Ending the war in multi-criteria decision analysis:

Taking the best from two worlds

The development and evaluation of guidelines for the use of MACBETH in multi-criteria group decision making for the assessment of new medical products

Lea Rietkötter
February 2014 - August 2014
University of Twente
Master of Science Thesis
Health Sciences
Specialization Health Technology Assessment and Innovation

Student: Lea Franziska Barbara Rietkötter
Studentnumber: s1404628
Mail: lea.rietkoetter@gmail.com

First Supervisor:
J.M. Hummel, PhD
Assistant professor | HTSR – University of Twente

Second Supervisor:
H. Broekhuizen, MSc
PhD Student | HTSR – University of Twente

Third Supervisor:
C.G.M. Groothuis-Oudshoorn, PhD
Assistant professor | HTSR – University of Twente

UNIVERSITY OF TWENTE.

Preface

This Master of Science Thesis aims at delivering new insights into the field of healthcare-related MCDA. The findings could be useful for the implementation of new approaches and improve the outcomes of decision making processes as a whole. In addition, gained data about the relative importance of frailty criteria will be helpful for the development of nutritional compounds, with the objective of an improvement in the quality of life of an ageing society.

This report is the final work of my Master studies of Health Sciences, specialization Health Technology Assessment and Innovation, at the University of Twente in Enschede. The past year has been absolutely memorable and the knowledge I have accumulated over such a short period of time is of great value for my career. Studying at a Dutch university has been a new experience for me and I am grateful for the new positive impressions and this great opportunity.

I would like to thank my supervisors Marjan Hummel, Henk Broekhuizen and Karin Groothuis-Oudshoorn for their support and numerous feedback. Marjan, you have taught me so much about MCDA and your expertise about and passion for this topic has been an inspiration. Henk, I am so thankful for your permanent help regarding questions of all kind, the insights you have given me to R and your patience concerning my work in R. Although in the end I was unable to use R very much, I am sure I will work with it in the future. Karin, you have assisted me a lot in modifying the structure and improving the writing of this report and I really appreciate it. I very much look forward to writing an article about this research with all of you and staying in contact. Furthermore, I thank João Bana e Costa from Bana Consulting for answering several questions concerning MACBETH and for supplying me with information about WISED. In addition, my deepest gratitude goes to my parents, who have helped me to finance my studies and supported me unconditionally throughout my academic track, and my husband Fraser. Fraser, you have proofread this work several times, you are partially responsible for my proficiency in English and you have given me encouragement in so many situations. I cannot thank you enough.

By completing this thesis, I finish my academic career. From September onwards I will start working at the Universitätsklinikum in Münster and I am excited about the challenges that work life holds ready for me. I am certain that my accumulated knowledge and the set of skills I have developed in the past years of studying will assist me in adapting to the new environment and being successful at my job.

Lea Rietkötter

August, 2014

Abstract

Background Information

Multi-criteria decision analysis (MCDA) has been applied in a variety of domains, including the healthcare market. With many multidisciplinary decision makers, uncertainties, complex structures and necessary trade-offs, the healthcare market is one field where multi-criteria group decision making could find many applications. The Analytic Hierarchy Process (AHP) is an MCDA method based on pairwise comparisons that has seen ongoing use in healthcare-related multi-criteria group decision making. It has proven to be an effective tool to support individual and group decision making with multiple criteria, however, the AHP has been criticised on a methodological level by several authors in the past. The "Measuring Attractiveness by a Categorical Based Evaluation Technique"-method (MACBETH) is also based on pairwise comparisons, but the execution of judgements and the mathematical foundations of MACBETH differ from the AHP. Although MACBETH has predominantly received positive feedback and is widely accepted, the approach has only been used very rarely in healthcare-related issues.

Objective

"What guidelines can be developed to use MACBETH as a feasible method for multi-criteria group decision making in the assessment of criteria related to the development of new medical products?"

The feasibility of the guidelines will be tested by means of a pilot panel session concerning the assessment of criteria related to the prevention of frailty in the elderly. To investigate the research question, five sub-questions were answered:

- *"What are the methodological differences between AHP and MACBETH?"*
- *"How can best practices of group decision making with the AHP in healthcare be used in the MACBETH-procedure?"*
- *"What is the relative importance of frailty criteria evaluated by elderly individuals?"*
- *"How are the methodology and the results of the group-assessment of frailty criteria using MACBETH perceived by elderly individuals?"*
- *"What recommendations should be proposed to enhance the use of MACBETH in healthcare-related questions?"*

Methods

A thorough literature study was conducted to compare the frequently used AHP and the MACBETH approach. This method determined specific structures, differences, advantages and disadvantages of both approaches. Best practices of the AHP were investigated and used for the development of guidelines for the use of MACBETH in healthcare-related group

decision making. Additionally, a test panel session and a pilot panel session were carried out. The test panel session was used for the assessment of the online tool “QuestionPress”, the pilot panel session assisted in testing the feasibility of the previously developed guidelines with elderly participants. In addition, data about the relative importance of frailty criteria was gained. Finally, a qualitative questionnaire was used to investigate the perception of the MACBETH methodology by elderly individuals.

Results

The literature study discovered several major methodological differences between AHP and MACBETH. Nonetheless, best practices of the AHP could be adapted in the guidelines for the use of MACBETH in healthcare-related multi-criteria group decision making. The guidelines demonstrated to be a feasible tool in the pilot panel sessions, minor issues were detected and the guidelines were modified. “Cognition” was chosen to be the most important frailty criterion by four out of six participants and the aggregated weights also identified “Cognition” to be of highest relative importance with a weight of 17.68. One of the major findings was the eligibility of the arithmetic mean for aggregated weights in group decision making. The questionnaire showed that elderly individuals overall found the methodology difficult. The high number of semantic categories was also criticised. Nonetheless, three participants liked the procedure and the user-friendliness and rated the assessment to be very valuable for future assessments.

Discussion/Conclusion

This work provides a comprehensive explanation and comparison of the AHP and the MACBETH approach. If software for group-decision making with MACBETH would be improved, this method is expected to be used more frequently in general and in healthcare. The developed guidelines for the use of MACBETH in healthcare-related multi-criteria group decision making are also likely to enhance the use of MACBETH in this field. It is recommended to use the arithmetic mean in group decision making with MACBETH, if no consensus can be found. The weights of frailty criteria can be helpful in the development of nutritional compounds for the prevention of frailty in the elderly. However, further research is needed to confirm and validate the findings.

Keywords

MCDA, MCDM, AHP, MACBETH, healthcare, frailty, elderly

Contents

List of Tables and Figures

Tables

Figures

List of Abbreviations

6MW	Six-Minute Walk Distance
AHP	Analytic Hierarchy Process
AMSM	Accelerometric Measurements on Sensor Mat
Bal	Balance
BBS	Berg Balance Scale
CDT	Clock Drawing Test
CGM	Continuous Glucose Monitoring
CI	Consistency Index
CL	Cholesterol Level
Cog	Cognition
COP	Condition of Order Preservation
CR	Consistency Ratio
CST	Chair Stand Test
DBP	Diastolic Blood Pressure
Dep	Depression
ELECTRE	Elimination Et Choix Traduisant la Realité (Elimination and Choice Expressing Reality)
Ene	Energy
FGL	Fluctuating Glucose Levels
FOSQ	Flight of Stairs Questionnaire
FSST	Four Square Step Test
GDS	Geriatric Depression Scale
GPCOG	General Practitioner Assessment of Cognition
Gri	Grip Strength
GS	Gait Speed
HRSC	Hamilton Rating Scale for Depression
Hyp	Hypertension
HypC	Hypercholesterolemia
JD	Jamar Dynamometer
LES	Leg Extension Strength
MACBETH	Measuring Attractiveness by a Categorical Based Evaluation Technique
MCDA	Multi-Criteria Decision Analysis
Mob	Mobility
PROMETHEE	Preference Ranking Organization Method for Enrichment Evaluation
RAVLT	Rey Auditory Verbal Learning Test
SBP	Systolic Blood Pressure
SD	Standard Deviation
SFQ	Subjective Fatigue Questionnaire

SMART	Simple Multi-Attribute Rating Technique
SMMSE	Standardized Mini-Mental State Examination
SPPB	Short Physical Performance Battery
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TUG	Timed Up and Go

1. Introduction

1.1 Background Information

Individuals and institutions face numerous decisions on a daily basis. In many of these cases, the decision making process involves not only one criterion, but a large number of points of comparison. This impedes a quick and easy choice of an alternative. For these issues, multi-criteria decision analysis (MCDA) was developed [1]. MCDA is both an approach and a number of techniques, used to provide priorities or rankings for several alternatives in complex decision problems [2]. By focusing on the decision maker and gaining subjective preference information, the most suitable solution for the decision maker to a problem is determined [3].

In general, the MCDA process starts with the identification of a problem, which will then be structured. This phase includes the definition of a decision goal, the identification of stakeholders, uncertainties, criteria and alternatives and the development of a collective understanding of the issue at hand. The next step is the model building, in which alternatives and criteria are explicitly defined and values are determined through preference modelling and measurement using a specific MCDA method. After gaining a first recommendation, the model is used to support the decision making process by analysing the sensitivity and the robustness of the results. In the end, a plan for further action is developed. [4]

There are several methods that can be used to solve multi-criteria problems. The most common ones are simple multi-attribute rating technique (SMART), analytic hierarchy process (AHP), technique for order preference by similarity to ideal solution (TOPSIS), outranking methods (such as elimination and choice expressing reality (ELECTRE) and preference ranking organization method for enrichment evaluation (PROMETHEE)) and goal programming [5, 3]. The choice of an appropriate technique depends on the decision maker and their cognition, the decision problem at hand, the possibility to gain necessary input data and the availability and user-friendliness of corresponding software as a decision support system [6].

Different MCDA methods have been widely implemented in a variety of domains. Also in the healthcare market, where the number of published articles about MCDA presents a near-exponential growth from 1960 to 2011, with a climax of 66 publications in 2010 [7]. Another just recently published literature review by Marsh, Lanitis [8] identified 40 studies concerning MCDA interventions in healthcare. More than 50 % of them were published in the past 3 years. However, to date many healthcare-related judgements and choices are still based on intuition or deliberate processes and therefore lack transparency and rationality

[9]. The healthcare market is very complex, decisions are usually multifaceted and impact and consequences are high, which increases the need for MCDA [7].

With many multidisciplinary decision makers, uncertainties, complex structures and necessary trade-offs, group decision making plays a major role in the healthcare field [10]. It is used to pool expert knowledge, level out individual differences and improve the quality of a decision as a whole [11, 12]. Most MCDA methods are applicable to both individual and group decision settings, but the group procedures differ. Belton and Pictet [13] distinguish between sharing (group members operate like a single decision maker), aggregating (pooling of individual judgements) and comparing (individual judgements are used as a basis for discussion).

The AHP is one MCDA technique that has been applied several times for group decision making. It was developed by Thomas L. Saaty in the 1970's with the aim, to convert subjective relative preferences to a set of overall weights [5]. It originally employs the eigenvalue method to gain priorities for alternatives and weights for decision criteria [14]. It has seen on-going use in the field of healthcare, from medical diagnosis to patient participation, treatment, technology and healthcare evaluation and human resource planning [15]. In group decision making, the AHP utilizes the previously explained aggregating approach, which is also supported by the accompanying software "Expert Choice"[1]. The AHP has received both positive and negative criticism in the past, which will be further evaluated in section 2.2.1, *Discussion about the AHP*.

MACBETH (Measuring Attractiveness by a Categorical Based Evaluation Technique) also uses pairwise comparisons, but the execution of judgements is carried out differently. Furthermore, the calculation of scores is achieved by employing linear programming instead of an eigenvalue method [17]. MACBETH is also suitable for group decision making, but the software M-MACBETH uses the sharing instead of the aggregating approach to gain group priorities. MACBETH has predominantly received positive feedback and is widely accepted [18, 19]. However, besides many public and private applications, MACBETH has very rarely been used in healthcare-related issues [20].

Both the AHP and MACBETH have a high number of supporters and the heated debate about the preferred method is ongoing. The similarities and major differences between both approaches therefore put MACBETH and AHP in the focus of this work.

[1] "Expert Choice" is a decision support software which uses the AHP to gain priorities for alternatives and weights for decision criteria. It is used commercially as well as for research at universities and government facilities. [16]

1.2 Research Objective and Research Question

The current lack of MACBETH applications in healthcare implicates the question, why such a widely accepted method has not yet been frequently used in this particular field. Furthermore, it is of interest to what extent this method can be adapted for group decision making and whether alterations are necessary to apply MACBETH in this specific sector.

Nutricia is a division of the company Danone and focuses solely on medical nutrition. It aims at providing a customized diet for individuals with specific requirements and needs [21]. The prevention of frailty in elderly is of high interest for Nutricia. However, the absence of weighting factors for different frailty criteria complicates the development of specific dietary supplements. The assessment of frailty criteria with MACBETH is therefore one specific goal of this study.

Based on this background, the research question is:

"What guidelines can be developed to use MACBETH as a feasible method for multi-criteria group decision making in the assessment of criteria related to the development of new medical products?"

