ORIGINAL RESEARCH ARTICLE

Patient 2012; 5 (4): 1-13
1178-1653/12/0004-0001/$49.95/0

Using the Analytic Hierarchy Process to Elicit Patient Preferences

Prioritizing Multiple Outcome Measures of Antidepressant Drug Treatment

Marjan J.M. Hummel,[1] *Fabian Volz,*[2] *Jeannette G. van Manen,*[1] *Marion Danner,*[2] *Charalabos-Markos Dintsios,*[2] *Maarten J. IJzerman*[1] and *Andreas Gerber*[2]

1 Department of Health Technology and Services Research, University of Twente, Enschede, the Netherlands
2 Department of Health Economics, Institute for Quality and Efficiency in Health Care (IQWiG; Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen), Cologne, Germany

Abstract

Background and Objective: In health technology assessment, the evidence obtained from clinical trials regarding multiple clinical outcomes is used to support reimbursement claims. At present, the relevance of these outcome measures for patients is, however, not systematically assessed, and judgments on their relevance may differ among patients and healthcare professionals. The analytic hierarchy process (AHP) is a technique for multi-criteria decision analysis that can be used for preference elicitation. In the present study, we explored the value of using the AHP to prioritize the relevance of outcome measures for major depression by patients, psychiatrists and psychotherapists, and to elicit preferences for alternative healthcare interventions regarding this weighted set of outcome measures.

Methods: Supported by the pairwise comparison technique of the AHP, a patient group and an expert group of psychiatrists and psychotherapists discussed and estimated the priorities of the clinical outcome measures of antidepressant treatment. These outcome measures included remission of depression, response to drug treatment, no relapse, (serious) adverse events, social function, no anxiety, no pain, and cognitive function. Clinical evidence on the outcomes of three antidepressants regarding these outcome measures was derived from a previous benefit assessment by the Institute for Quality and Efficiency in Health Care (IQWiG; Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen).

Results: The most important outcome measures according to the patients were, in order of decreasing importance: response to drug treatment, cognitive function, social function, no anxiety, remission, and no relapse. The patients and the experts showed some remarkable differences regarding the relative importance of response (weight patients = 0.37; weight experts = 0.05) and remission (weight patients = 0.09; weight experts = 0.40); however, both experts and patients

agreed upon the list of the six most important measures, with experts only adding one additional outcome measure.

Conclusions: The AHP can easily be used to elicit patient preferences and the study has demonstrated differences between patients and experts. The AHP is useful for policy makers in combining multiple clinical outcomes of healthcare interventions grounded in randomized controlled trials in an overall health economic evaluation. This may be particularly relevant in cases where different outcome measures lead to conflicting results about the best alternative to reimburse. Alternatively, AHP may also support researchers in selecting (primary) outcome measures with the highest relevance.

Key points for decision makers

- The identification of relevant outcome measures in the treatment of major depression can be supported by multi-criteria decision analysis
- It is feasible to estimate the overall added value of a drug regarding various outcome measures using the analytic hierarchy process (AHP)
- Prioritization of outcome measures differs between patients and professionals; participants in the AHP need to be carefully selected to be relevant to the decision problem

Introduction

Health technology assessment (HTA) examines healthcare technologies from multiple perspectives to support implementation and reimbursement decisions. Nowadays, the patient perspective in HTA is increasingly considered essential to gain, for instance, a better understanding of what outcomes are deemed important to patients.[1,2] Investigators of clinical trials usually select primary and secondary outcome measures based on the expected clinical effects as judged by healthcare professionals. It can, however, be questioned whether the judgments of healthcare professionals reflect patient preferences. The relevance of the diverse outcome measures to patients is often not systematically assessed in the clinical studies that may contribute data on the effects of the technologies to health economic evaluations.

Several techniques can be used to weigh the relevance of multiple outcome measures. One is the analytic hierarchy process (AHP), a frequently used technique for multi-criteria decision analysis (MCDA).[3,4] The AHP works well to support a variety of healthcare decisions.[5-7] It has been used frequently to assist patients and healthcare professionals in understanding and making complex decisions.[8,9] The AHP guides these participants to prioritize multiple, even competing, clinical outcome measures. If conducted in a group setting, the AHP can support the efforts of panel members to share information about their beliefs, attitudes, and knowledge underlying the priorities they are to assign to the outcome measures.[10] HTA agencies are exploring the use of MCDA in health economic evaluations. In its search for adequate techniques to include patient preferences in their health economic evaluations, the German Institute for Quality and Efficiency in Health Care (IQWiG; Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen) was interested in an application of the AHP in a group setting.

A relevant clinical area in which the AHP could be applied is the treatment of major depression, which has a lifetime prevalence of 13–17% in Western countries[11,12] and can have a severe impact on quality of life.[13,14] In moderate to severe cases of major depression, pharmaceutical agents

are indicated either alone or in combination with cognitive therapy.[15] Venlafaxine, duloxetine, (both serotonin-norepinephrine reuptake inhibitors [SNRIs]), bupropion, and mirtazapine (both atypical antidepressants) are relatively new antidepressants. In multiple clinical trials, the effects of these antidepressants on patients have been scrutinized with respect to various outcome measures: response, remission, relapse, diverse accompanying effects, and (serious) adverse events.[16-20] However, the relevance of these measures to patients is yet to be investigated.

