings. Amid all the attempts to improve decision making and information communication, a central concept - expected value - has not received the attention needed even if the objective is to argue against it. We attribute this to the overarching reluctance to address the question of how information should be synthesised with preferences in any explicit way. Such an approach represents a form of reverse symbolic violence, implying that a proper person possesses high quality synthesising ability as an intuitive competence.

While these comparison sites increasingly include ratings and scores for medical devices and health products apps, they avoid evaluations of healthcare options that would involve weightings for criteria such as length of life. That is what our alternative approach, where the options become ones such as lifestyle change, medications and surgery and the attributes ones such an quantity and quality of life and treatment burden. While suggesting that health care decisions may be appropriately approached in the same way as buying a washer-dryer will be surprising if not appalling to some, there are three very good reasons for this extension to healthcare. (It is hopefully clear why the example must not be a healthcare one.)

Since it recognises and exploits a widely possessed type of generic literacy, the alternative not only has less potential to produce symbolic violence but simultaneously greater potential to empower the person. Such empowerment is a precondition of the person owning the decision (whether or not it is in some way shared), which increases the likelihood that the option decided upon will be adhered to subsequently. Whether there is greater concordance in relation to that chosen option is an open question. This will be determined by many things including the clinician's attitude and commitment to person-centred care, as well as quality of both the aid and the discourse surrounding it.

The orthodox approach cannot deliver person-centred care. In person-centred healthcare the relative importance of the considerations that matter to the person in their life is elicited and combined, at the point of decision, with the best estimates available on the performance of the available options on those criteria. This integration is performed in an explicit way which can be communicated to the person. Any prior comparative option evaluations, such as those that constitute the conventional 'evidence base' cannot be part of this process. The ethics of transparent person-centred care require 'evidence base' to be reconceptualised as the unsynthesised matrix of option performance rates for the person-important criteria mapped against the person's criterion preferences [10]. Our approach is therefore not only compatible with person-centred healthcare, it is actually the only way we can see transparent and direct decision support for it being delivered.

Emphasising the generic character of all healthcare decisions enables the individual to visualise any healthcare decision, whatever the condition (or set of conditions) in the same way, rather than it being implied that they need to know a lot about their breast or prostate cancer or whatever. They can then exploit their social and cultural capital which exists because their friends and contacts 'speak the same language' at a decision level. Irrespective of the biological specifics. And that generic competence extend through the life course, so that a sequence of decisions about contraception, birthing technique, and menopause management, as well as any morbidities that arise in the life course, can all be thought of and discussed socially within the same graphic structure.

Professionals already possess this generic decision literacy, so the task should be the simple one of recognising that it should be applied to their area of professional expertise, not just in their domestic life as a consumer. This does not mean writing off their other 'knowledge capital', but it does mean complementing it in order to engage with persons who do not possess it and are at risk of symbolic violence.

3. Reflections

While our focus is on the micro and meso levels, we can speculate about the wider systemic origins of the focus on this particular type of functional health literacy, rather than generic decision literacy. Among the most important macro origins would seem to be the demands for methodological rigour in studies used to justify policy level decisions with financial implications, such as on drug reimbursement or decision aid provision. The dually-personalised measures appropriate for person-centred care do not provide 'hard' criteria, able to be aggregated for groups. Possession, or not, of a proposed set of essential facts, especially about the improvements offered by a new drug or device, is eminently fit for purpose, *given this purpose*. But we question who should define what and how much information is important in person-centred care [11] and sug-

gest reconceptualising the person - previously known as patient [12] - as a researcher engaged in an n-of-1 study for optimal health behaviour choices [10].

Health informaticians interested in supporting person-centred decision making and care at all points in patient pathways, including health records and decision aids, need to acknowledge, accept, accommodate, and adopt MCDA-based approaches to transparently document, support, and evaluate healthcare decisions.