The feasibility of the guidelines will be tested by means of a pilot panel session concerning the assessment of criteria related to the prevention of frailty in the elderly. To investigate the research question, five sub-questions will be answered:

- *"What are the methodological differences between AHP and MACBETH?"*
- *"How can best practices of group decision making with the AHP in healthcare be used in the MACBETH-procedure?"*
- *"What is the relative importance of frailty criteria evaluated by elderly individuals?"*
- *"How are the methodology and the results of the group-assessment of frailty criteria using MACBETH perceived by elderly individuals?"*
- *"What recommendations should be proposed to enhance the use of MACBETH in healthcare-related questions?"*

1.3 Structure of the Thesis

Due to the fact that AHP unlike MACBETH has been used frequently in the healthcare market and in group decision making, in which it employs a different approach, this work will first execute a thorough literature study to explain and compare both techniques. The objective of this method is the determination of specific structures, differences and advantages and

disadvantages between AHP and MACBETH. Furthermore, the results of the literature study will result in the creation of guidelines for the use of MACBETH in the group assessment of criteria related to the development of new medical products.

The practical section of this work will focus on the planning, the conduction and the evaluation of an assessment using MACBETH concerning the importance of different frailty criteria of the elderly. A test panel session will be used for the assessment of an online audience response system. The following pilot panel session has several objectives: First it will test the feasibility of the previously developed guidelines for healthcare-related group decision making with MACBETH. Further, it will determine potential issues during the group decision making with elderly individuals using MACBETH. Third it will help to identify possibilities and recommendations for the broader use of MACBETH in healthcare. And finally, the assessment will result in the collection of valuable data about criteria regarding the prevention of frailty of the elderly.

After presenting the results of the literature study and the panel sessions, the discussion is used to evaluate the applicability of MACBETH as a tool for group decision making in healthcare. Recommendations for the use of MACBETH in the assessment of healthcare technology and potentially necessary modifications will be proposed and the validity of collected data will be interpreted.

2. Theory

2.1 Method Literature Research

The first step of this study was a comprehensive literature research, which was performed on 3rd March 2014. This method was chosen to explain the frequently applied AHP and the less commonly used MACBETH procedure and to investigate methodological differences between both approaches. This comparison was conducted predominantly on methodological criteria, namely the MCDA model, the structure, the decision hierarchy, the number of pairwise comparisons, the nature of the judgements, the scale, the consistency, the calculation of priorities and weights, the application in healthcare, the available group decision support software and the criticism concerning the methods. Furthermore, the goal was to identify best practices of group decision making with the AHP, which could then be used for the development of guidelines and in the practical assessment of frailty criteria with MACBETH.

The databases Web of Science and Scopus were searched for articles regarding AHP, MACBETH, MCDA, healthcare-related MCDA and group decision making with MCDA. Search

terms were defined as presented in table 1. After eliminating duplicates and inapplicable documents, the titles and abstracts were scanned. Studies were excluded if they were not available in neither the English nor the German language, if they were not concerned with MCDA and either AHP or MACBETH or group decision making as a focus or if they were mainly concerned with developing a new MCDA method. Furthermore, not yet deleted duplicates and articles without accessibility of a full text format were excluded. Due to the abundance of AHP applications, studies using fuzzy AHP were not taken into consideration. The quality of the articles was no exclusion criterion. A total of 116 articles were identified to be potentially relevant and were reviewed.

2.2 Results Literature Research

The results of the literature research are shown in table 1.

Table 1: Search terms and results of the literature research.

	Search terms	Hits Web of Science		Hits Scopus
		TOPIC	TITLE	TITLE-ABS-KEY
1	MACBETH	1,975	1,686	904
2	AHP	9,081	2,105	13,114
3	MCDA	781	89	965
4	MCDM	1,836	377	2,405
5	multi criteria	23,685	2,605	39,090
6	health	1,126,052	387,724	2,955,181
7	healthcare	73,257	17,704	180,884
8	Group decision	57,997	3,060	110,222
9	1 AND 2	5	1	7
10	1 AND (3 OR 4 OR 5)	31	3	47
11	2 AND (3 OR 4 OR 5)	1,408	29	1,862
12	1 AND 6	4	0	12
13	1 AND 7	0	0	1
14	2 AND 6	220	10	465
15	2 AND 7	39	4	66
16	1 AND 8	6	0	7
17	2 AND 8	732	49	843
18	(3 OR 4 OR 5) AND (6 OR 7)	1,424	39	2,524

The number of articles related to the AHP was higher compared to the number of articles regarding MACBETH. A total of 607 titles were screened, of which 412 articles were excluded with respect to the previously mentioned exclusion criteria. Of the 195 remaining articles, the abstracts were screened and after further elimination of 79 articles due to the exclusion criteria, 116 articles remained. These 116 articles were identified to be potentially relevant and were reviewed. Significant references of these articles were also examined for further information. An overview of the most important articles is given in appendix A.

2.2.1 AHP

General Information

The Analytic Hierarchy Process is a value measurement model[2] employed to derive ratio scales from judgements. It has been developed by Thomas L. Saaty in the 1970's to assess a finite set of alternatives in complex decision problems and to support the decision making process. The first step in the AHP is the structuring of the decision problem. Therefore a hierarchy is built, containing the decision goal on the highest hierarchical level, followed by the decision criteria, the sub-criteria (if applicable) and finally the alternatives on the lowest hierarchical level. An example of such a hierarchy is given in figure 1. [22, 14]

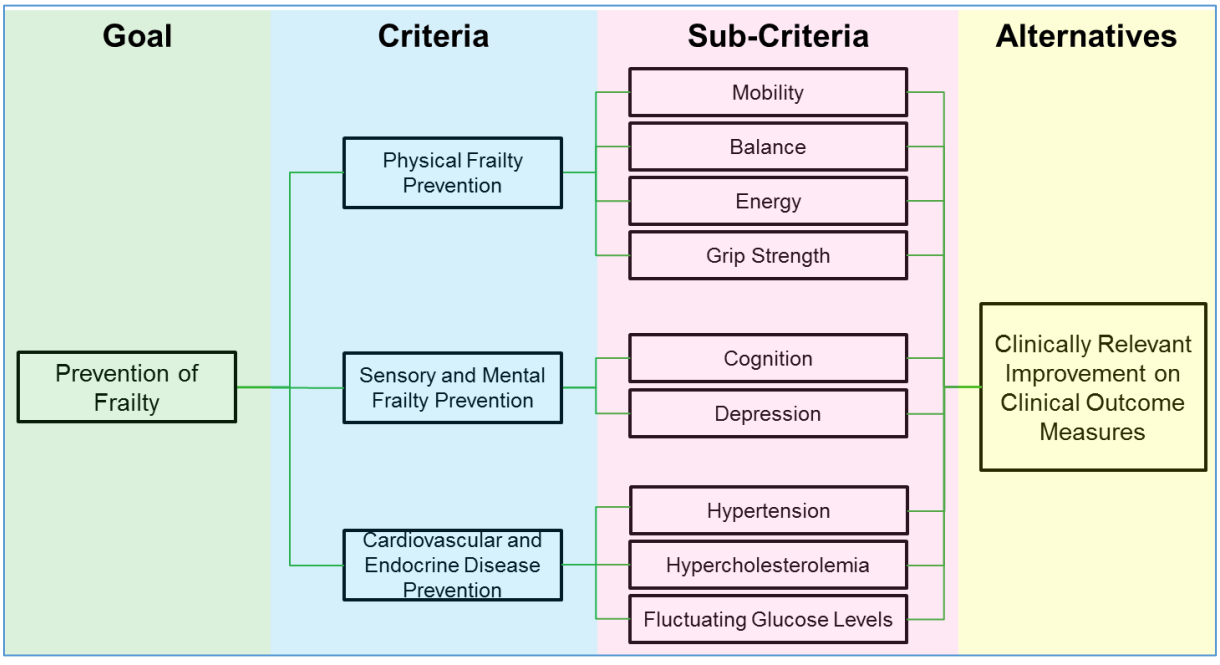


Figure 1: Example of a decision hierarchy in the AHP.

The choice of the decision criteria and the sub-criteria is crucial. They have to be well-defined, relevant, mutually exclusive and operational to avoid confusion and bias and to ensure comparability [2]. According to Bahurmoz [23], the number of criteria and sub-

[2] Value measurement models aim at assigning a real number or value to the alternatives and therefore providing a rank or preference order. [4]

criteria on each level should not exceed seven, to reduce the number of necessary pairwise comparisons in the next step. Miller [24] mentioned that seven ± two elements is the limit for individuals to process information, which was confirmed with respect to the AHP by Saaty and Ozdemir [25]. The number of necessary judgements can be calculated with the formula $n(n-1)/2$, with n being the number of criteria, sub-criteria or alternatives. In the hierarchy presented in figure 1, six pairwise comparisons would be necessary to assess "Physical Frailty Prevention", one pairwise comparison would be needed for evaluating "Sensory and Mental Frailty Prevention" and three judgements would be requested for the criterion "Cardiovascular and Endocrine Disease Prevention". Three additional pairwise comparisons concerning the criteria with respect to the goal "Prevention of Frailty" would finally be required, which would result in a total number of 13 pairwise comparisons without taking alternatives into consideration. [22, 14]

After the decision problem has been structured, pairwise comparisons are used to derive weights for the (sub-) criteria and priorities for the alternatives. This can be carried out in two different ways. In the Top-Down approach, the criteria are first assessed to gain weights, followed by the priorities of the different alternatives. The order of judgements in the Bottom-Up valuation is reversed. Two alternatives with respect to a specific criterion are evaluated at a time regarding the decision makers' preference. Then, two criteria of the same hierarchical level are assessed at a time with respect to their importance [26]. The original measurement scale for the judgements proposed by Saaty [27] is the fundamental scale, which reaches from 1 (equal importance) to 9 (extreme importance/9-fold higher importance). However, this scale has been discussed multiple times in the past and alternative scales have been evaluated and proposed. An overview of different scales that have been investigated and could be applied is presented in table 2.

Table 2: Overview of measurement scales in the AHP. Following Ishizaka, Balkenborg [28].

Scale	Definition	Parameters	Recommendation by author/s
Linear [27]	$c = a \cdot x$	$a > 0; x = 1,2,\dots,9$	Yes
Power [29]	$c = x^a$	$a > 1; x = 1,2,\dots,9$	No, Saaty's 1-9 scale preferred
Geometric [30]	$c = a^{x-1}$	$a > 1; x = 1,2,\dots,9$	Yes
Logarithmic [31]	$c = \log_a (x+1)$	$a > 1; x = 1,2,\dots,9$	Yes, recommended for high values
Root Square [29]	$c = \sqrt[a]{x}$	$a > 1; x = 1,2,\dots,9$	No, Saaty's 1-9 scale preferred
Inverse Linear [32]	$c = 9/(10-x)$	$x = 1,2,\dots,9$	Yes
Balanced [33]	$c = w/(1-w)$	$w = 0.5, 0.55, 0.6, \dots, 0.9$	Yes

Once a set of necessary pairwise comparisons is finished, a consistency check is executed. This step is necessary to detect contradictions in the judgements, which could be caused by several factors such as uncertainty, bounded rationality or missing concentration. Therefore a consistency index is calculated using the formula $Consistency\ Index\ (CI) = \frac{\lambda_{max}-n}{n-1}$, with λ_{max} being the maximal eigenvalue and n being the dimension of the matrix. A consistency index of 0 represents complete consistency. This consistency index is then used for the determination of the consistency ratio, which follows the formula $Consistency\ Ratio\ (CR) = \frac{CI}{RI}$. RI is the random index, which stands for the average consistency index of 500 reciprocal matrices randomly filled with values from the fundamental scale. The consistency ratio should be below the threshold of 0.1, which would mean 10 % inconsistency compared to the average inconsistency of the random reciprocal matrices. Otherwise it is necessary to check and revise the pairwise comparisons. [34, 4]

The next phase of the AHP is the calculation of criteria priorities and local alternative priorities. Criteria priorities or weights represent the importance of each specific criterion. Local alternative priorities show the preference of alternatives with respect to a certain criterion [34]. The original AHP by Saaty [27] uses the principal right eigenvalue method to derive these priorities. Therefore, the equation $A\omega = \lambda_{max}\omega$ needs to be solved, in which A represents the reciprocal comparison matrix, ω stands for the eigenvector/priority vector and λ_{max} is the principal eigenvalue. In a fully consistent matrix, the eigenvalue complies with the dimension n of the matrix, otherwise it is greater than n. The approach can be interpreted as a way of averaging and normalizing different possibilities of comparing the elements of a judgement matrix [29]. Nonetheless, other methods for this calculation have been proposed. Due to the large amount of criticism concerning Saaty's original principal right eigenvalue method, an overview of alternative approaches is given in table 3.

Table 3: Overview of different methods used to derive priorities.
Following Ishizaka and Lusti [35], Golany and Kress [36], Srdjevic [37].