Previous studies on patient preferences in the area of major depression compared counseling, treatment with pharmaceutical agents, and no treatment.[21-23] Outcome measures of relevance to patients are not only related to the effectiveness and adverse effects of the treatments but also to stigma, social support, the relationship with healthcare providers, possible addiction to medication, and the length of time needed to complete the treatment.[21,22] These preferences are likely to vary by gender, ethnicity, income, knowledge about treatments, prior recent treatments, and co-morbidities.[23] These previous studies provide more insight into patient preferences to compare cognitive therapy, treatment with pharmaceutical agents, or no treatment of major depression. However, the relevance of all clinical outcome measures remains unclear when comparing solely pharmaceutical agents for those patients that prefer antidepressant treatment with pharmaceutical agents.

This study was undertaken to explore the value of the AHP in health economic assessments and its contribution to assist reimbursement decisions. The research questions were as follows:

1. What are the priorities and weights assigned by the patients and experts to the outcome measures of antidepressant treatment with pharmaceutical agents?
2. If there are differences between the two panels, how could they be explained?
3. How can the priorities and weights of the patient panel contribute to health economic assessments of, in this case, antidepressant treatment with pharmaceutical agents?

The preliminary results on the prioritization of the outcome measures have been described by Danner et al.[24] This study adds to these results by (i) strengthening the analysis by exploring the consistency among judgments of individual group members in depth; (ii) elaborating on the research methodology, and the differences in judgments among patients and the psychiatrists and psychotherapists; and (iii) adding the comparison of the performance of three antidepressants on this weighted set of outcome measures. Accordingly, this study presents a full MCDA to elicit preferences for three alternative drug treatments of major depression.

Methods

Analytic Hierarchy Process Procedures

Decision Structure

As a first step, the AHP structures a decision task into a hierarchy of factors, including the objective, criteria, sub-criteria, and alternatives. The objective was to prioritize the outcome measures of antidepressant treatment according to their relevance to patients. The sub-criteria were the outcome measures or clinical endpoints about which clinical evidence had been collected in clinical trials. The main criteria were categories of conceptually related outcome measures. These categories were defined by the researchers. We included three antidepressant drugs or drug classes about which the most evidence was available, namely duloxetine, venlafaxine, and selective serotonin reuptake inhibitors (SSRIs).

Selecting Outcome Measures

Existing outcome measures were identified from the IQWiG benefit assessment of antidepressants.[25,26] In this study, the outcome measures were selected by the researchers according to the following rationale:

- the overall availability of clinical evidence on a specific outcome measure of the antidepressants;
- statistically significant differences in outcomes between antidepressants and placebo or another antidepressant;
- other outcome measures potentially relevant to patients that were reported in the literature.

The main methodological constraint was that outcome measures needed to be mutually exclusive,

clear, comprehensive, and of importance within the same order of magnitude.

Table I provides the definitions of the (categories of) outcome measures selected.

Pairwise Comparisons

In the next step, the AHP offers a pairwise comparison approach to estimate the weights of the outcome measures. Each group member judges how important an outcome measure is compared with another outcome measure. Even though more research on the most appropriate scale to be used in HTA would be beneficial,[27] we applied a double nine-point scale, which is the most widely used AHP scale, to score this judgment. The scale ranges from 1, reflecting equal importance of the outcome measures, to 9, reflecting extremely greater importance for one of the two outcome measures. Similarly the performance of the alternatives can be compared pairwise on a nine-point scale that ranges from an equal performance up to an extremely higher performance. This scale can then be used to make subjective estimations of the relative performance of the antidepressants. Alternatively, as has been conducted in this study, the relative performance of the alternatives can be rated on the basis of clinical evidence.

An example of a pairwise comparison is shown in figure 1.

Consistency

For each matrix of pairwise comparisons, the AHP provides a measure of consistency to show if each pairwise comparison is logically sound with regard to the remainder of the comparisons.[3]

Table I. Outcome measures included

Outcome measure	Definition
1. Efficacy	The desired effect on depressive symptoms. These depressive symptoms include: depressed mood, loss of interest in activities, insomnia or hypersomnia (sleeping problems), lack of energy, change in weight and/or appetite, loss/gain or increased/decreased appetite, feelings of worthlessness or guilt, psychomotor agitation or inhibition/retardation, decreased concentration and/or suicidal thoughts
a. Response	Decrease in the quantity/quality of depressive symptoms by at least 50% as measured on a depression scale (e.g. HAM-D, MADRS)
b. Remission	Decrease in depressive symptoms to a degree that the patient no longer fulfills criteria of a depressive episode as measured on a depression scale (e.g. HAM-D, MADRS)
c. No relapse	Depressive symptoms do not increase to 'depressive episode levels' as measured on a depression scale (CGI-S) and on the basis of a diagnosis according to the DSM-IV within 6–12 months of remission
2. Serious adverse events	These undesired side effects are life threatening, lead to permanent/severe disability or death, or require hospitalization (increase the length of a hospitalization)
a. Suicide and attempted suicide	This serious adverse event focuses on suicidality that is evoked by antidepressants. It does not focus on the increased risk to commit suicide due to the major depression itself
b. Other serious adverse events	Each adverse event can become a serious adverse event if it leads to one of the following: death/threat to life, permanent/severe disability, or hospitalization
3. Adverse events	These undesired side effects are non-life threatening or do not require hospitalization
a. Sexual dysfunction	Loss of interest in sexual activities and/or limitations in sexual functioning due to antidepressants
b. Other adverse events	Examples of other adverse events are hypertension, restlessness, sedation, dizziness, nausea, dry mouth, sweating, increase in weight
4. Accompanying effects on quality of life	Effects of treatment on patient-relevant outcomes that go beyond the effects on the depressive symptoms
a. Social function	Taking part in work or school, social life, leisure, and home life. The antidepressants improve social function in the case of short-term acute therapy
b. No anxiety	Not feeling (and/or behaving in a) fearful, anxious, or tense (manner)
c. No pain	Not feeling pain, such as a headache
d. Cognitive function	Ability to concentrate, think logically, and perform routine intellectual tasks

CGI-S = Clinical Global Impression-Severity scale; **DSM-IV** = *Diagnostic and Statistical Manual of Mental Disorders* (4th Edition); **HAM-D** = Hamilton Rating Scale for Depression; **MADRS** = Montgomery-Åsberg Depression Rating Scale.