References

[1] N.R.Adkins, C. Corus. Health literacy for improved health outcomes: Effective capital in the marketplace. *Journal of Consumer Affaiirs* **43** (2009) 199–222.
[2] K.Sørensen, S. van den Brouke, J. Fullam, et al. Health literacy and public health: a systematic review and integration of definitions and models. *BMC Public Health* **12** (2012) 80.
[3] G. Elwyn, A.M. O'Connor, C. Bennett, et al. Assessing the quality of decision support technologies using the International Patient Decision Aid Standards instrument (IPDASi). *PLoS One* **4** (2009) e4705.
[4] N. Joseph-Williams, G. Elwyn, A. Edwards. Knowledge is not power for patients: a systematic review and thematic synthesis of patient-reported barriers and facilitators to shared decision making. *Patient Education and Counselling* **94** (2014) 291–309.
[5] K.R. Sepucha, S. Feibelmann, W.A. Abdu, et al. Psychometric evaluation of a decision quality instrument for treatment of lumbar herniated disc. *Spine* 37 (2012) 1609–16.
[6] J. Dowie, M.K. Kaltoft, G. Salkeld, M. Cunich. Towards generic online multicriteria decision support in patient-centred health care. *Health Expectations* (2013) online: 02.08.13 doi: 10.1111/hex.12111
[7] Ø. Eiring, L. Slaughter. An assessment of the potential for personalization in patient decision aids. *Electronic Healthcare Lecture Notes Institute for Computer Science Social Informatics and Telecommunication Engineering* **91** (2012) 51–7.
[8] M.K. Kaltoft, M. Cunich, G. Salkeld, J. Dowie. Assessing decision quality in patient-centred care requires a preference-sensitive measure. *Journal of Health Services Research and Policy* **19** (2014) 110–7 online first 12.12.13.
[9] eDigitalResearch. *Comparing comparison sites: a report for Consumer Focus*. Southampton; 2012.
[10] M.K. Kaltoft, J.B. Nielsen, G. Salkeld, J. Dowie. Increasing user involvement in health care and health research simultaneously: A proto-protocol for "Person-as-Researcher" and online decision support tools. *JMIR Research Protocols* **3** (2014) e61.
[11] M.K. Kaltoft, J.B Nielsen, G. Salkeld G, J. Dowie. Who should decide how much and what imformation is importaant in person-centred care. *Journal of Health Services Research and Policy* (2015) online first (doi:10.1177/1355819614567911).
[12] J.A.M. Kremer, M van der Eijk, J.W.M. Aarts, B.R. Bloem. The Individual Formerly Known As Patient, TIFKAP. *Minerva Medica* **102** (2011) 505.

Context Sensitive Health Informatics: Many Places, Many Users, Many Contexts, Many Uses
E.M. Borycki et al. (Eds.)

**doi:10.3233/978-1-61499-574-6-74*
**This document includes correction of referencing from the published version.*

Enhancing Healthcare Provider Feedback and Personal Health Literacy: Dual Use of a Decision Quality Measure

Mette Kjer KALTOFT[a11] Jesper Bo NIELSEN[a], Glenn SALKELD[b], Jack DOWIE[c]
[a] *Department of Public Health, University of Southern Denmark*
[b] *School of Public Health, University of Sydney*
[c] *London School of Hygiene and Tropical Medicine*

Abstract In this protocol for a pilot study we seek to establish the feasibility of using a web-based survey to simultaneously supply healthcare organisations and agencies with feedback on a key aspect of the care experience they provide and increase the generic health decision literacy of the individuals responding. The focus is on the person's involvement in decision making, an aspect of care which is seriously under-represented in current surveys if one adopts the perspective of person-centred care. By engaging with an instrument to assess decision quality the person can, in the one action, provide a retrospective evaluation of a past decision making experience in a specific provider context and enhance their competency in future decision making in any setting. We see this as an exercise in context-sensitive educational health informatics.

Keywords. Informed choice; health literacy; person-centred care; empowerment; patient experience surveys; Patient-Reported Outcome Measure

Introduction

Against the wider backdrops of the Aarhus convention (http://www.unece.org/env/pp/treatytext.html) and other efforts to promote individual, societal and environmental health there are significant moves to increase person and citizen involvement in the promotion of health and provision of healthcare services. They take two broad forms.