Method	Supporter(s)
Principal Right Eigenvalue	Saaty [27]
Principal Left Eigenvalue	Johnson, Beine [38]
Modified Eigenvalue	Cogger and Yu [39]
Approximate Method/ Mean of Normalized Values	Saaty [40]
Geometric Mean/ Logarithmic Least Squares	Crawford and Williams [41]
Least Squares	Jensen [42]
Weighted Least Squares	Chu, Kalaba [43], Blankmeyer [44]
Logarithmic Least Absolute Values	Cook and Kress [45]

Once criteria and local alternative priorities are determined, global alternative priorities can be calculated using aggregation, to gain a ranking of the alternatives. Saaty [27] employs an additive value function for this synthesis. However, Belton and Gear [46], Holder [47], Holder [48] strongly criticised this approach due to the possibility of a rank reversal in case of an introduction of another option in additive models that include a normalisation step. To avoid this issue, two possibilities were introduced. Instead of the so-called distributive mode the ideal mode is recommended, if rank reversal is not accepted and alternatives might be added. Then normalisation is achieved "by dividing the score of each alternative by the score of the best alternative under each criterion" [34]. The other option is a multiplicative aggregation, which was proposed by Lootsma [49] and Barzilai and Lootsma [50]. [34]

The next step in the AHP is the interpretation of the results, followed by a sensitivity analysis. This last analysis is of high importance to investigate the impact of changes in criteria weights on the ranking of the alternatives. If the results are robust, recommendations about the preferred alternative can be made. A summary of the different steps in the AHP is provided in figure 2. [34]

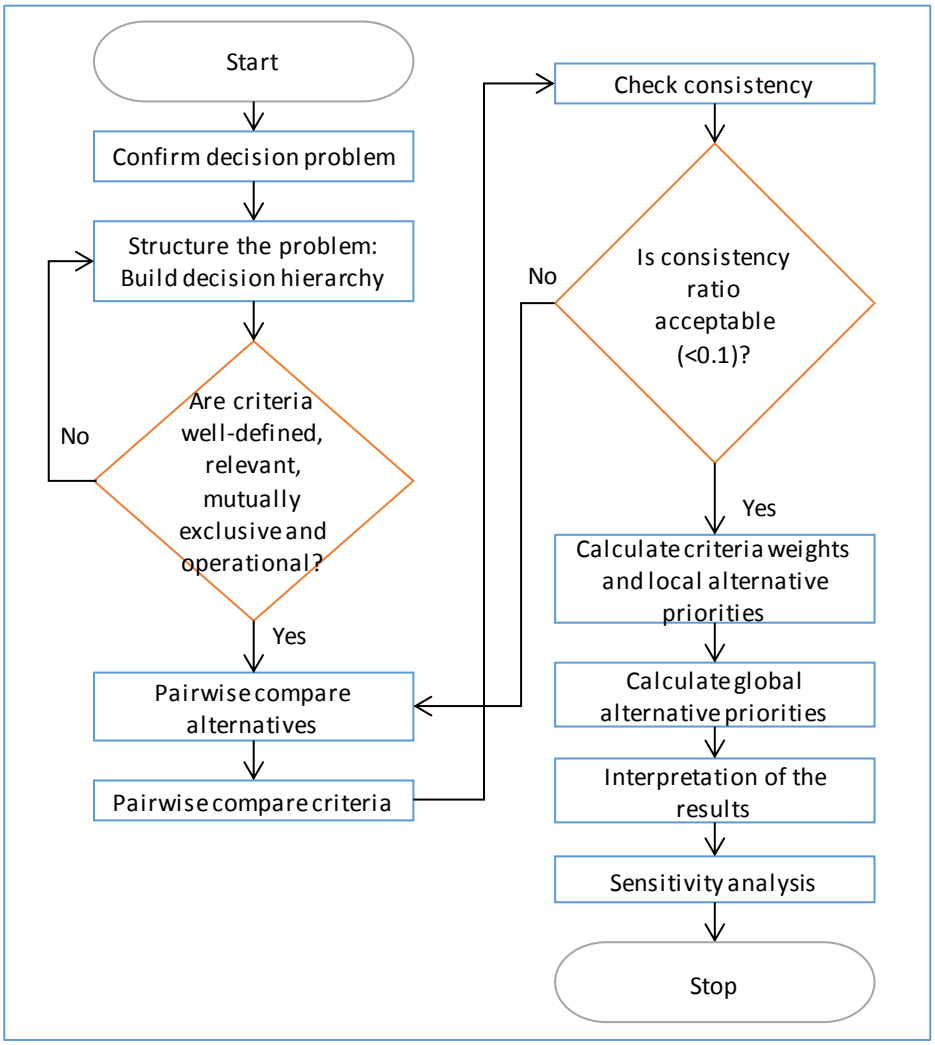


Figure 2: Summarizing flow chart of the AHP.

Group Decision Making

The AHP has been used several times for group decision applications. To ensure a structured, balanced and significant procedure, a facilitator is recommended to guide the decision makers and to encourage their participation during the debates [26]. Furthermore, the purchase of suitable software should be considered, which offers a user-friendly interface and alleviates the conduction and presentation of results [26]. Once these preliminary preparations have been organised, group decision making with the AHP can be employed in many different ways. The sharing approach from Belton and Pictet [13] aims at reaching a consensus between participants, if necessary by discussion and reduction of differences. The group members should act as one decision maker. The aggregating approach can be used, if a compromise can be reached by either a vote or an aggregation of individual judgements. Differences between opinions are not necessarily discussed. And finally the comparing approach acknowledges individual differences and uses them for a discussion, without necessarily reducing them.

All of these three approaches can be utilised with the AHP, but in most cases the disagreement between decision makers is rather high and therefore a mathematical aggregation is employed [51]. The software "Team Expert Choice" also follows this method. The aggregation of judgements can take part in two different ways during the group decision making. The first option is the calculation of a group average using the geometric mean after each pairwise comparison, which is then used to derive group weights and priorities [26]. The second option is the conduction of individual judgements and individual outcomes, the latter is then used to calculate aggregated results with the weighted arithmetic mean [26]. Other aggregation methods such as linear programming [52] and the Bayesian approach [53] have been proposed, but are not used on a regular basis.

In case of differing expertise between decision makers, it is possible to assign individual weights to the stakeholders to reflect the differences in knowledge. Decision makers can evaluate each other using pairwise comparison matrices to obtain these weights, based on the relative importance of the decision makers. Another possibility would be a supra decision maker, which is solely responsible for the individual stakeholder weights. The use of the consistency index as a measure for the expertise of the decision makers as proposed by Cho and Cho [54] has been criticised, due to the upper limitation of 9 in the fundamental scale, which can force inconsistency to arise. [34]

AHP in Healthcare

Due to the high complexity, impact, consequences and diversity, MCDA has been increasingly applied in the healthcare market [7]. A literature review by Marsh, Lanitis [8] investigated the use of MCDA applications in healthcare, with the result that in 26.8 % of the applications, AHP was the method used for weight elicitation. Especially in the assessment of

prescriptions and investments the AHP was largely adopted. Another literature review by Liberatore and Nydick [15] investigated the use of the analytic hierarchy process in medical and healthcare decision making. This article showed that the AHP is one MCDA method which has been used for a variety of healthcare applications, including medical diagnosis, patient participation, treatment options, organ transplantation, project and technology evaluation, human resource planning and healthcare policy. The high number of applications, the wide range of application areas and the steady use of the AHP in healthcare-related issues indicate the power of this method as a tool for multi-criteria decision analysis in the healthcare sector. However, the authors point out that although the AHP has been shown to be helpful and useful in shared decision making, the current acceptance of physicians is low. The scepticism seems to be related to formalized methods in general [15], and therefore to date many healthcare-related judgements and choices are still based on intuition or deliberate processes and therefore lack transparency and rationality [9].

Best Practices

Several tutorials and instructions can be found in the literature about the use of the AHP as an MCDA method. This part of the thesis will determine best practices of the AHP and recommendations with respect to group decision making and healthcare-related issues. Saaty [51] presented observations and suggestions about group decision making with the AHP and focused on the composition of the group, the decision-making session itself and the implementation of the results. He recommended to find participants with an equal responsibility and power, to avoid very strong members with a high influence on the other participants as well as people without any accountability and knowledge. Furthermore, Saaty proposed to either use a consensus vote or individual judgements aggregated with the geometric mean to fill out the pairwise comparison matrix, depending on the disagreement between decision makers. Finally, the group should evaluate the results and discuss possibilities for the implementation of the preferred alternative.

Dolan, Isselhardt [55] prepared a tutorial for the use of the AHP in medical decision making. They explained the typical steps in the AHP process and applied them to a medical decision problem. The structuring of the problem and the preparation of a hierarchy is mentioned to be of high priority and should be both sufficient and clear. During the so-called scoring step, where the pairwise comparisons are conducted, the importance to ordinarily judge about two elements using the fundamental scale is pointed out, even if an exact value is present. This is paramount due to the relative importance that is dependent on the decision maker. As an example, costs of 10,000 € compared to costs of 20,000 € for two alternatives would probably be considered of very strong importance to an individual. However, for a big private enterprise with a major annual revenue the costs would assumedly be of equal or weak importance. This shows that the verbal judgements are determined by the individual

perception and importance of the decision makers. In the third step, the weighting procedure, the authors describe four different possibilities to calculate the eigenvector in case of consistency. If the comparison matrix has a high number of inconsistencies, a different approximation method is proposed and the use of a suitable software is recommended. Lastly, two different methods are acknowledged concerning the sensitivity analysis. The “what-if” method is suggested in case of an analysis of the effects of changes in the judgements concerning the decision alternatives. In this case the comparison matrix would be reconstructed each time the input is varied, which would then result in recalculated weights and a change in the overall scores of the options. The proportional weight method is used for investigating a variation of the criteria weights. The weight of one specific criterion is varied and the other weights change with respect to their relative original values.

Hummel, Bridges [26] published a tutorial for group decision making in benefit-risk assessment with the AHP. In addition to a step-by-step guide for the use of the AHP in supporting group decision making, several recommendations can be found. The authors mention the advantage of having an independent facilitator, who motivates the panel members to partake in the discussions and who offers guidance in the different steps of the AHP. Supporting software is another factor that is recognised to ease the decision making process. Regarding the determination of the decision hierarchy structure, three possibilities are specified. A brainstorming group session is helpful for complex issues and unknown factors, a previous preparation of the hierarchy is favoured if criteria are already known from sufficient literature. A combination of both techniques includes an already developed hierarchy which can be discussed and modified in a subsequent group setting. With respect to the pairwise comparisons, J. M. Hummel et al. point out to measure all criteria positively, to ensure comparability. In addition, if large amounts of alternatives are assessed, values can be directly rated using “qualitative or quantitative intensity scales” [26] to avoid a very high number of pairwise comparisons. Furthermore, the authors recommend to use the bottom-up approach, which first evaluates the alternatives, followed by the criteria. The reason for this is the dependence of the decision criteria on the performance of the alternatives. In the overall prioritization of the alternatives, the method of synthesis is dependent on different factors. But in case of the comparison of new technologies with the gold standard, the previously explained ideal mode should be preferred for the synthesis. And finally, according to Hummel, Bridges [26] the examination of heterogeneity could be useful to detect relevant differences between subgroups.

In conclusion, the most important recommendations and best practices of the AHP that are relevant for this work are the introduction of a facilitator, an evenly composed group of participants, the discussion of the hierarchy, a questioning structure following the bottom-up approach, an aggregation of judgements through discussions or mathematical

aggregation, a thorough sensitivity analysis and an examination of heterogeneity between subgroups.

Discussion about the AHP

The AHP has been frequently applied in a variety of settings over the past 35 years [14]. It was appraised to be a very powerful and flexible tool that can be very useful for structuring a problem, for group decision making and for very complex decision problems [56]. Nonetheless, the AHP has been severely criticised by various authors in the past. This criticism is evaluated in this part of this work.

The first object of criticism is the possibility of rank reversal in case of the introduction of an additional alternative to an existing decision problem [46, 57]. Belton and Gear [46] suggested a modified normalisation to preserve the preference order. Saaty and Vargas [58] on the other hand, stated the acceptability of rank reversal. They mentioned that if a new alternative is added, it depends on the relationship between alternatives whether ranks are reversed, which is a natural development. Despite this legitimacy of rank reversal, Saaty [59] accepted Belton's and Gear's "revised AHP" and called it the ideal mode (explained in chapter 2.2.1, *General Information*).

Furthermore, the fundamental scale has received negative appraisal. Donegan, Dodd [60] claimed some ambiguity in the English verbal scale. Barzilai [61] asserted that preferences should not be measured using a ratio scale due to a missing absolute zero. Saaty [62] on the other hand, mentioned that a ratio scale is the only possibility for an aggregation of judgements. Only with a ratio scale, summation, subtraction, multiplication and division is possible. Furthermore Saaty stated that units of measurements cancel out in the common questioning procedure and therefore the absolute values would lead to an automatically derived relative ratio scale [63]. Other numerical scales have been proposed by a variety of authors, which have been previously mentioned in section 2.2.1, *General Information* (table 2). The fundamental scale from one to nine has yet been applied the most to date.

Priorities in the AHP can be derived in different ways. Originally, the right principal eigenvector was used by Saaty [27]. Due to a possibility of rank reversal related to the eigenvalue method, the geometric mean found many supporters and has been used several times to derive priorities [64]. Bana e Costa and Vansnick [65] analysed the eigenvalue method and mentioned that it violates the "Condition of Order Preservation (COP)"[3]. However, several authors reacted on this article with the claim that the COP is not justified

[3] The COP is an extension of the criticism about the rank reversal. „For all alternatives x_1,x_2,x_3,x_4 such that x_1 dominates x_2 and x_3 dominates x_4, if the evaluator's judgements indicate the extent to which x_1 dominates x_2 is greater that the extent to which x_3 dominates x_4, then the vector of priorities w should be such that, not only $w(x_1)>w(x_2)$ and $w(x_3)>w(x_4)$ […] but also that $w(x_1)/w(x_2)>w(x_3)/w(x_4)$" [65]

[66-68]. Different methods for the derivation of priorities have been indicated in section 2.2.1 *General Information* and an overview is given in table 3.

Another point of criticism is the consistency index by Saaty [27]. Karapetrovic and Rosenbloom [69] have shown that in several cases the consistency ratio was above 10%, although judgements were made logically and non-random. Bana e Costa and Vansnick [65] presented an example with judgmental inconsistency, which satisfied the consistency ratio with $0.03 < 0.1$. Some alternative methods have been developed to measure consistency or to discover inconsistent judgements, but the explanation of these would go beyond the scope of this thesis [70, 66, 71, 41, 72, 33, 73-75].