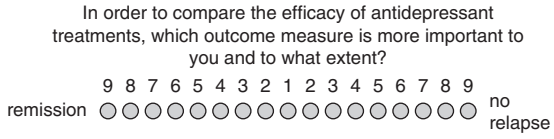

Fig. 1. Example of a pairwise comparison. For example, a 9 on the side of no relapse means that no relapse is thought to be extremely more important than remission.

The consistency ratio (CR) indicates how strongly the pairwise judgments resemble a purely random set of pairwise comparisons. Saaty's rule of thumb is that the pairwise judgments are allowed to have 10% of the inconsistency of a random set of pairwise comparisons.[3] This implies that a value of the CR ≤0.1 can be considered acceptable, and generates plausible outcomes.[3] Judgments that have a CR lower than 0.1 are reasonable, lower than 0.2 is tolerable, and higher than 0.2 should be revised or discarded.[28] In case of higher inconsistency, the decision makers are urged to check for accidental mistakes and to reconsider their pairwise comparisons, until the consistency measure is below the threshold indicated.[29]

Weights and Priorities

In cases of acceptable degrees of inconsistency, weighting factors and performance priorities are calculated. The principal right eigenvector approach is recommended by Saaty.[3] This eigenvector method can be interpreted as being a simple averaging process by which the final weights are the average of all possible ways of comparing the importance of the criteria. In cases of an acceptable degree of inconsistency, the weight factors assigned to the outcome measures are plausible in representing the relevance of each of these outcome measures to the patients. The performance priorities are plausible in representing how well the antidepressants perform on each outcome measure. The overall performance priority of an antidepressant is the weighted average of the performance priorities on all outcome measures.

Group Average

When calculating a group average to reflect the opinion of the group as a whole, the use of the geometric mean of all pairwise comparisons is recommended.[30]

Weights and CRs are then calculated for the group as a whole. If the differences between weights assigned by the individual group members are to be analyzed, these individual weights can be calculated based on each group member's pairwise comparisons. In cases of non-perfectly consistent judgments, the arithmetic mean of the weights assigned by the individual group members in general will differ slightly from the group weight as calculated from the geometric mean of the individual pairwise comparisons.

Recruitment of Participants

For the prioritization of outcome measures, we recruited two different panels. One panel included patients and the other was an expert panel with psychiatrists and psychotherapists. Around 20 patient organizations and self-help groups invited their member patients to participate in the AHP study. In addition, an invitation to participate was published on a German depression website. A sample of 12 patients volunteered to participate with an age range between 30 and 70 years (nine female, three male), suffering from moderate to severe depression, and currently experiencing a phase of remission or recovery. A patient undergoes a phase of remission when he or she has a depression symptom score (e.g. on a scale such as the Montgomery-Åsberg Depression Rating Scale [MADRS] or the Hamilton Rating Scale for Depression [HAM-D]) that is below the threshold considered as a major depressive episode. A patient is in recovery when remission holds for a longer period of time (about 6–12 months). Being in a phase of remission or recovery meant that patients would probably remember their experiences in the acute phase of depression, but also be aware of their preferences regarding their current state of depression. In this application of the AHP, no additional sociodemographic or patient characteristics were asked to keep patients comfortably anonymous. In an IQWiG statement officially signed by the Head of the Department of Health Economics, anonymity and confidentiality were guaranteed. Furthermore, patients were explicitly informed at the outset of the session that they could withdraw at any time during the process.

For the recruitment of the expert panel, experts were contacted who were involved in the development of clinical guidelines in the field of depression. Additional experts were selected from the websites of German scientific societies in the field of mental health, and websites of local private practices and hospitals. A group of seven experts were willing to participate, including psychiatrists and psychotherapists.

Data Collection and Inclusion

Weights of the Outcome Measures

Using hand-held radio-controlled keypads, the patients and the experts judged on the double nine-point scale of the AHP the pairwise comparisons between the outcome measures. Individual judgments were projected on a screen, allowing the members of the panel to discuss the rationales behind their individual scores. These discussions were meant to share information to make a more informed judgment, not for reasons of consensus formation. For reasons of privacy and to support giving honest judgments on sensitive issues, the patients were able to anonymously prioritize the patient-relevant outcome measures. Their names were not projected on the screen with their judgments, so they were not forced to discuss diverging judgments. For the same reasons, their discussions were not recorded. The experts prioritized the patient-relevant outcome measures non-anonymously, since they were not discussing individual patients. During the discussions, the panel members could alter their judgments. On the basis of the final individual judgments provided after the discussions, CRs and weighting factors were calculated for the group as a whole.

Inconsistencies in the group judgments were checked during the panel sessions. In order not to deter panel members from revealing personal judgments, inconsistencies in individual judgments were checked in a *post hoc* analysis. If group members had one cluster of pairwise comparisons with a CR higher than 0.4, these single pairwise comparisons were excluded from further analysis. Group members with an overall CR that remained higher than 0.2 were excluded on account of structural inconsistencies in judging the pairwise comparisons.