On the one hand are initiatives emanating from providers responsible for health services at a community or national level, seeking to gain more and better information and feedback from patients viewed collectively, as a whole or as members of subgroup. Anonymised feedback in the form of satisfaction surveys has been the traditional source and these are now becoming even more prominent, while undergoing the much-needed revisions that take advantage of web-based technologies and rapidly increasing access to the internet. Most bodies now accept that self-reported 'satisfaction' is not an appropriate concept and replace it with requests for reports on the person's experience of specified events or actions. In recent years these wider surveys have been accompanied by efforts to increase 'user involvement' in top-level organisational and

[1] Corresponding author. E-mail: mkaltoft@health.sdu.dk

research settings, representatives of patients or patient groups, or lay persons, being invited to the table. [1–3] Citizen juries, focus groups, and similar community-based arrangements, provide an intermediate mechanism, giving the possibility of deeper, if narrower, feedback than a survey, but remaining outside the responsible body.[4]

On the other hand are the initiatives that focus on the individual, seeing him or her as a person/patient seeking optimal health and healthcare within the existing system and organisational arrangements. These efforts have been initiated mainly by professional and academic groups, often in collaboration with patient organisations. Their aim is to provide better support to the person in the context of their personal health journey, some taking the form of information or decision aids, some mechanisms for emotional or social support.

There is clear overlap between the two and a few national organisations are now moving into the second area of personalised support through decision aids. However, the basic distinction remains valid and the following study protocol is based on the assumption that a connection can be made so that the individual can simultaneously contribute to the higher-level feedback process *and* benefit personally. This dual strategy is designed to minimise both cost and respondent fatigue and maximise the return to healthcare provider and person in relation to decision making quality.

The protocol focuses on decision making, because we see individual involvement in decisions as a central aspect of the quality of the person's care experience and a key indicator of any organisation's commitment to person-centred care. Using the MyDecisionQuality (MDQ) instrument we seek to show how the individual can, in one online survey, simultaneously contribute enhanced feedback to providers on past decisions and benefit personally from the increased generic health decision literacy that may improve the quality of their future health decisions.

1. Limitations of existing surveys

Surveys seeking patient feedback or assessments of patient experience typically suffer from at least three limitations from the perspective of person-centred care.

First, they are typically confined to eliciting ratings on a number of indicators. If these are weighted to produce an overall index, rather than left as a profile, the weights are supplied by the instrument developers. They are quite often simple equal weights as in the Patient Experience Questionnaire (PEQ) [5] subsequently cluster-analysed in Bjerknaes. [6] Only those built within the Dutch Consumer Quality Index (CQI) framework incorporate patient weightings into the assessment. [7] The condition-specific CQI instrument is [8] in fact two instruments. CQI Experience elicits ratings on each item. CQI Importance elicits importance weightings for each item, both on four point Likert scales. The percentage of respondents giving the lowest experience rating to an indicator is multiplied by the percentage giving it the highest weighting to produce a Quality Improvement Score for use in prioritisation. These are clearly group level results and we learn nothing about the individual level relationship between experience and importance.

Second, surveys underemphasise the person's participation in decision making. Remarkably neither the PEQ nor Bjerknaes paper contains the words 'decision' or 'preference'. The defence that this may not emerge from literature reviews or patient focus groups is not convincing. It is the product of long socialisation into the largely passive and disempowered status as a patient of a provider, a patient who is to be 'informed', 'communicated with', 'have things explained clearly', 'listened to attentively', 'treated with respect', 'taken seriously', etc.

The third limitation involves the restriction to patients' treatment experience within an illness care context and provider facility. This means omitting invitations issued to persons regarding screening, vaccination and other preventive actions. Our protocol, which involves dissemination to community residents as well as patients, rectifies this.

The protocol has been developed initially for the Danish context, where we already observe large scale and successful efforts in making Patient-Reported Outcome Measures the centre of an integrated electronic system [9]. But we see this Danish study as just one instantiation of a higher level 'proto protocol', adaptable and sensitive to other countries and settings, through translation to the professional, legal and ethical circumstances in the jurisdiction. In the Danish piloting we will offer both Danish and English versions of the DQ4ALL survey, embedding the MDQ instrument.

2. Objectives

To explore the feasibility and acceptability of the MDQ instrument to persons in the community to (i) provide feedback to providers on self-rated dually-personalised decision quality as a key aspect of the person's health and healthcare experience, and (ii) increase the health decision literacy of the person in relation to evaluating past decisions and preparing for future ones.

.

3. Methods

The DQ4ALL is a randomised survey with two arms one of which includes MDQ. The randomization occurs at the point of access to the anonymous survey. Both arms elicit year of birth, sex and health status measure (EQ-5D) before responding to the Control Preferences Scale [10] and to recall one healthcare decision, taken in any setting (primary/secondary/community). They are then asked when this recalled decision happened (4 ranges), and whether it was about testing/screening), treatment (initiation, change, discontinuation), rehabilitation, or prevention (e.g. vaccination, lifestyle/behaviour change). At this point, they respond to the Satisfaction With Decision instrument [11] and the Control Preference Scale, both modified to apply to the recalled decision.