The synthesis of the priorities following Saaty [27] is determined through additive aggregation, also called the distributive mode. Due to the previously explained phenomenon of rank reversal in the distributive mode, the ideal mode and the multiplicative aggregation have been proposed by Belton and Gear [46] and Lootsma [49], respectively. Nonetheless, Vargas [76] presented an example to verify that in the case that weights are known, the additive aggregation is the only method to determine these exact weights.

As a conclusion, the majority of criticism is based on a missing adherence to axioms of utility theory[4]. Supporters of AHP on the other hand state that even though some of the axioms of the AHP are similar to the ones in utility theory, it is "historically and theoretically a different and independent theory of decision making from utility theory" [78].

Despite the criticism about different aspects of the AHP, this method is rather likely to be increasingly used in different applications in the future [64]. This might be motivated by the intuitive structure, the simplicity of the application and its flexibility, and the large amount of supporting software [64]. However, the discussion about the theoretical background of the AHP is ongoing. For that reason, the sub-chapter 2.2.3 will compare the AHP and MACBETH as two slightly similar yet very different MCDA methods.

2.2.2 MACBETH

General Information and Structure

MACBETH stands for 'Measuring Attractiveness by a Categorical Based Evaluation Technique'. It is another value measurement approach from the early 1990's that uses non-numerical judgements about the difference of attractiveness in pairwise comparisons to gain

[4] Utility theory assigns a utility function to a decision maker through preference relations and the adherence to specific axioms (for detailed information, see [77])

scores for options and weights for criteria in MCDA. The corresponding software M-MACBETH[5] is commonly used as a decision support system in several applications. [80-83]

The first step of the decision-aiding process with MACBETH is to structure the problem. This can be facilitated with a hierarchy or decision tree that includes the goal, multiple criteria and different options. It is important to mention that non-criteria nodes can be added for informative reasons but do not influence the decision, and that different criteria levels are not taken into consideration. [17]

The previously mentioned decision tree (see figure 1) could then be modified to the one presented in figure 3. In MACBETH, the non-criteria nodes are only added to categorize the frailty criteria into frailty domains, however, a specific assessment of them is not carried out.

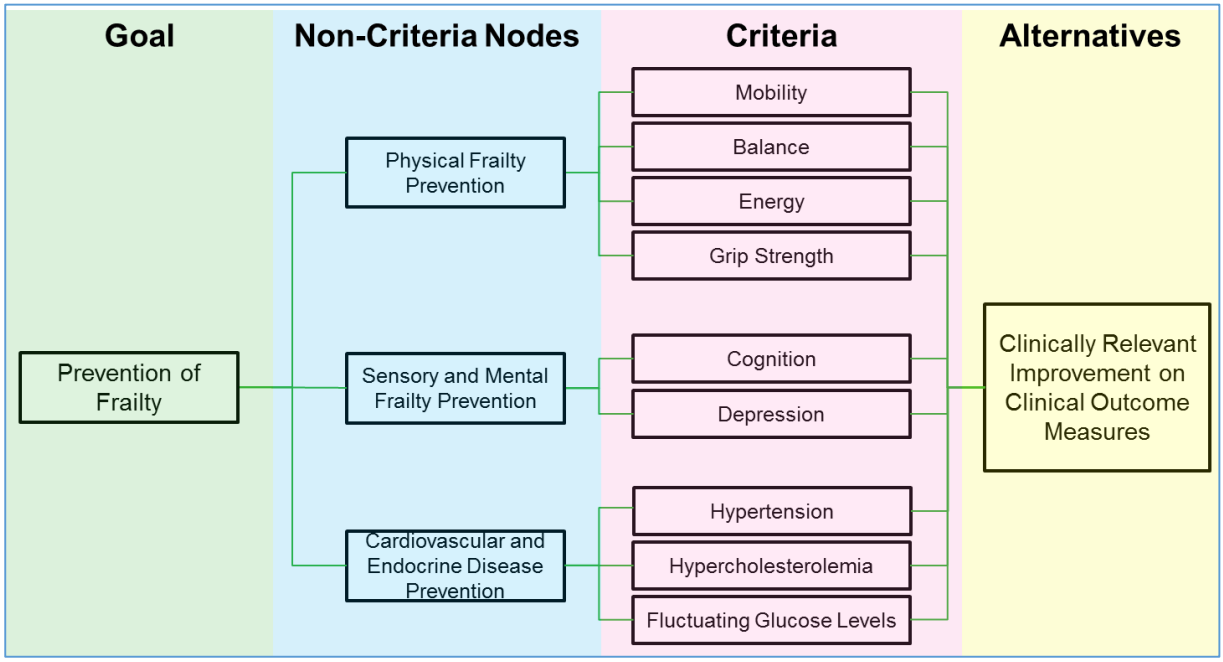


Figure 3: Example of a decision hierarchy in MACBETH.

Subsequently, the evaluation model is built by using pairwise comparisons. The decision maker is first asked whether there is a difference of attractiveness between two options with regard to a specific criterion, and if so, which of the two options is more attractive than the other. The answers provide the researcher with ordinal preference information. If one option is preferred, the decision maker is then asked to judge about the difference of attractiveness between these two options, which offers cardinal preference information. The decision maker can choose between the semantic categories “very weak”, “weak”, “moderate”, “strong”, “very strong” or “extreme” to express their judgement. However, a range of semantic categories in case of hesitation is also accepted. This questioning

[5] M-MACBETH is a decision support system developed by BANA Consulting, which uses the MACBETH approach to facilitate complex MCDA [79]

procedure is repeated until all options have been compared pairwise with respect to each of the criteria. A consistency check takes part simultaneously and if inconsistency is detected, judgements need to be revised. By means of linear programming (see sub-chapter 2.2.2, *Mathematical Background*), the qualitative judgements of the decision maker are then used to generate values on an interval scale. The least attractive option is grounded with a score of 0, and the most attractive option is given a score of 100. Reference levels such as "neutral" and "good" can be chosen to help the decision maker to gain a better understanding of the comparisons, but they are not necessary for the creation of a value scale. [84, 79, 85]

The next task is the weighting of the criteria, which is done with a similar questioning procedure. First, "good" and "neutral" reference levels have to be determined for each criterion. The decision maker is then asked to rank the criteria in order of importance of an improvement from neutral to good on each criterion. Next, the decision maker is asked to judge the importance of an improvement from neutral to good with respect to the semantic categories mentioned above on each criterion. The following step is to pairwise compare the importance of improvements from neutral to good on two criteria at a time using the semantic categories. By applying the same linear program as before, the weights are calculated, with 0 as the weight of the neutral option and a sum of weights of 100. It is important to mention that asking only about the importance of criteria is not sufficient, relations in value presented by a neutral and good level are needed to make an informed decision. [85]

The final part in the creation of the evaluation model is the determination of scores for each option, taken into account values in each criterion and the criteria weights. Nonetheless, the decision maker is able to adjust the priorities and weights within a certain range, which is dependent on the compliance with the constraints in the linear program. Once adjustments are executed, the overall score for the options is calculated with an additive aggregation model. The option with the highest overall score is most attractive for the decision maker and should be chosen. [17]

A subsequent sensitivity and robustness analysis is recommended, to explore the power and strength of the results. The sensitivity analysis enables to analyse the impact of a change in the weight of a specific criterion. The robustness analysis is particularly helpful, if uncertainty plays an important role in the decision making process. It evaluates whether the result of the best option changes in case of a variation in the weights up to a predefined percentage (e.g. ± 5 %). An overview of the different steps in MACBETH is provided in figure 4. [17]

With respect to group decision making, MACBETH and the software M-MACBETH originally employ the sharing procedure, which aims at finding a consensus between the decision

makers who act as one person [13]. However, lately MACBETH has also been used for so called decision conferences, where a common element is found through discussions and negotiations, therefore employing the aggregating approach mentioned by Belton and Pictet [13] [86].

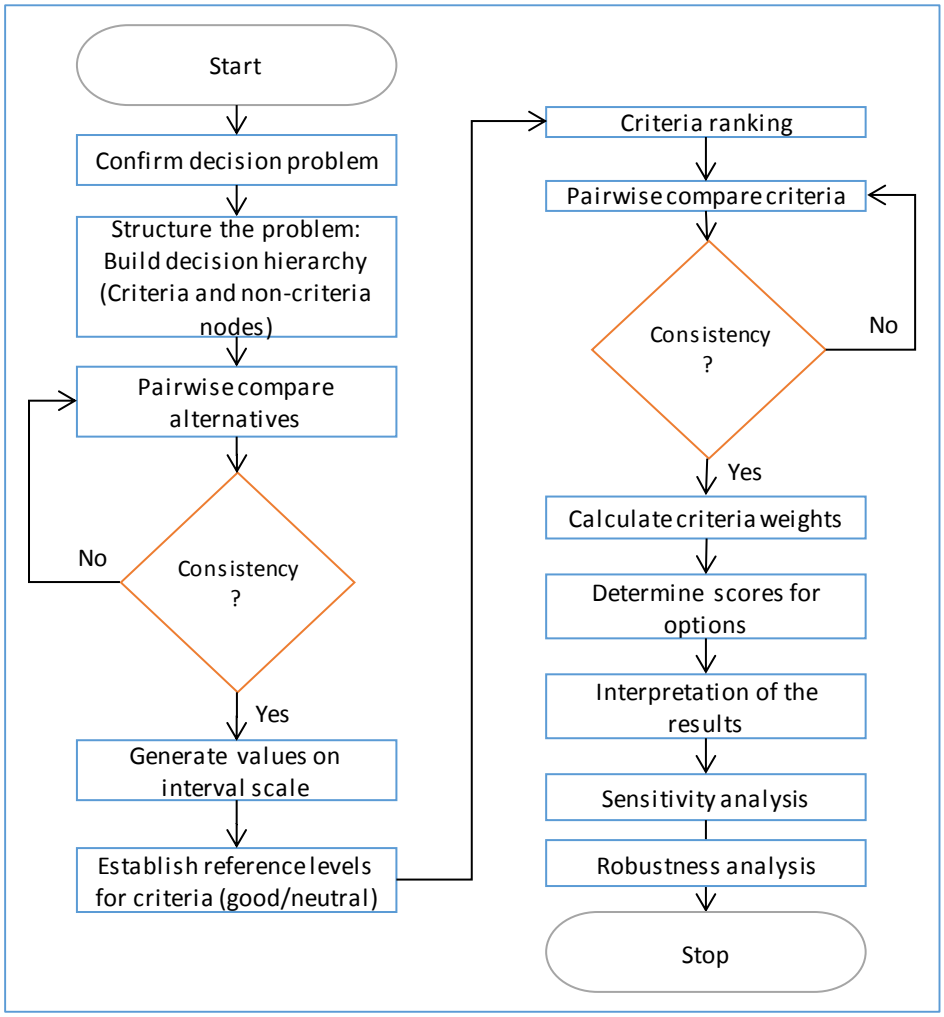


Figure 4: Summarizing flow chart of MACBETH.

Mathematical Background

This sub-chapter focuses on giving a short introduction towards the mathematical foundations of MACBETH once consistency is reached, following Bana e Costa and Vansnick [80], Bana e Costa, De Corte [20], Bana e Costa [85]. If a more detailed explanation is needed, Bana e Costa, De Corte [20] is recommended.

If the judgements are obtained as explained in the previous section and consistency is reached, the following linear program (LP-MACBETH) is solved in M-MACBETH to gain the basic MACBETH scale [20]:

$$\min x_1$$

subject to

$$(1)\quad x_p - x_r = 0 \qquad \forall\,(a_p, a_r) \in I \; with \; p < r$$

$$(2)\quad \sigma_i + \frac{1}{2} \leq x_p - x_r \qquad \forall\, i,j \in N_{1,Q} \; with \; i \leq j, \forall\,(a_p, a_r) \in C_{ij}$$

$$(3)\quad x_p - x_r \leq \sigma_{j+1} - \frac{1}{2} \qquad \forall\, i,j \in N_{1,Q-1} \; with \; i \leq j, \forall\,(a_p, a_r) \in C_{ij}$$

$$(4)\quad \sigma_1 = \frac{1}{2}$$

$$(5)\quad \sigma_{i-1} + 1 \leq \sigma_i \qquad \forall\, i \in N_{2,Q}$$

$$(6)\quad x_i \geq 0 \qquad \forall\, i \in N_{1,n}$$

$$(7)\quad \sigma_i \geq 0 \qquad \forall\, i \in N_{1,Q}$$

The x stands for the different scores assigned to the elements (different options or criteria), the σ represent the difference in attractiveness (also called thresholds or intervals) between two elements. The I stands for a binary relation of indifference between two elements, the $N_{1,Q}$ represents the range of non-negative natural numbers between 1 and Q, Q is the number of semantic categories, n serves as the number of variables and C_{ij} stands for the semantic categories which apply in case of a dominance relation.

The objective function of this linear program is the minimization of the score of the most preferred element x_1. The first constraint (1) concerns the scores in case of identity, which means that no difference in attractiveness exists between two elements. In this case, the difference in scores and the difference in attractiveness has to be zero. The second (2) and the third (3) equation stand for the dominance constraints. With element "p" being preferred to element "r", the order in the element rankings has to be preserved and the order between the thresholds must be maintained. Constraint number four (4) acts as an anchor for the intervals of the semantic categories, with 0.5 being arbitrarily assigned. Equation number five (5) symbolises the minimal sigma differences between the intervals and finally constraints six (6) and seven (7) are non-negativity constraints for the scores of the elements and the thresholds of the semantic categories.