Performance Priorities of the Antidepressants

The comparison of the performance of the three antidepressants, venlafaxine, duloxetine, and SSRIs, was based on clinical evidence, as reported in a benefit assessment by IQWiG.[25] The reported pooled odds ratios for evaluating the performance priorities of the antidepressants were used to estimate the efficacy and adverse events of one antidepressant compared with the other.[25] Their performance on the quality-of-life measure was prioritized on the basis of the antidepressants' effect on quality of life in comparison with the effect of a placebo (SF-36, mental health component). These priorities reflect the relative performance of the antidepressants on each specific outcome measure. The overall performance priority of an antidepressant is the sum of the performance priority (p_n) regarding each outcome measure multiplied by the weight of this outcome measure (w_n): $\Sigma(p_n \times w_n)$ for $n = 1,2,3$.

We applied a sensitivity analysis to analyze the impact of the weights of the outcome measures on the overall performance priorities of the antidepressants.

Statistical Analysis

The weighting factors of the outcome measures, the performance priorities of the antidepressants, and the CRs were calculated by means of the software package Expert Choice version 11 (Expert Choice, Arlington, VA, USA). We compared patients and expert judgments by means of t-tests using SPSS version 18.0 (SPSS Inc., Chicago, IL, USA).

Results

Weights of the Outcome Measures

The sessions of both panels lasted around 3–4 hours in order to go through all 15 pairwise comparisons. The inconsistencies at the group level were all acceptable; no revisions were necessary during the panel sessions. Based on the *post hoc* analysis of the individual inconsistencies, the

judgments of 11 patients and five psychiatrists and psychotherapists were included in the AHP analysis. Their judgments met the overall consistency threshold level of 0.2. The overall CR is 0.02 for the patients and 0.01 for the psychiatrists and psychotherapists. Supplementary table S1 in the Supplemental Digital Content, http://links.adisonline.com/PBZ/A42, shows the weights and CRs of the patient and expert panels before and after the exclusion of inconsistent pairwise comparisons.

Figure 2 graphically presents the weights of the lowest level patient-relevant outcome measures as assigned by the patients and experts. These group weights are calculated by using the geometric mean of the panelists' pairwise comparisons.

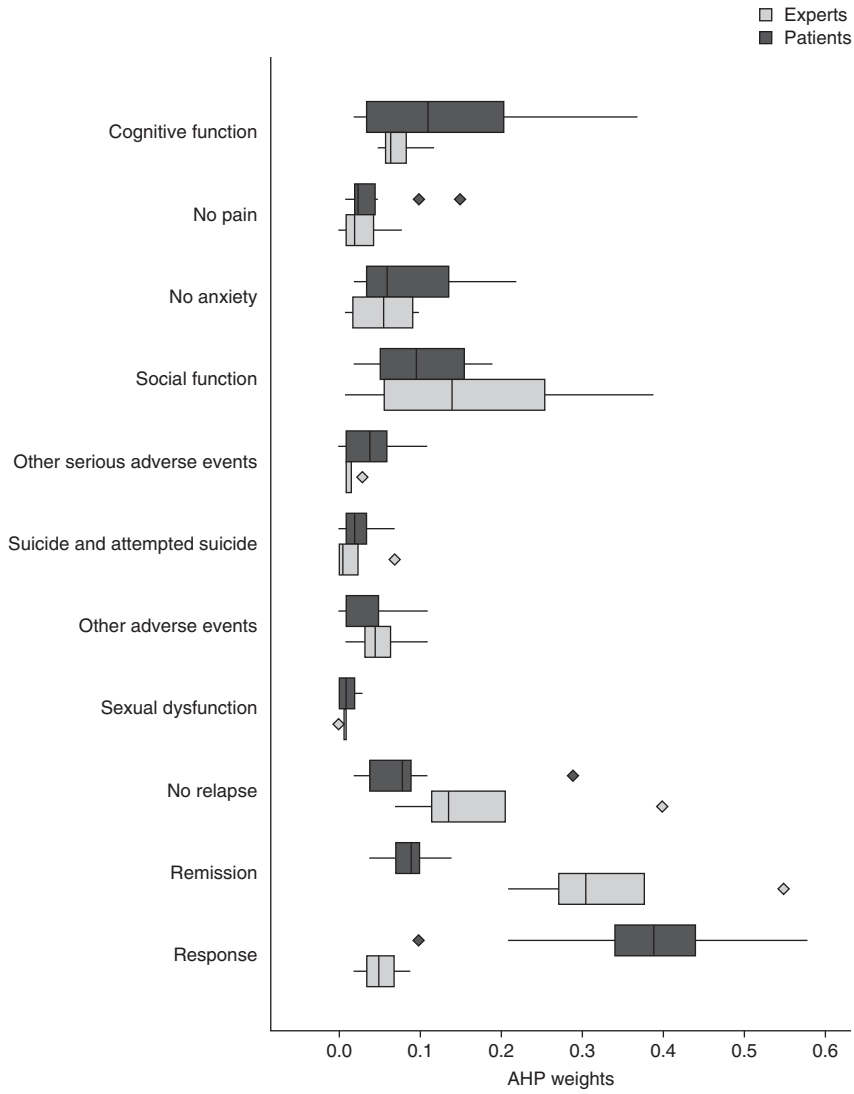

Fig. 2. Box plots of weights assigned by patients and experts to the outcome measures. The diamonds are outliers. The ends of each box represent the 25th and 75th percentiles, and the ends of each line show the 95% confidence interval. Lines within boxes represent medians (i.e. 50th percentiles). **AHP** = analytic hierarchy process.