3.1 MyDecisionQuality (MDQ)

The MDQ instrument is then responded to in respect of the recalled decision.
MDQ is a dually-personalised instrument based on Multi-Criteria Decision Analysis. [12] MDQ is generic in the sense that the criteria are phrased without reference to any particular decision or context. Information relating to the specific decision, must be provided outside the MDQ instrument, such as in the wider condition-decision support system in which MDQ will often be situated. [13]
The Ratings items for MyDecisionQuality appear below. (The Weightings are phrased as the importance of each criterion. Both are elicited on a 0 to 10 scale.)
OPTIONS: I was clear about the possible options for me and what they involve;
EFFECTS: I was clear about the possible effects and outcomes of the options for me;
IMPORTANCE: I was clear about the relative importance of the different effects and outcomes for me;
CHANCES: I was clear about the chances of the different effects and outcomes happening to me, including the uncertainties surrounding the best estimates;
TRUST: I trusted the information I have been given is the best possible;

SUPPORT: I was satisfied with the level of support and consideration I received throughout the decision process, especially in regard to communicating at my level;
CONTROL: I felt in control of the decision to the extent I wish.
COMMITMENT: I was committed to acting on the decision

As with all implementations of the simple 'weighted-sum' version of MCDA, MDQ combines a set of importance weights for multiple criteria with performance ratings for each option on these criteria, and calculates the overall score as the expected value of eight criteria of decision quality. The MDQ Score, unique to the person and to the particular occasion, is shown with the partial contributions of each criterion to it displayed in segments; its weighting and rating are highlighted when the segment is touched or cursor is rolled over it. The resulting visual picture appears in Figure 1.

Figure 1 MDQ screen (in Annalisa implementation [12]) showing 8 criteria, Weightings, Ratings, and Score, with Score breakdown by criterion.

The respondent is also provided with insight into the priorities for future quality improvement by being shown the quality gains possible from improved rating on each criterion, weightings unchanged. For example, in figure 1 we can inform the person of the effect on their decision quality score of improving their rating on Importance, lowly rated at 0.3, given the relatively high weight of 0.188 they have assigned it. Achieving perfect rating on this criterion would increase their score by 0.7 x 0.188 or 0.132, equivalent to a 20% improvement. Feeding back the result of the same calculation for each of the criteria generates a personalised list of future priorities for decision making.

MDQ has been used as the primary outcome in a trial of two decision aids for the PSA screening decision in Australia [14]. Most relevantly here, the initial Danish version of the survey underwent some limited pre-piloting through a patient organization and medical department.

We will approach the Danish Knowledge Center for User Involvement in Health Care (ViBIS) to achieve a wide distribution of the survey among the residents of Denmark, including migrants.

3.2 Ethics

Since the survey is being distributed to persons in the community rather than patients, consent is by opting into its completion, and all data is anonymous, we expect no ethics approval will be required. Respondents will be able to give meta-consent to being approached in relation to this research by providing an e-mail address.

3.3 Health decision literacy

A final set of questions in DQ4ALL seek to determine whether completing it in relation to a recalled decision has helped evaluate or reevaluate that decision, and increased their perceived ability to enter into future decision making processes more fully and competently. In other words we seek to establish whether their perceived health decision literacy has been enhanced, by an implicit nudge of how to think proactively and more slowly. We do this by administering a subset of 6 items of the Preparation for Decision Making Scale relevant to this generic setting [15].

Health decision literacy is a wider and more diffuse concept than Decision Making Competence, though it can be seen as a background contributing factor. It has been the subject of extensive theorisation and measurement, notably by Fischhoff and colleagues. [16] They see it as a multidimensional construct, but show it is capable of being differentiated from general cognitive ability.

4. Analysis and Results

For feedback to provider purposes a range of descriptive statistics relating to the rating, weighting and scores for MDQ will be produced at group and subgroup level. These will be subjected to latent class analysis to determine the existence of preference-based clusters. Both the individual and clustered results will be regressed on sociodemographic and other characteristics, including type and location of the recalled decision, as part of a hypothesis generation, not hypothesis testing, process.

To assess the impact on perceived effect on generic health decision literacy we compare the responses to the subset of items of the preparation for decision making scale.