As an example, suppose the following matrix has been filled out:

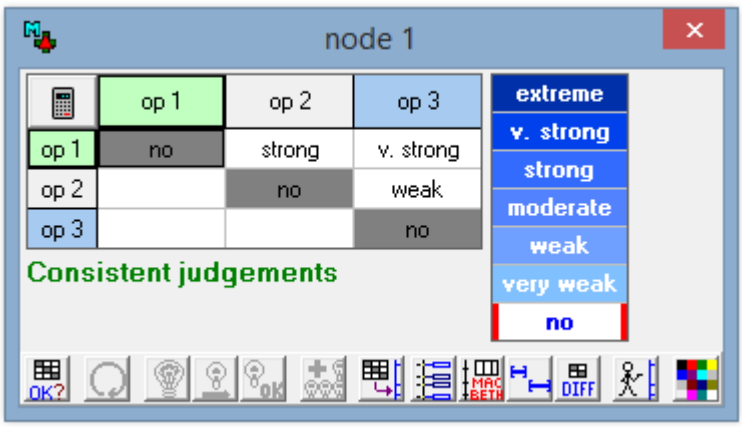


Figure 5: Example pairwise comparison matrix.

The constraints could be represented with the following inequalities:

Comparison op 1/ op 2:	$x_1 - x_2 - \sigma_4 \geq 0.5$	*Constraint 2*
	$-x_1 + x_2 + \sigma_5 \geq 0.5$	*Constraint 3*
Comparison op 1/ op 3:	$x_1 - x_3 - \sigma_5 \geq 0.5$	*Constraint 2*
	$-x_1 + x_3 + \sigma_6 \geq 0.5$	*Constraint 3*
Comparison op 2/ op 3:	$x_2 - x_3 - \sigma_2 \geq 0.5$	*Constraint 2*
	$-x_2 + x_3 + \sigma_3 \geq 0.5$	*Constraint 3*
Anchor:	$2\,\sigma_1 = 1$	*Constraint 4*
Minimal σ-difference:	$-\sigma_1 + \sigma_2 \geq 1$	*Constraint 5*
	$-\sigma_2 + \sigma_3 \geq 1$	*Constraint 5*
	$-\sigma_3 + \sigma_4 \geq 1$	*Constraint 5*
	$-\sigma_4 + \sigma_5 \geq 1$	*Constraint 5*
	$-\sigma_5 + \sigma_6 \geq 1$	*Constraint 5*
Non-negativity scores:	$x_1 \geq 0$	*Constraint 6*
	$x_2 \geq 0$	*Constraint 6*
	$x_3 \geq 0$	*Constraint 6*
Non-negativity thresholds:	$\sigma_1 \geq 0$	*Constraint 7*
	$\sigma_2 \geq 0$	*Constraint 7*
	$\sigma_3 \geq 0$	*Constraint 7*
	$\sigma_4 \geq 0$	*Constraint 7*
	$\sigma_5 \geq 0$	*Constraint 7*
	$\sigma_6 \geq 0$	*Constraint 7*

The result of this linear program would be a basic MACBETH scale with a score of 6 for option 1, a score of 2 for option 2 and a score of 0 for option 3, the anchored normalized MACBETH scale would be a score of $\frac{6-0}{6-0} \times 100 = 100$ for option 1, a score of $\frac{2-0}{6-0} \times 100 = 33.33$ for option 2 and a score of $\frac{0-0}{6-0} \times 100 = 0$ for option 3. In case of an assessment of criteria weights instead of option priorities, the same linear program LP-MACBETH is used, but an additional criterion "all lower" is entered, which is grounded to zero. The normalization at the end of the calculation then follows the formula $\frac{x_i}{\sum x_i}$.

2.2.3 Comparison AHP and MACBETH

Comparisons of the frequently used AHP and the less commonly applied MACBETH method have only rarely been conducted and if so without much detail. Bana e Costa and Vansnick [80] mention "significant fundamental differences" between both approaches. Ertay, Kahraman [87] obtained ranking orders of alternatives concerning renewable energy using both MACBETH and fuzzy AHP with the outcome that both methods yield the same results.

Salomon [88] gives an example for a cash flow analysis in which the results of MACBETH do not favour the most attractive option opposed to the same assessment with the AHP. Due to the major lack of thorough comparisons between MACBETH and AHP in literature to date, this section of the thesis will provide a detailed overview of similarities and differences.

Both the AHP and MACBETH are value measurement models based on pairwise comparisons. They follow a similar decision making process. First, the problem is structured by means of a decision hierarchy, then pairwise comparisons are conducted through absolute judgements and a comparison matrix is filled out. A consistency check takes part, priorities and weights are calculated and aggregated with an additive value function. Finally, the results need to be interpreted and a sensitivity analysis is recommended. At first, both methods seem to be fairly similar. However, if both methods are investigated in detail, some major methodological differences become apparent.

MACBETH is based on utility theory, which assigns a utility function to a decision maker through preference relations and the adherence to specific axioms (for detailed information, see [77]). Although some of the axioms of the AHP are similar to the ones in utility theory, it is "historically and theoretically a different and independent theory of decision making from utility theory" [78].

The decision hierarchies of the AHP and MACBETH slightly differ, if sub-criteria are considered. MACBETH distinguishes between criteria and non-criteria nodes and only assesses one level of criteria, whereas in the AHP sub-criteria are evaluated with respect to the specific criterion in the level above. This difference is accountable for a higher number of pairwise comparisons in the MACBETH procedure, if sub-criteria are considered. In addition, the AHP questioning procedure applies verbal judgements concerning a ratio of importance or priority, whereas absolute judgements in MACBETH are concerned with the difference of attractiveness between two options. [17, 34]

Another important point of distinction is the scale used for the pairwise comparisons. The fundamental scale from 1 (equal importance or preference) to 9 (extreme importance or preference) applied by Saaty is a ratio scale that associates a fixed number with every semantic category, which means that an extreme importance of one option over another is related to a 9-fold higher importance. MACBETH, by contrast, determines the numerical values of the six semantic categories on an ordinal scale during the linear programming procedure (see previous sub-chapter 2.2.2, *Mathematical Background*), which means they "are not a priori fixed" [80]. The semantic categories reach from "very weak difference in attractiveness" to "extreme difference in attractiveness", plus an additional expression of "no difference in attractiveness". This procedure in MACBETH, of constructing an interval

scale on basis of the semantic categories during the decision making process, is rather unique. [80]

The consistency in MACBETH is simultaneously checked with the software M-MACBETH during the completion of the comparison matrix. If inconsistency is detected, a warning appears and judgements need to be revised, or else a calculation of the MACBETH scale is impossible and priorities and weights cannot be generated. In the AHP, consistency is checked with the consistency ratio. Opposed to MACBETH, the AHP allows a consistency ratio of up to 10%, justified by an inconsistent human nature [89].

The calculation of priorities and weights in the original AHP by Saaty is achieved with the previously explained and highly discussed eigenvalue method. MACBETH uses a set of complex linear programs to gain these results. The AHP and the eigenvalue method in particular have been criticised by the developers of MACBETH with respect to the "Condition of Order Preservation" and the consistency ratio (previously explained in chapter 2.2.1, *Discussion about the AHP*).

In addition, priorities and the interpretation of criteria weights differ between the AHP and MACBETH. Priorities in MACBETH do not sum up to 1, opposed to the priorities in the classical AHP. However, if the ideal mode in AHP is used, the priorities would also not equate to 1 [90]. In the common MACBETH questioning procedure, the importance of a specified change from neutral to good in the performance is assessed [91]. Saatys' original AHP analysed the importance of one criterion compared to another with respect to the overall goal, but to date many applications have considered the importance of a change in the average performance of two criteria [92].

Besides these methodological differences, the AHP has been frequently applied in a variety of sectors, including the healthcare market, whereas MACBETH has been used predominantly in the fields of agriculture, manufacturing, services, energy, environment, human resource management and in the public sector [85]. Only very few applications used MACBETH in partially healthcare-related issues [93-96]. Moreover, AHP offers a decision support software called "Team Expert Choice" to support group decision making with this method, and MACBETH and the accompanying software M-MACBETH mainly focus on the sharing approach. Nonetheless, a low number of applications have been found in which decision conferencing was employed, used to negotiate and discuss judgements between several decision makers [97, 86, 98].

Finally, it should be mentioned that the AHP has received constant criticism about different aspects of this method, while the methodology of MACBETH seems to have been widely accepted. Only the choice of semantic categories proposed by MACBETH and the higher

number of pairwise comparisons have received negative comments [88]. An overview of the comparison of MACBETH and AHP is given in table 4.

Table 4: Overview of similarities and differences between AHP and MACBETH.

	Category	AHP	MACBETH
Similarities	MCDA model	Value measurement model	
	Structure	1. Decision hierarchy 2. Pairwise comparisons 3. Consistency check 4. Derivation of weights and priorities 5. Synthesis of weights and priorities 6. Interpretation of results 7. Sensitivity analysis	
Differences	Decision hierarchy	Sub-criteria evaluated with respect to criteria	Only one criteria-level is assessed
	Number of pairwise comparisons	Lower (in case of sub-criteria)	Higher (in case of sub-criteria)
	Judgements	Ratio of importance/priority/preference	Difference of attractiveness
	Scale	9-point fundamental scale (ratio scale)	6 semantic categories (ordinal scale)
	Consistency check	10 % inconsistency accepted	No inconsistency accepted
	Calculation of priorities and weights	Eigenvalue method	Linear programming
	Sum of priorities	Distributive mode: 1 Ideal mode: ≠ 1	In general: ≠ 1
	Interpretation criteria weights	Importance of criteria or importance of change in average performance	Importance of specified change in performance (e.g. of previously defined levels e.g. neutral and good)
	Application in healthcare	Frequently	Only in a few healthcare-related issues
	Group decision support software	Yes (Team Expert Choice)	No (Only negotiations/discussions possible with M-MACBETH)
	Criticism	Numerous criticism over the past decades	Only very limited criticism

2.2.4 Guidelines

With respect to the results from the literature review, the following guidelines were developed for the use of MACBETH as an MCDA tool for group decision making in the assessment of criteria related to the development of new medical products.

1. <u>Preparations</u>

The conduction of a panel session is recommended for group decision making with MACBETH in healthcare.

1.1. Analysis of the decision problem
1.2. Preparation of a decision hierarchy
  1.2.1. Determination of the overall goal
  1.2.2. Determination of alternatives
  1.2.3. Determination of criteria and non-criteria nodes (if applicable)
    - Criteria and non-criteria nodes should be determined by experts
    - Criteria should be well-defined, relevant, mutually exclusive and operational to avoid bias and to ensure comparability [2]
    - Choice of maximum seven ± two positively measured criteria [25]
  1.2.4. Choice of clinical outcome measure for each criterion, validation by experts
  1.2.5. Thorough discussion of the hierarchy, until a satisfying structure is found
1.3. Purchase of M-MACBETH software to ease the evaluation
1.4. Insertion of the decision hierarchy into M-MACBETH
1.5. Choice of a facilitator
    - Facilitator should be independent and experienced to support the group decision process [26]
1.6. Choice of panel members
    - Choice of 5-10 panel members to ensure the gain of sufficient information without the risk of having too complex and timely discussions
    - Equal knowledge/conditions/influence [51]
1.7. Preparation and provision of necessary information (semantic categories, criteria…)
    - This step should be done very thoroughly, to provide the panel members with a well-researched and equal amount of background information about the methodology and the topic
1.8. Choice of an audience response system for the conduction of individual judgements
    - Criteria for the choice of an audience response system: flexibility (input, number of possible answers), possibility to choose non-anonymous mode and to revise judgements, no additional hardware needed
1.9. Preparations in the audience response system concerning the project

2. Assessment

The panel session should be conducted in an atmosphere of trust to support the discussions.

2.1. Introduction by the facilitator
  2.1.1. Introduction of the facilitator and the topic (e.g. prevention of frailty in the elderly)
  2.1.2. Explanation of the structure of the panel session
  2.1.3. Introduction of MCDA
  2.1.4. Introduction of MACBETH and M-MACBETH
  2.1.5. Explanation of criteria and clinical outcome measures
2.2. Bottom-Up approach [26]
  2.2.1. Assessment of the alternatives (pairwise comparisons)
  2.2.2. Assessment criteria (pairwise comparisons)
2.3. Collection of individual judgement
2.4. Group discussion
  - During the group discussions, panel members should be motivated to participate in the debates by the facilitator, to ensure an even distribution of knowledge
2.5. Possibility for revision
2.6. Finding a consensus or choice of a semantic category/a range of semantic categories by application of the majority rule
2.7. Calculation of overall priorities/weights

3. Conclusion/Interpretation

3.1. Discussion of overall priorities/weights
3.2. Acceptance of overall priorities/weights
3.3. Conduction of a questionnaire to analyse the perception of the method, the atmosphere and the results by the decision makers
3.4. Dismissal of the decision makers
3.5. Sensitivity analysis: Proportional weight method
3.6. Robustness analysis
3.7. Analysis of consequences: Choice of an option

The feasibility of these guidelines or recommendations concerning the use of MACBETH in healthcare-related group decision making was then tested and revised according to the results from the practical assessments (see following section). An overview of the guidelines is presented in figure 6.