The outcome measures that the patients ranked to be the six most important are confirmed in the experts' ranking of outcome measures. The experts only added 'other adverse events' as one of the most important outcome measures. The six confirmed patient-relevant outcome measures related to efficacy: response, remission, and no relapse; and to the accompanying effects on quality of life: social function, cognitive function, and no anxiety. The weights assigned to these six outcome measures cover 86% of the overall weight assigned by the patient group. The priority weight of the combined category of 'all adverse events' (0.098 as rated by the patients) is in the same order of magnitude as the weight of each of the accompanying symptoms. These seven outcome measures, including the combined outcome measure 'all adverse events,' cover 96% of the overall weight as assigned by the patients.

In addition, the box plots in figure 2 show the differences in weights within the patient and expert panels. Within the two panels, judgments differed little in terms of standard deviations regarding the importance of the different (serious) adverse events. Significant differences were displayed in the importance of response and remission, and smaller differences were found regarding the accompanying effects on quality of life. Patients particularly differed in opinion among themselves about the importance of response and no relapse, as well as the accompanying effects on quality of life: cognitive function, no anxiety, and social function. The largest individual differences in the patient panel were that two patients more strongly emphasized the importance of no relapse, and one patient the importance of no anxiety. Regarding the accompanying effects on quality of life, the psychiatrists and psychotherapists differed more in opinion about the importance of social function than the patients did.

The patient and expert groups differed significantly on response (weight patients = 0.37; weight experts = 0.05), remission (weight patients = 0.09; weight experts = 0.40), and the less important outcome measure of other serious adverse events (weight patient = 0.06; weight experts = 0.04) [$p \leq 0.05$]. The patients considered the response to antidepressants to be of utmost importance. One statement illustrated the high importance of response to patients: *''I would rather live the rest of my live with a mild depression, than to have no hope in the acute state of depression that the medicine will give me some relief.''* The experts had a different perspective on the importance of response. One argument they mentioned was *''response is not that important. What is important is full remission. If there is no full remission, long-term prognosis for the patient is not so good.''*

Performance Priorities of the Antidepressants

As an illustration, table II presents an approximation of the performance priorities of the antidepressants on each outcome measure. These priorities are derived from clinical evidence on pairwise comparisons of the performance of the antidepressants as described in the benefit assessment report by IQWiG.[25] For example, the pooled odds ratio of response due to venlafaxine versus SSRIs is 1.20 (95% confidence interval [CI] 1.07, 1.35), due to duloxetine versus SSRIs is 0.96 (95% CI 0.80, 1.15), and due to duloxetine versus venlafaxine is 0.75 (95% CI 0.52, 1.08). The pooled odds ratios were used to score the pairwise comparisons of the antidepressants. By means of the principal right eigenvector approach, the AHP approximated the performance priorities of duloxetine, venlafaxine, and SSRIs on response to be 0.30, 0.39, and 0.31, respectively. A higher performance priority of one of the antidepressant treatments reflects a more positive outcome.

The overall performance priorities of the antidepressants are the weighted average of the performance priorities on each outcome measure. The weights of the outcome measures as assigned by the patient panel are used to calculate the patients' overall performance priorities, and the weights of the expert panel are used to calculate the experts' overall performance priorities.

Sensitivity Analysis

We performed a sensitivity analysis on the weights of the outcome measures as assigned by the patients. As one weight is heightened, the remainder of the criteria weights are lowered in

Table II. Performance priorities of the antidepressants

Antidepressants	Outcome measures						Overall performance priority	
	Efficacy			AEs		QOL ($w_p = 0.37$; $w_e = 0.28$)	Patients	Experts
	Response ($w_p = 0.37$; $w_e = 0.05$)	Remission ($w_p = 0.09$; $w_e = 0.40$)	No relapse ($w_p = 0.07$; $w_e = 0.17$)	AE ($w_p = 0.03$; $w_e = 0.07$)	SAE ($w_p = 0.06$; $w_e = 0.04$)			
Duloxetine	0.298	0.344	0.333[a]	0.316	0.387	0.244	**0.291**	**0.312**
Venlafaxine	0.387	0.347	0.333[a]	0.296	0.238	0.362	**0.358**	**0.344**
SSRIs	0.315	0.310	0.333[a]	0.388	0.375	0.394	**0.351**	**0.345**

a Clinical data are not available; antidepressants were assumed to be equally effective with respect to preventing a 'relapse.'

AE = adverse event; **QOL** = quality of life; **SAE** = serious adverse event; **SSRI** = selective serotonin reuptake inhibitor; $\mathbf{w_e}$ = the weight assigned by the expert panel; $\mathbf{w_p}$ = the weight assigned by the patient panel.

proportion, for the weights to keep summing up to one. Changing the weight of quality of life or efficacy will not reverse the rank order of duloxetine; this antidepressant still has the lowest overall performance priority. Increasing the weight of adverse events to 0.45 or remission to 0.54 will, however, alter the overall performance priority of duloxetine so that it becomes higher than that of either venlafaxine or SSRIs. However, these weights are significantly higher than the weights as assigned by both the patient panel and the expert panel ($p > 0.05$). These overall priorities suggest that, despite their conflicting judgments on the importance of response and remission, both the patient panel and the panel with psychiatrists and psychotherapists appear to slightly prefer venlafaxine and SSRIs over duloxetine.