For those who have experienced the MDQ arm there will be further analysis of the perceived usefulness of the MDQ score and prioritisation suggestions.

Since all the responses are online, web-logging will enable analysis of the time spent on individual pages of the survey, as well as total time spent. This data will supply additional variables for analysis in both the feedback and literacy contexts.

5. Conclusion

In this pilot study we seek to establish the feasibility of using a web-based survey to simultaneously supply healthcare organisations and agencies with feedback on a key aspect of the care experience they provide, and increase the generic health decision literacy of the individuals responding. The focus is on the person's involvement in decision making, an aspect of care which is under-represented in current surveys from the perspective of person-centred care. By engaging with an instrument to assess decision quality the person can, in the one action, provide a retrospective evaluation of a past decision making experience in a specific provider context and enhance their competency in relation to future decision making in any provider setting. We seek to combine organisational and educational health informatics in a context-sensitive way.

Acknowledgments

Mette Kjer Kaltoft's PhD study is funded by the Region of Southern Denmark, The Health Foundation, and the University of Southern Denmark.

References

[1] Barber R. Exploring the meaning and impact of public involvement in health research. University of Sheffield PhD e-thesis; (2014).

[2] Boote J, Wong R, Booth A. "Talking the talk or walking the walk?" A bibliometric review of the literature on public involvement in health research published between 1995 and 2009. Health Expectations **4** (2012), 1–14.

[3] Kaltoft MK, Nielsen JB, Salkeld G, Dowie J. Increasing user involvement in health care and health research simultaneously: A proto-protocol for "Person-as-Researcher" and online decision support tools. JMIR Research Protocols **3** (2014).

[4] Mooney G. Communitarian claims and community capabilities: furthering priority setting? Social Science and Medicine **60** (2005), 247–55.

[5] Pettersen KI, Veenstra M, Guldvog B, Kolstad A. The Patient Experiences Questionnaire: development, validity and reliability. International Journal for Quality in Health Care **16** (2004), 453–63.

[6] Bjertnaes O, Skudal KE, Iversen HH. Classification of patients based on their evaluation of hospital outcomes: cluster analysis following a national survey in Norway. BMC Health Services Research **13** (2013), 1:73.

[7] Delnoij DMJ, Rademakers JJDJM, Groenewegen PP. The Dutch consumer quality index: an example of stakeholder involvement in indicator development. BMC Health Services Research **10** (2010), 88.

[8] Van Der Veer SN, Jager KJ, Visserman E, et al. Development and validation of the Consumer Quality index instrument to measure the experience and priority of chronic dialysis patients. Nephrology Dialysis Transplantation **27** (2012), 284–91.

[9] Hjollund NHI, Larsen LP, Biering K,et al. Use of patient-reported outcome (PRO) measures at group and patient levels: Experiences from the generic integrated PRO system, WestChronic. Interactive Journal of Medical Research **3** (2014), e5.

[10] Deber RB, Kraetschmer N, Irvine J. What Role Do Patients Wish to Play in Treatment Decision Making? Archives of Internal Medicine **156** (1996),1414-20.

[11] Holmes-Rovner M, Kroll J, Schmitt N,et al. Patient satisfaction with health care decisions: The Satisfaction with Decision scale. Medical Decision Making **16** (1996),1:58-64.

[12] Kaltoft MK, Cunich M, Salkeld G, Dowie J. Assessing decision quality in patient-centred care requires a preference-sensitive measure. Journal of Health Services Research and Policy **19** (2014), 110–17 online first 12.12.13.

[13] Kaltoft MK. Nursing Informatics AND Nursing Ethics: addressing their disconnect through an enhanced TIGER vision. Studies in Health Technology and Informatics **192** (2013), 879–83.

[14] Cunich M, Salkeld G, Dowie J, Henderson J, et al. Integrating evidence and individual preferences using a web-based Multi-Criteria Decision Analytic tool: An application to Prostate Cancer screening. Patient **4** (2011), 1–10.

[15] Bennett C, Graham ID, Kristjansson E et al. Validation of a Preparation for Decision Making scale. Patient Education and Counseling **78** (2010), 130-133.

[16] Parker, AM, & Fischhoff, B. (2005). Decision-making competence : External validation through an individual-differences approach. Journal of Behavioral Decision Making **18** (2005), 1–27.

***Note.**

This document includes correction of referencing from the published version.