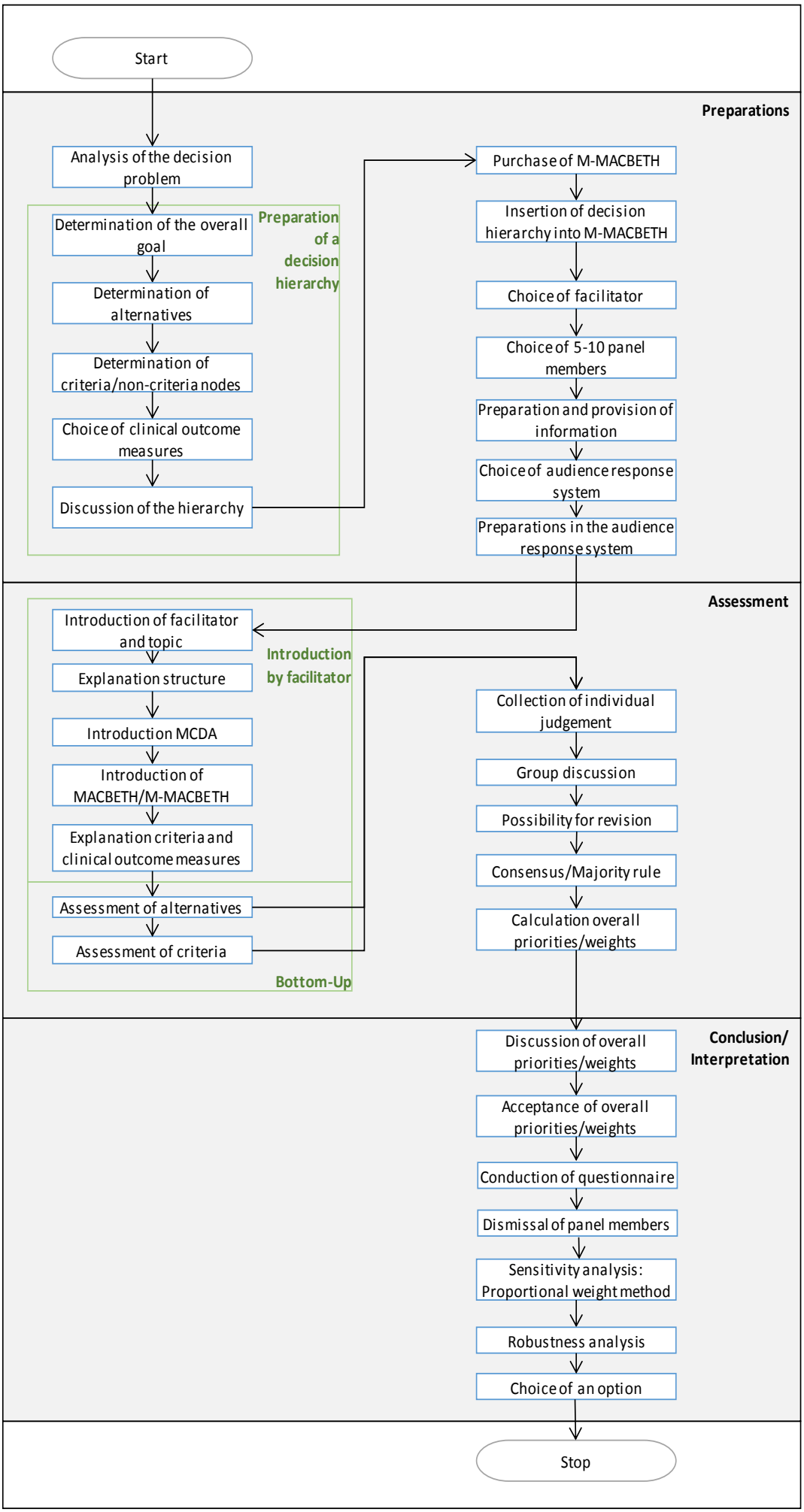


Figure 6: Summarizing flow chart of guidelines for MACBETH.

3. Practical Assessments

3.1 Methods

3.1.1 Preparations

The practical part of this study was used to test and evaluate the feasibility of the previously developed guidelines concerning healthcare-related group decision making with MACBETH. More specifically, the applicability of the guidelines in general, the group decision making process, MACBETH and M-MACBETH, upcoming inconsistencies and the sensitivity of the results were investigated. In addition, the purpose was to gain weights for criteria related to frailty in the elderly and to acquire information about the perception of MACBETH methodology and results by elderly individuals. On this account, a test panel session and a pilot panel session were conducted and a qualitative questionnaire was developed.

Prior to these assessments, Gooskens [99] determined relevant frailty criteria and a limited number of suitable outcome measures for each frailty criterion and validated them by experts (figure 7 and table 5). The former was achieved by means of thorough literature research with the focus on systematic reviews, the latter was carried out through interviews with four gerontology experts [99]. In addition, an information sheet, which explains the different clinical outcome measures, was prepared (see Appendix B).

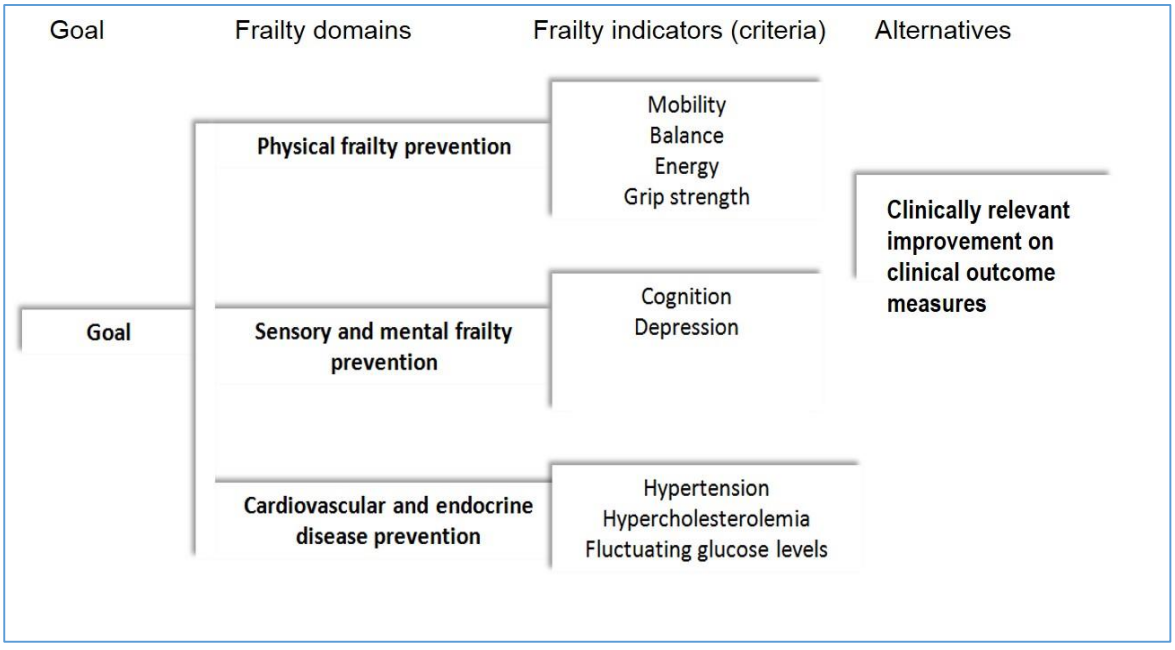


Figure 7: Decision structure for the prevention of frailty in the elderly.

Two patient profiles were created (see Appendix C). The first profile showed typical performances of a frail elderly person on all clinical outcome measures. The second profile presented a frail elderly person with clinically relevant moderate improvements on each

outcome measure. These profiles were generated to act as references in the assessment of outcome measures and frailty criteria. Furthermore, a thorough analysis of MACBETH was performed (see section 2.2.2) and the questioning procedure was set up in the online audience response tool "QuestionPress"[6].

Table 5: Overview of clinical outcome measures with respect to the frailty criteria.

Criteria	Clinical Outcome Measures
Mobility	Gait Speed (GS)
	Timed Up and Go (TUG)
	Flight of Stairs Questionnaire (FOSQ)
	Short Physical Performance Battery (SPPB)
Balance	Berg Balance Scale (BBS)
	Four Square Step Test (FSST)
	Accelerometric Measurements on Sensor Mat (AMSM)
	Chair Stand Test (CST)
Energy	Six-Minute Walk Distance (6MW)
	Subjective Fatigue Questionnaire (SFQ)
Grip Strength	Jamar Dynamometer (JD)
	Leg Extension Strength (LES)
Cognition	General Practitioner Assessment of Cognition (GPCOG) incl. Clock Drawing Test (CDT)
	Rey Auditory Verbal Learning Test combined with CDT (RAVLT)
	Standardized Mini-Mental State Examination (SMMSE)
Depression	Geriatric Depression Scale (GDS)
	Hamilton Rating Scale (HRSC/HAM-D17/7)
Hypertension	Blood Pressure (SBP+DBP)
Hypercholesterolemia	Cholesterol Level (CL)
Fluctuating Glucose Levels	Continuous Glucose Monitoring (CGM)

3.1.2 Test Panel Session

The first assessment, a test panel session, was carried out on 16th of May 2014. It was conducted with students and employees of the Health Technology and Services Research department at the University of Twente. The goal was to test the online tool

[6] QuestionPress is an online audience or classroom response system. It allows polls, surveys and assessments to be answered in real-time by an audience via internet. [100]

"QuestionPress" as an audience response system for the assessment of clinical outcome measures and criteria related to frailty in the elderly. Several audience response tools were inspected, but "QuestionPress" was chosen due to its high flexibility regarding the input and number of possible answers, the possibility to choose a non-anonymous mode, the chance to revise judgements and the fact that no additional hardware such as keypads is needed, only a computer or mobile device is necessary.

Four participants took part in the assessment. There were no specific in- or exclusion criteria, because the focus was on the user-friendliness and reliability of the online tool "QuestionPress". After a short introduction about the clinical outcome measures, the frailty criteria and the audience response system "QuestionPress", the moderator presented the first pairwise comparison and participants were able to vote for their preferred answer using "QuestionPress". The results of the voting were then directly shown on screen. Participants were able to revise their judgement if necessary, otherwise the moderator proceeded to the next pairwise comparison. The same procedure was followed during the assessment about the frailty criteria. Once several judgements and revisions were tested, participants of the test panel session were dismissed.

3.1.3 Pilot Panel Session

A pilot panel session was conducted on 14th of June 2014 in Aachen, Germany. This session was paramount for testing MACBETH and the feasibility of the previously developed guidelines with elderly participants. The feasibility of the guidelines was assessed regarding the following aspects: Acceptance, validity and sensitivity of the results, upcoming inconsistencies or problems, perception of the structure and methodology by elderly participants and time consumption. Furthermore, the pilot panel session was used to identify possibilities and recommendations for the enhanced use of MACBETH in healthcare. And finally the pilot panel session resulted in preliminary data about the weights of frailty criteria from the perspective of the elderly.

In- and exclusion criteria concerning the participants are represented in table 6. The goal was to find a minimum of 5 and a maximum of 10 participants for the pilot panel session as recommended in the guidelines. This number was chosen to ensure the gain of sufficient information concerning the weight of frailty criteria without having the risk of too complex and timely discussions.

Table 6: In- and exclusion criteria for participants of the pilot panel session.

Inclusion Criteria	Exclusion Criteria
Age: 65-80 years	Cognitive impairment
Knowledge of the German language	Dementia
Literacy	Poor overall health status
Resident of Aachen region (Germany)	
Availability on 14th of June 2014	

Previously to the pilot panel session, people complying with the in- and exclusion criteria (table 6) were approached in Aachen, Germany, regarding the participation in the pilot panel session. Some background information was given and six people agreed to partake in this study on 14th of June 2014. As preparation, one specific clinical outcome measure was chosen for each frailty criterion (selection see Appendix D). The clinical outcome measures were chosen due to the quality to represent the frailty criteria as well as the ease to explain the measures to elderly individuals. With respect to the results of the test panel session, it was decided that "QuestionPress" would not be employed due to the high age of the participants and to avoid difficulties and bias due to a lack of computer literacy. Instead, information sheets with the explanations about the chosen clinical outcome measures and a qualitative questionnaire were prepared in German and the decision tree was entered into the software M-MACBETH in German (see Appendices B, E and F). Furthermore, the pairwise comparisons and the semantic categories were translated into German and printed out for the participants. All translations were validated by a native English speaker.

On the day of the pilot panel session, participants (in the following also named panel members or decision makers) arrived at 2 pm. The moderator greeted the panel members and name badges were handed out to support the ease the communication during the discussions. After a short introduction about the course of action of the panel session, the moderator explained the frailty criteria and handed the information sheets to the participants. Each clinical outcome measure was demonstrated and explained. Participants were able to ask questions. Then, the two patient profiles were presented. Each improvement from patient profile 1 to patient profile 2 was explained. Questions by the participants were answered by the moderator. The pairwise comparisons were handed to the panel members, the name and age was requested and the questioning procedure and the semantic categories were explained. Participants were then asked to fill out the first round of the pairwise comparisons following the example in figure 8.