Discussion

In this study, the AHP technique was used to elicit patient preferences for multiple outcomes of antidepressant treatment. The weights calculated offered a quantitative overview of the relevance of the outcome measures. Both patients and experts agreed upon the high importance of outcome measures such as response, remission, no relapse, cognitive function, no anxiety, and social function. The group discussions offered insight in the question why these outcome measures were considered to be important for patients suffering from a major depression.

The assessment also revealed outcome measures that are of lesser importance for this specific group of patients being studied. For example, patients' and experts' views coincided on the irrelevance of the negative effect of antidepressants on sexual function. It may thus be concluded that the study not only provides weights for important outcome measures, but also can identify outcome measures, or clinical endpoints, that are perceived to have less meaning in both clinical and health policy decision making. Furthermore, prioritization by means of the AHP can draw attention to the fact that combined or aggregated outcome measures need to be used. Even though each of the adverse events was given a low priority separately, a combined outcome measure of all (serious) adverse events was sufficiently relevant to be taken up in an HTA.

The patients and experts most strongly disagreed on the importance of response and remission. The strikingly high weight that the patients assigned to response to antidepressants in the acute phase of their depression suggests that, even though none of the patients were in the acute phase of depression anymore, they were able to take their past experiences into account. In contrast, the experts emphasized the importance of remission. This could be explained by their longer-term perspective, their focus on the long-term chances for complete recovery. For future research, the authors recommend recruiting, as in this study, patients in remission as well as patients who have been in a recovery phase for an extensive period of time in order to take account of the shorter- and longer-term perspectives. In our case, the differences in priorities between patients

and experts did not result in different conclusions on the overall performance of the antidepressants.

The outcome measures included were derived from existing clinical trials and were selected by researchers. It is also possible to add a brainstorming phase into the AHP procedures. Besides the outcome measures applied in clinical studies, the patients and experts can then add more patient-relevant outcome measures to the AHP structure if necessary. One only needs to be careful not to add outcome measures that overlap the clinical outcome measures already included. Clear conceptual definitions of all outcome measures are essential in avoiding overlap. In our case, it might have been possible that patients would have wanted to add criteria, for example the possibility of addiction to medication as identified by Wittink et al.[22] The patients in this case, however, confirmed after the session that they agreed with the criteria selected.

In psychiatry and other clinical fields, many clinical outcome measures are related or do overlap with intermediate or surrogate endpoints. For example, social function is related to other accompanying symptoms affecting quality of life, such as anxiety. Concerning the use of any technique of MCDA, overlap among criteria may have an undesired impact on the outcomes and result in rank reversal. Rank reversal would mean that a therapeutic alternative is falsely chosen over another that would actually be most preferred if any potential overlap between criteria could be prevented. In general, some overlap between outcome measures is less problematic if one is to gain insight in selecting the most important outcome measures; the set of most important outcome measures can still be identified. More problematic is this overlap in cases where multiple outcome measures are aggregated to estimate the overall preferences for the alternative treatments. Then, the performance of treatments on the overlapping outcome measures is too strongly weighted in the aggregated performance priorities of the alternative treatments. If this overlap were to change the ranking of treatments in order of decreasing performance priority, this rank reversal can affect the policy or clinical decisions on these treatments. Sensitivity analysis can be applied to check the possibility of a rank reversal of treatments. A procedure for sensitivity analysis on the weights of outcome measures has been suggested by Mareschal,[31] and a more comprehensive sensitivity analysis that includes altering the performance priorities as well has been suggested by Triantaphyllou and Sanchez.[32] In our case, the sensitivity analysis showed that changing the weights of the main outcome measures is not likely to evoke a rank reversal of the least preferred antidepressant. In cases in which a rank reversal is likely, researchers can identify overlap among the outcome measures, for example, in discussion with a focus group with patients and experts using clear conceptual definitions of the outcome measures. In cases where outcome measures are interrelated, the use of the analytic network process could be considered, which takes into account the interactions between criteria.[33]

The group judgments about the relevance of the outcome measures showed a high level of overall consistency. At the individual level, relatively more patients than experts were consistent in their judgments. Even though it can be argued that inconsistent judgments in general should be removed from the analysis,[34] the calculations can be adapted to warn for highly inconsistent individual judgments during the panel session as well. This would enable the correction of at least the accidental mistakes. For future studies, we recommend the correction of these accidental mistakes during the panel session, and the inclusion of all judgments in case the overall inconsistency is acceptable.

One point of criticism regarding the patient panel involved is that it might not be representative of the larger group of patients with a major depression. A relatively homogeneous group of patients volunteered to participate in the evaluation. All patients were from the same ethnic and cultural background, in a state of remission or recovery, likely to be knowledgeable about the effects of antidepressant medicaments, and likely to have had antidepressant treatment with pharmaceutical agents. Such characteristics by themselves can impact the patients' preferences for antidepressant treatments.[23] For example, patients rated cognitive function to be more important than the experts did. The impact of a major de-

pression on cognitive function, or the importance attached to a strong cognitive function, might differ strongly among patients, and could possibly be slightly overstated in our sample of patients. One explanation could be that the group of patients willing to participate was highly educated and well informed, and hence is less representative. Moreover, the judgments revealed in this homogeneous group may not reflect the full diversity in judgments of the overall patient population. The difference in opinions within the patient panel about other accompanying effects on quality of life could be explained by the heterogeneity in the patient group.

Should values be derived for actual reimbursement decisions, we strongly recommend that sociodemographic characteristics and characteristics of the major depression are adequately considered in the selection of the patient sample on the basis of epidemiologic data. Characteristics such as education and the current stage of the major depression as shown in this study, as well as gender, ethnicity, and co-morbidities, need to be taken into account when selecting participants.[23] If it is not possible to select one panel large enough to be representative and small enough to have effective group discussion, multiple panels may be organized to capture the diversity of patients. Another possible solution is to gather additional information from a larger group of patients by means of (online) surveys. No specific guidelines exist for the minimal sample size in AHP experiments to assure for stable or robust weights. Power analysis can provide a rough indication of the minimum amount of patients required to weigh outcome measures for a given CI. If the results were based on a representative panel of patients, we would certainly use the weights derived from patient panels to support health economic evaluations, despite any deviations from values elicited in an expert panel. This panel needs to be informed of the performance of the antidepressants in order to be able to assign the appropriate weights to the outcome measures.[35]

As an example, we used clinical evidence to prioritize three antidepressants with regard to the weighted set of outcome measures. A drawback is that clinical evidence was not available on all outcome measures. Clinical evidence on the impact of the antidepressants on relapse and the outcome measures regarding quality of life were missing. In this illustration, the differences in the overall performance priorities of the antidepressants were small. This would indicate that neither the patient panel nor the experts appeared to have a strong preference for one of the antidepressants. Nevertheless, individual patients might differ in their preferences for the antidepressants. For example, the patient who strongly emphasized the importance of no anxiety might specifically prefer an antidepressant that outperforms the other antidepressants in reducing anxiety. When the AHP is used to support clinical decision making, it can be essential to take such individual preferences into account, while in supporting health policy decision making these individual preferences could be aggregated.

Formal guidelines on how to integrate clinical evidence into the AHP analysis still need to be developed, and we recommend further research on this topic. When comparing, for example, three healthcare interventions, you either need a three-arm randomized controlled trial with these three interventions, or an adjusted indirect comparison for each endpoint of interest with placebo as a common comparator or another common comparator. In cases where the estimators are derived from different placebo controlled trials, there is a risk of heterogeneity and even inversion of the effects if these studies differ in structural equity. Another methodological consideration is that odds ratios that are larger than nine or smaller than one-ninth do not fit the nine-point AHP scale. In our illustration, the odds ratios are not violating the magnitude of this original AHP scale. An alternative solution could be to first transform the relative estimators on a logarithmic scale as distances, and then to normalize them. Accordingly, the performances of the healthcare interventions are shown in a span of 0–1, like all performance priorities calculated with the AHP. Depending on the CI of each relative estimator, sensitivity analysis needs to be conducted for these estimators of the performance priorities, as well as sensitivity analysis for the patient weights of the outcome measures.

Conclusions

The AHP can be used to weight outcome measures of clinical studies, even in this complex application area of major depression. In applying the AHP, overlap between outcome measures is to be minimized and a representative patient panel needs to be selected. The weights derived provide an overview of the relevance of the outcome measures to patients. The AHP can be used to (i) support clinical investigators in selecting patient-relevant outcome measures to be used in clinical assessments, and (ii) to weight the clinical outcomes regarding the different outcome measures to support clinical and health policy decision making. This is particularly relevant in case different outcome measures lead to conflicting results about the best alternative to reimburse. Although AHP may be used for other purposes, the present study was intended to demonstrate how patients' preferences could feed into a policy decision on reimbursement in Germany for the case of the antidepressants.

Acknowledgments

The work reported in this paper was financially supported by IQWiG. No conflicts of interest are declared, either for the authors or for the funding organization.

Marjan Hummel is the main author of the paper and conducted the AHP analyses. Fabian Volz and Marion Danner prepared the questionnaire, organized the panel sessions, and contributed to writing the paper. Jeannette van Manen contributed to the (statistical) analyses and writing the paper. Markos Dintsios took part in the preparations of the project and contributed to writing the paper. Maarten IJzerman facilitated the project and contributed to writing the paper. Andreas Gerber as Head of Department formally supervised the whole process and also contributed to writing the paper. Marjan Hummel acts as guarantor for the content of the paper.