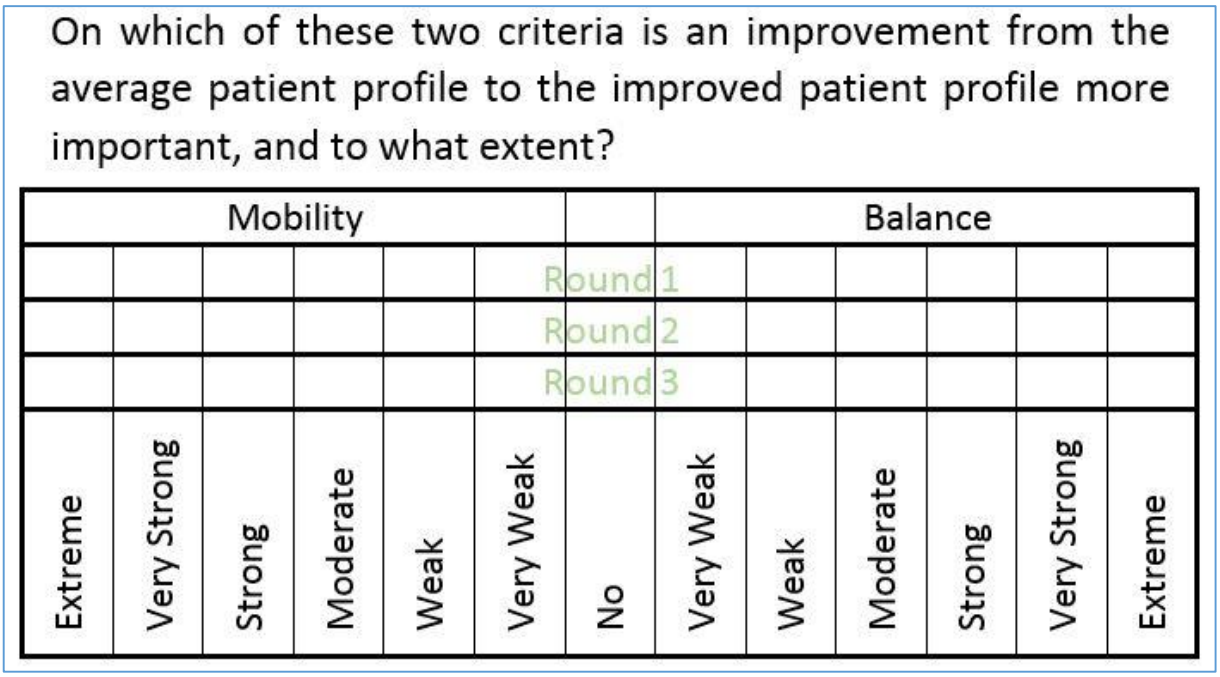


Figure 8: Example questioning procedure.

24 pairwise comparisons were requested. Once panel members made their judgements, the moderator collected every individual judgement by asking the participants clockwise and a discussion was initiated if a big discrepancy occurred. Participants were able to revise their judgements afterwards. Once the second round was over, the moderator collected the sheets and inserted the individual judgements into M-MACBETH. In the meantime, participants had a break. Once inconsistency was detected, the moderator presented the source of inconsistency to the corresponding participant and the participant decided about how to proceed. If an additional judgement was needed, the participant was asked to assess the specific pairwise comparison. Once all the weights were generated, the individual histograms were printed out and handed to the panel members. A discussion about the results was initiated and panel members were able to adjust their specific weights. Once the adjustments were made and every participant was satisfied with the individual weights, overall weights were calculated using the arithmetic mean. The overall weights were presented to the panel members and a discussion took part. In the end, the participants were asked to discuss the methodology and a qualitative questionnaire (see section 3.1.4) was handed out. After filling out the questionnaire, panel members were dismissed at 6:20 pm, after 4 hours and 20 minutes. The whole panel session was filmed to capture the arguments during the discussions.

3.1.4 Questionnaire

A qualitative questionnaire was prepared to receive information about the perception of the methodology and the conduction of the pilot panel session by elderly individuals. The questionnaire consisted of twelve open questions which were developed by the author with respect to the points of interest in this research. The first question concerned the choice of frailty criteria, the following two questions examined the atmosphere during the discussions. Question number four investigated the semantic categories, whereas questions number five and six dealt with the information provided and the work of the facilitator. In question seven the procedure of pairwise comparisons was assessed and in the questions eight and nine the expenditure of time and the user-friendliness of the MACBETH group decision making procedure were evaluated. Question ten was concerned with the validity of the results of the pilot panel session and question number eleven was used to gain information whether the MCDA assessment of criteria could be of value in the future. In question number twelve, participants were able to enter additional comments and thoughts about the pilot panel session and the MACBETH procedure. Participants were not obliged to fill out every question in the questionnaire.

3.2 Results

3.2.1 Test Panel Session

One female and three male individuals participated in the test panel session. Some alterable flaws were detected concerning the size of the font and the background colour. Besides these minor issues, the online tool “QuestionPress” showed to be reliable and valuable for the assessment of frailty criteria and clinical outcome measures. Therefore, the online tool “QuestionPress” was chosen to be used in the Nutricia expert panel session planned for September. Nonetheless, it was decided that “QuestionPress” would not be employed in the pilot panel session due to the high age of the participants and to avoid difficulties and bias due to a lack of computer literacy.

3.2.2 Pilot Panel Session

Five female participants and one male participant took part in the pilot panel session on 14th of June 2014 in Aachen, Germany. Their age was between 65 and 75 years, mean age was 68.5 years. The characteristics of the participants are displayed in table 7.

Table 7: Characteristics of panel members.

Characteristics	Gender	Age	Educational level	(Former) job
Participant 1	Female	66	General certificate of secondary education	Housewife
Participant 2	Female	67	General certificate of secondary education	Medical technical assistant
Participant 3	Male	72	Doctor's degree	Orthopaedic surgeon
Participant 4	Female	65	General certificate of secondary education	Receptionist at a general practitioner
Participant 5	Female	75	General certificate of secondary education	Bank assistant, afterwards housewife
Participant 6	Female	66	General certificate of secondary education	Hospital nurse

The rank order of frailty criteria differed among all participants, which resulted in a high discrepancy and a low level of revisions of individual judgements during or after the discussions. Although the discussions were lively, no consensus was found. Therefore the weights for frailty criteria were calculated individually using the M-MACBETH software. After every participant was able to adjust their weights, the overall weights were calculated using the arithmetic mean. These final individual and overall weights are shown in table 7 and figure 9.

According to the calculated overall weights, "Cognition" would be the most important frailty criterion with 17.68 % and a standard deviation (SD) of 7.51, followed by "Balance" (16.18 %, SD 6.31), "Energy" (16.18 %, SD 7.64), "Mobility" (13.07 %, SD 3.07), "Depression" (9.73 %, SD 8.01), "Hypertension" (9.54 %, SD 2.43), "Grip Strength" (8.87 %, SD 5.45), "Fluctuating Glucose Levels" (6.66 %, SD 4.07) and "Hypercholesterolemia" as the least important frailty criterion with 4.25 % and a SD of 5.81. However, individual weights and rank orders were rated rather differently. Four participants ranked the criterion "Cognition" to be most important, the remaining participants chose "Balance" and "Energy" to be most important. Also four participants chose "Hypercholesterolemia" to be the least important criterion, whereas the other two panel members chose "Depression" and "Grip Strengths" to be of lowest importance.

Table 8: Individual and overall weights of the frailty criteria.

Criterion	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5	Participant 6	Sum	Weights (Mean)	SD
Cognition	20.08	20.32	20.38	21.79	21.09	2.4	106.06	**17.68**	**7.51**
Balance	26.78	12.64	9.45	14.78	13.08	20.36	97.09	**16.18**	**6.31**
Energy	18.17	14.27	6.49	7.78	10.55	26.93	84.19	**14.03**	**7.64**
Mobility	14.35	15.92	12.93	13.23	14.77	7.2	78.4	**13.07**	**3.07**
Depression	3.84	12.63	17.4	19.84	0.43	4.22	58.36	**9.73**	**8.01**
Hypertension	8.15	8.25	9.95	6.62	13.5	10.76	57.23	**9.54**	**2.43**
Grip Strength	5.73	10.99	16.91	10.12	8.86	0.63	53.24	**8.87**	**5.45**
Fluctuating Glucose Levels	1.93	4.41	5.97	5.45	8.44	13.75	39.95	**6.66**	**4.07**
Hypercholesterolemia	0.97	0.57	0.52	0.39	9.28	13.75	25.48	**4.25**	**5.81**
Sum	**100**	**100**	**100**	**100**	**100**	**100**	**600**	**100**	

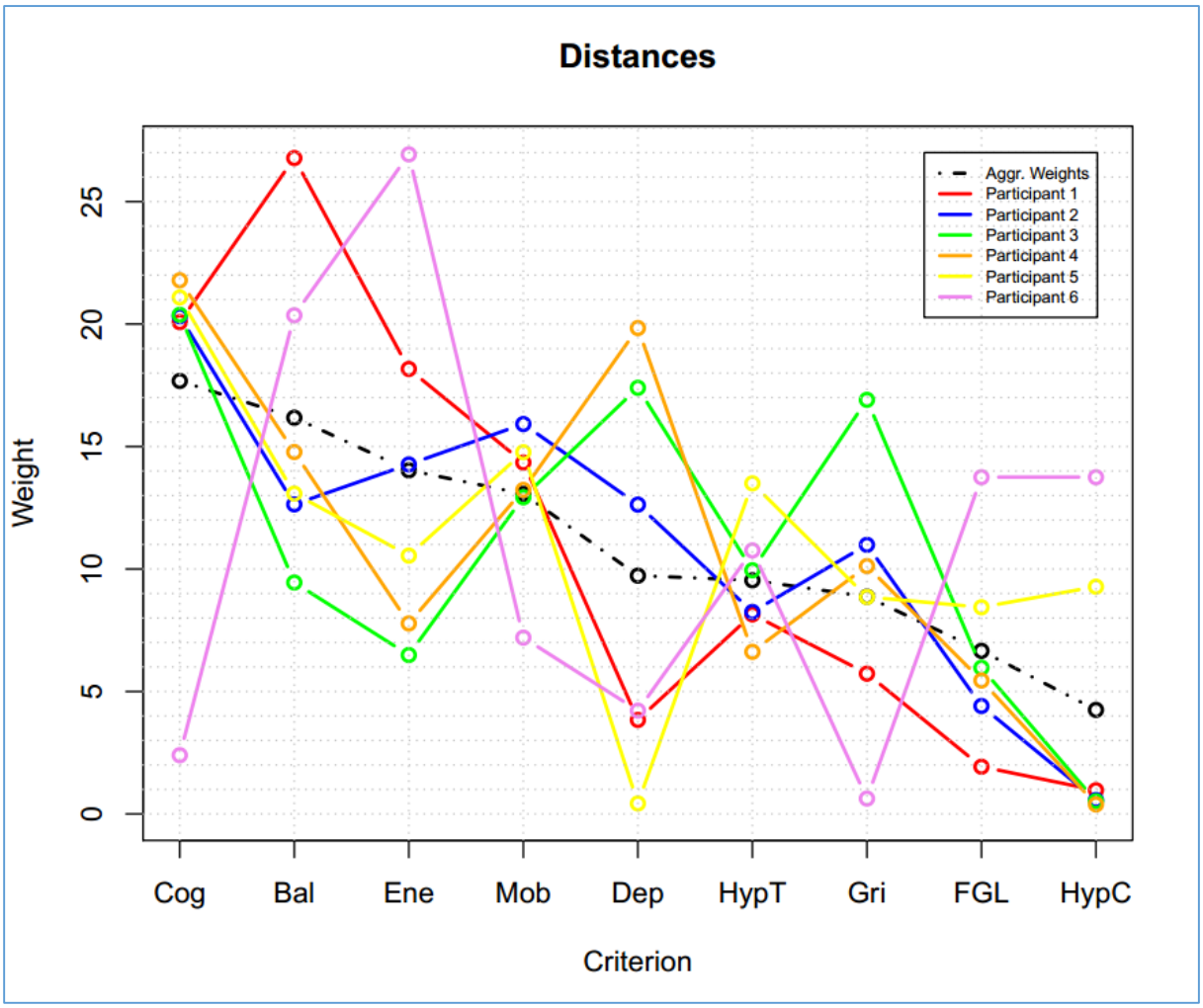


Figure 9: Individual and aggregated weights, presented in a line graph. Preparation in R. Abbreviations: Cog=Cognition, Bal=Balance, Ene=Energy, Mob=Mobility, Dep=Depression, HypT=Hypertension, Gri=Grip Strength, FGL=Fluctuating Glucose Levels, HypC=Hypercholesterolemia.

During the discussion at the end of the panel session, participants mentioned the feeling that if they were struggling with a specific issue, the corresponding frailty criterion got a higher weight. For example participant 1 stated to have balancing problems on a regular basis and therefore the criterion "Balance" was chosen to be of highest importance. Participant 6 specified in the discussion to have diabetes and that this would quite possibly be related to a higher weight of the frailty criterion "Fluctuating Glucose Levels". Although individual weights of the participants partially differed strongly with the calculated overall weights, five out of six participants accepted the overall weights and stated they were satisfied with the overall outcomes of the panel session. The remaining participant number 6 disagreed with the high weight of "Cognition" with the reason that elderly patients usually would not realise and worry about their cognitive state and therefore would not find an improvement in this criterion to be of high importance.

During and after the introduction, panel members asked several questions concerning the methodology. Some were unsure about the execution of pairwise comparisons, others had

problems understanding the clinical outcome measures used for assessing the frailty criteria. This was especially apparent among the oldest participants. In one instance the oldest participant stated “I am unable to do this, I am too old for this”. During the questioning procedure many panel members expressed doubts about staying consistent with the judgements and the high load of concentration needed for the panel session. After the first problems, participants were advised to prepare a rank order of the frailty criteria to ease the decision making process. There was an overall consensus that this was very helpful. During entering the individual judgements into M-MACBETH, inconsistencies were detected in each judgement matrix. Two types of inconsistency can be distinguished. The first type is called “no ranking” in M-MACBETH and prohibits a ranking of the elements [82]. This occurred in judgements of two participants. Participant number 1 had one “no ranking” inconsistency, participant number 5 had three inconsistencies of type one. The second type of inconsistency is called “inconsistent judgement”, in which a ranking by order of relative attractiveness is possible, but the semantic judgements conflict each other [82]. This type two inconsistency was detected in fifteen cases. Participants number 1, 2, 3 and 4 had two “inconsistent judgements” each that needed revision, participants number 6 had three “inconsistent judgements” and participant 5 had four inconsistencies of type two. The panel members with the highest amount of inconsistencies was the oldest participant, which also mentioned insecurity and a lack of self-confidence at the start of the assessment.