References

1. Bridges J. Stated preference methods in health care evaluation: an emerging methodological paradigm in health economics. Appl Health Econ Health Policy 2003; 2 (4): 213-24
2. Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ Policy Law 2009; 4 (4): 527-46
3. Saaty TL. Highlights and critical points in the theory and application of the analytic hierarchy process. Eur J Oper Res 1994; 74: 426-47
4. 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 (4): 229-48
5. Dolan JG, Bordley DR. Individualized patient decision-making using the analytic hierarchy process (AHP): reliability, validity, and clinical usefulness [abstract]. Med Decis Making 1991; 11 (4): 322
6. IJzerman M, van Til JA, Bridges JFP. A comparison of analytic hierarchy process and conjoint analysis methods in assessing treatment alternatives for stroke rehabilitation. Patient 2012; 5 (1): 45-56
7. Hilgerink MP, Hummel JM, Manohar S, et al. Assessment of the added value of the Twente Photoacoustic Mammoscope in breast cancer diagnosis. Medical Devices (Auckl) 2011; 4: 107-15
8. Liberatore MJ, Nydick RL. The analytic hierarchy process in medical and health care decision making: a literature review. Eur J Oper Res 2008; 189 (1): 194-207
9. Dolan JG, Bordley DR. Diagnostic strategies in the management of acute upper gastrointestinal bleeding: patient and physician preferences. J Gen Intern Med 1993; 8 (10): 525-9
10. Hummel JM, van Rossum W, Verkerke GJ, et al. The effects of team expert choice on group-decision making in collaborative new product development: a pilot study. J Multi-Criteria Decis Anal 2000; 9 (1-3): 90-8
11. Bloom BS. Prevalence and economic effects of depression. Manag Care 2004; 13 (6): 9-16
12. O'Connor EA, Whitlock EP, Beil TL, et al. Screening for depression in adult patients in primary care settings: a systematic evidence review. Ann Intern Med 2009; 151: 793-803
13. Masand PS, Gupta SG. Long-term effects of newer-generation antidepressants: SSRIs, venlafaxine, nefazodone, bupropion, and mirtazapine. Ann Clin Psych 2002; 14 (3): 175-82
14. Nuevo R, Leighton C, Dunn G, et al. Impact of severity and type of depression on quality of life in cases identified in the community. Psychol Med 2010; 11: 1-9
15. Fournier JC, DeRubeis RJ, Hollon SD, et al. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 2010; 303 (1); 47-53
16. Williams Jr JW, Mulrow CD, Chiquette E, et al. A systematic review of newer pharmacotherapies for depression in adults: evidence report summary. Ann Intern Med 2000; 132 (9): 743-56
17. Cunningham LA. Once-daily venlafaxine extended release (XR) and venlafaxine immediate release (IR) in outpatients with major depression. Ann Clin Psychiatry 1997; 9 (3): 157-64
18. Brannan SK, Mallinckrodt CH, Brown EB, et al. Duloxetine 60 mg once-daily in the treatment of painful physical symptoms in patients with major depressive disorder. J Psychiatr Res 2005; 39 (1): 43-53
19. Hewett K, Chrzanowski W, Schmitz M, et al. Eight-week, placebo-controlled, double-blind comparison of the antidepressant efficacy and tolerability of bupropion XR and venlafaxine XR. J Psychopharmacol 2009; 23 (5): 531-8
20. Smith WT, Glaudin V, Panagides J, et al. Mirtazapine vs. amitriptyline vs. placebo in the treatment of major depressive disorder. Psychopharmacol Bull 1990; 26 (2): 191-6

21. Cooper-Patrick L, Powe NR, Jenckes MW, et al. Identification of patient attitude and preferences regarding treatment of depression. J Gen Intern Med 1997; 12: 431-8
22. Wittink MN, Cary M, Tenhave T, et al. Towards patient-centered care for depression: conjoint methods to tailor treatment based on preferences. Patient 2010; 3 (3): 145-57
23. Dwight-Johnson M, Sherbourne CD, Liao D, et al. Treatment preferences among depressed primary care patients. J Gen Intern Med 2000; 15: 527-34
24. Danner M, Hummel JM, Volz F, et al. Integrating patients' views into health technology assessment: analytic hierarchy process (AHP) as a method to elicit patient preferences. Int J Technol Assess Health Care 2011; 27 (4): 1-7
25. IQWiG. Selective serotonin and norepinephrine reuptake inhibitors (SNRI) for patients with depression: executive summary [IQWiG reports – commission no. A05-20A]. Cologne: IQWiG, 2009 [online]. Available from URL: https://www.iqwig.de/download/A05-20A_Executive_Summary_SNRI_for_patients_with_depression.pdf [Accessed 2012 Jun 11]
26. IQWiG. Bupropion, mirtazapine, and reboxetine in the treatment of depression: executive summary [IQWiG reports – commission no. A05-20C]. Cologne: IQWiG, 2009 [online]. Available from URL: https://www.iqwig.de/download/A05-20C_Executive_summary_preliminary_report_Bupropion_mirtazapine_and_reboxetine_in_the_treatment_of_depression.pdf [Accessed 2012 Jun 11]
27. Triantaphyllou E, Lootsma FA, Pardalos PM, et al. On the evaluation and application of different scales for quantifying pairwise comparisons in fuzzy sets. J Multi-Criteria Decis Anal 1994; 3 (3): 133-55
28. Saaty TL. The analytic hierarchy process: planning, priority setting, resource allocation. New York: McGraw-Hill, 1980
29. Harker PT. Derivatives of the Perron root of a positive reciprocal matrix: with application to the analytic hierarchy process. Appl Math Comput 1997; 22 (2-3): 217-32
30. Forman E, Peniwati K. Aggregating individual judgments and priorities with the analytic hierarchy process. Eur J Oper Res 1998; 108: 165-9
31. Mareschal B. Weight stability intervals in multicriteria decision aid. Eur J Oper Res 1988; 33: 54-64
32. Triantaphyllou E, Sanchez A. A sensitivity analysis approach for some deterministic multi-criteria decision making methods. Decision Sci 1997; 28: 151-94
33. Saaty TL, Vargas LG. Decision making with the analytic network process: economic, political, social and technological applications with benefits, opportunities, costs and risks. New York: Springer Science and Business Media, 2006
34. Lancsara E, Louviere J. Deleting 'irrational' responses from discrete choice experiments: a case of investigating or imposing preferences? Health Econ 2006; 15: 797-811
35. Holder RD. Some comments on the analytic hierarchy process. J Opl Res Soc 1990; 41 (11): 1073-6

Correspondence: *Marjan Hummel*, PhD, Department of Health Technology and Services Research, University of Twente, PO Box 217, 7500 AE Enschede, the Netherlands.
E-mail: J.M.Hummel@utwente.nl