At the end of the panel session, participants discussed the different criteria. Most participants mentioned that the assessment would have been very difficult due to interdependencies between the criteria, such as “Mobility” and “Balance”, “Mobility” and “Hypertension”, or “Depression” and “Cognition”, “Mobility” or “Balance”. Furthermore, the semantic categories of MACBETH were debated intensively. Most participants found the number of categories too high and would have wished for only three or four categories as opposed to six. They were unsure whether the individual interpretation of the semantic categories would manipulate the outcome, until the facilitator explained that the actual range of each category is determined in the linear program in MACBETH. Participants found this information very helpful and interesting.

A sensitivity analysis using the proportional weight method would usually be recommended after the calculation of the weights and priorities. However, in the practical assessment at hand the alternatives were represented by the two patient profiles, which only acted as references. Since the improved patient profile would always be preferred, no additional information can be gathered. Nonetheless, the overall weights with respect to the individual weights can be investigated. If participant 1 would decrease the weight of criterion “Cognition” by 7% to 13.08% and increase the other weights proportionally, “Balance” would be the most important criterion with 16.57%. “Balance” would also be the most important criterion, if participant 2 or participant 4 or participant 5 would reduce their weights of the

criterion “Cognition” by 8%. In the case that participant 3 would reduce the weight of “Cognition” by 8%, “Balance” and “Cognition” would both be the most important criteria with 16.34% each. If participant 6 would lower the weight of criterion “Cognition” to zero, the overall ranking would not be affected. With respect to the lowest rated criterion “Hypercholesterolemia”, “Fluctuating Glucose Levels” would become the criterion with the lowest importance only if participant 1 would increase the importance of “Hypercholesterolemia” by 15%. For participant 2, participant 3, participant 4 and participant 5 an increase of 14%, for participant 6 an increase of 13% would be necessary, to change the rank of “Hypercholesterolemia”.

Summing up the results of the criteria weights, “Cognition” was considered the most important criterion in the assessment by most participants, whereas “Hypercholesterolemia” was rated to be of lowest priority. As a conclusion about the methodology, the pilot panel session has shown that an assessment of frailty criteria with MACBETH using the previously developed guidelines is feasible in a group setting without any major complications. Nonetheless, some important modifications of the guidelines are necessary with respect to the group decision making process. If the discrepancy between individual judgments is too high and no consensus can be reached, the calculation of individual weights with a subsequent calculation of overall weights using the arithmetic mean seems to be recommended. These aggregated results were accepted by five out of six participants. Inconsistencies occurred, but were detected and eliminated and the methodology was applicable with elderly individuals. A thorough explanation of the MACBETH method, the criteria and the clinical outcome measures is needed for the participants to understand the questioning procedure and to ease the decision making. For this reason it appears to be necessary to schedule sufficient amount of time for the introduction as well as the evaluation phase of the panel session. Finally, the number of semantic categories has been criticised and should be the object of further research. With respect to the points of criticism discovered in the assessment, a modified version of the guidelines was developed.

1. Preparations

The conduction of a panel session is recommended for group decision making with MACBETH in healthcare. The preparation phase should require approximately two weeks.

1.1. Analysis of the decision problem
1.2. Preparation of a decision hierarchy
    1.2.1. Determination of the overall goal
    1.2.2. Determination of alternatives
    1.2.3. Determination of criteria and non-criteria nodes (if applicable)
        - Criteria and non-criteria nodes should be determined by experts

- Criteria should be well-defined, relevant, mutually exclusive and operational to avoid bias and to ensure comparability [2]

1.2.4. Choice of maximum seven ± two positively measured criteria [25]

1.2.5. Choice of clinical outcome measure for each criterion, validation by experts

1.2.6. Thorough discussion of the hierarchy, until a satisfying structure is found

1.3. Purchase of M-MACBETH software to ease the evaluation

1.4. Insertion of the decision hierarchy into M-MACBETH

1.5. Choice of a facilitator

- Facilitator should be independent and experienced to support the group decision process [26]

1.6. Choice of panel members

- Choice of 5-10 panel members to ensure the gain of sufficient information without the risk of having too complex and timely discussions
- Equal knowledge/conditions/influence [51]

1.7. Preparation and provision of necessary information (semantic categories, criteria…)

- This step should be done very thoroughly, to provide the panel members with a well-researched and equal amount of background information about the methodology and the topic

1.8. If panel members have sufficient computer literacy: Choice of an audience response system for the conduction of individual judgements and preparation of the assessment in this audience response system

1.9. If panel members have no or insufficient computer literacy: Preparation of the pairwise comparisons as hardcopies

2. <u>Assessment</u>

The panel session should be conducted in an atmosphere of trust to support the discussions. The assessment phase should require approximately three to four hours.

2.1. Thorough introduction by the facilitator

2.1.1. Introduction of the facilitator and the topic (e.g. prevention of frailty in the elderly)

2.1.2. Explanation of the structure of the panel session

2.1.3. Introduction of MCDA

2.1.4. Introduction of MACBETH and M-MACBETH

2.1.5. Explanation of criteria and clinical outcome measures

2.2. Advice for panel members to first prepare a rank order of the options/criteria

2.3. Bottom-Up approach [26]

2.3.1. Assessment of the alternatives (pairwise comparisons)

2.3.2. Assessment criteria (pairwise comparisons)

2.4. Choice of method for group decision making

2.4.1. If agreement between decision makers is high: Finding a consensus
2.4.2. If agreement between decision makers is low: Mathematical aggregation of individual judgements with arithmetic mean

2.5. Consensus:
2.5.1. Collection of individual judgement
2.5.2. Group discussion
- During the group discussions, panel members should be motivated to participate in the debates by the facilitator, to ensure an even distribution of knowledge

2.5.3. Possibility for revision
2.5.4. Finding a consensus or choice of a semantic category/a range of semantic categories by application of the majority rule
2.5.5. Calculation of overall priorities/weights

2.6. Mathematical aggregation:
2.6.1. Collection of all individual judgements
2.6.2. Group discussion
- During the group discussions, panel members should be motivated to participate in the debates by the facilitator, to ensure an even distribution of knowledge

2.6.3. Possibility for revision
2.6.4. Calculation of individual priorities/weights
2.6.5. Discussion of individual results
2.6.6. Possibility for adjustments
2.6.7. Calculation of aggregated priorities/weights using the arithmetic mean

3. <u>Conclusion/Interpretation</u>

The final discussion is expected to require thirty minutes to one hour. The subsequent analyses is expected to take approximately two to three hours.

3.1. Discussion of overall priorities/weights
3.2. Acceptance or rejection of overall priorities/weights
3.3. Conduction of a questionnaire to analyse the perception of the method, the atmosphere and the results by the decision makers
3.4. Dismissal of the decision makers
3.5. Sensitivity analysis: Proportional weight method
3.6. Robustness analysis
3.7. Analysis of consequences: Choice of an option

The flow chart of the modified guidelines can be found in appendix H.

3.2.3 Questionnaire

The questionnaire (appendix F) showed that every participant felt comfortable to answer the questions individually and that nobody felt manipulated or suppressed. Three participants would have wished for additional criteria to be evaluated, namely “Capacity of Reaction”, “Independence in Everyday Life” and more criteria concerning physical deficiencies. The criticism about the scale being too comprehensive was also addressed in the questionnaire, with only one participant stating the scale being “just right”. With respect to the facilitator, participants mentioned that enough information was provided to make the judgements and that the facilitator was professional, motivating and gave the same attention to all participants. The methodology of pairwise comparisons was not judged uniformly. Two panel members said it would be very difficult to pairwise compare the criteria, two other panel members liked it and one of them mentioned that it would reduce the cognitive load. The user-friendliness was evaluated to be good by two panel members and three panel members said it was not easy to understand, difficult for elderly people and that a comprehensive introduction is needed. However, the expenditure of time was rated to be fine by every participant and most participants thought that the procedure of MACBETH would be valuable for future assessments. One participant wrote in the questionnaire that the whole evaluation seems to be difficult for people of higher age, that criteria dependencies are difficult to take into account and that the clinical outcome measures would not directly represent what most people would understand of the specific criteria.

3.2.4 Summary of Findings

With respect to the different aspects investigated to test the feasibility of the developed guidelines, the pilot panel session has shown that the introduction about MCDA and MACBETH is of major importance, if participants are of a higher age. In general, criteria should be well chosen and independent in MCDA, to avoid issues concerning the pairwise comparisons. Several cases of inconsistency have arisen during the session, however, all of them could be resolved. In group decision making with MACBETH, the arithmetic mean of the individual weights was applied, because no consensus could be reached. The results were accepted by five of the six panel members. The pairwise comparisons and the MACBETH methodology was perceived unequally by the participants, two found it user-friendly and three mentioned it was difficult to understand. With four hours twenty minutes, the pilot panel session can be evaluated as time-consuming, but participants approved of the expenditure of time. Due to these findings and with respect to the modifications, the guidelines can be assessed as feasible.

4. Discussion

4.1 Findings and Interpretations

Although MACBETH has been widely accepted in the MCDA-community, it has only rarely been applied to healthcare-related decision making. The focus of this work was therefore the development of guidelines for the use of MACBETH as an MCDA method for group decision making in the assessment of criteria related to the development of new medical products.

On this account, a thorough literature study was performed to gain information about MACBETH, the frequently used AHP and its best practices in group decision making and healthcare-related MCDA. In addition, similarities and differences between AHP and MACBETH were investigated. To the knowledge of the author, until this point, no detailed and objective comparison of these approaches had been carried out. Although the AHP has received much criticism in the past, it is still the method that has been applied the most in healthcare-related MCDA. The reasons for this extensive use should be subject to further investigations.

Using the results of the literature study, guidelines for the application of MACBETH in group decision making in healthcare were developed. It was interesting to recognise that even though the AHP and the MACBETH methodology have several major differences, it was possible to apply best practices of the AHP in these recommendations. The feasibility of these guidelines has been tested in a practical assessment of frailty criteria with elderly individuals. The assessment has shown that the guidelines are feasible for applying MACBETH in group decision making for the assessment of criteria related to the development of new medical products. In addition, difficulties were detected and the guidelines were modified (see section 3.2.2 and appendix H). It is important to mention that it is not possible to calculate the group average using the geometric mean after each pairwise comparison with MACBETH. This is based on the variable nature of the semantic scale, which is dependent on the individual perception of the semantic categories by the decision makers. The range of the categories is determined in the linear program at the end of the scoring step. Instead, the only possibility to follow the mathematical aggregating approach in group decision making with MACBETH is to employ the arithmetic mean at the end of the assessment by aggregating the individual outcomes. It seems to be beneficial to apply this approach in group decision making with decision makers that are unlikely to reach a consensus. A subsequent discussion about the acceptance or rejection of the overall weights and priorities appears to be of major importance. In the practical assessment of this work, five out of six panel members agreed to the overall weights and accepted it as an

adequate compromise. The participant who did not agree to the overall weights also had the highest total and Euclidian distance (see violet line in figure 9 and Appendix I).

Besides gaining information about the quality of the previously developed guidelines, the pilot panel session resulted in valuable data about the relative importance of different frailty criteria evaluated by elderly individuals. Although individual weights differed strongly among participants, five out of six decision makers accepted the overall weights and ranking of the frailty criteria. The results show that an improvement in “Cognition” was chosen to be of highest importance, followed by an improvement in “Balance”, “Energy” and “Mobility”. These four criteria are almost responsible for 61% of the weights. Therefore future research about the development of medical products should focus on nutritional compounds that improve frailty with respect to “Cognition”, “Balance”, “Energy” and “Mobility”.

The modified sensitivity analysis has shown that rather radical changes would be needed before the criterion “Cognition” would be ranked second instead of first in the ranking of relative importance. This criterion can therefore be described as robust. The same applies to the ranking of “Hypercholesterolemia” as the criterion with the lowest relative importance. This criterion can also be assessed as robust. It is interesting to see that criteria of the domain “Cardiovascular and Endocrine Disease Prevention” were assessed to be of lower importance. This could be related to an underestimated risk of chronic diseases such as diabetes, if people are not directly affected [101].

During the discussions in the pilot panel session, it became apparent that the assessment demands a great cognitive load, which is a significant burden for participants of high age. Therefore, future evaluations should consider to either focus on panel members below the age of 70 or to provide carers that can help elderly individuals if needed. In addition, a comprehensive introduction and competent responses towards upcoming questions are of major importance. The patient profiles should be provided at all time, to ensure an equal understanding of the frailty criteria at hand. The scale was also subject of the discussions. Participants criticised the large amount of options and mentioned to have preferred less semantic categories. They stated to have doubts about staying consistent using the scale and that a choice was difficult. Although the possibility to choose a range of categories for the pairwise comparisons was mentioned twice during the panel session, participants did not select this possibility. One would have expected that ranges would be selected in case of difficulties of choosing an option, but this was not the case. This could have several reasons. Either panel members wanted to make a clear decision for their own process during the scoring step, or the provided hard copies with separate boxes for each judgement prevented the choice of a range. In future assessments, this should be taken into consideration.

The qualitative questionnaire was used to investigate the perception of the MACBETH methodology and to determine recommendations for the broader use of MACBETH in healthcare-related group decision making. Two participants had suggestions for missing criteria, namely “capacity of reaction” and “independence in everyday life” and one participant was missing additional criteria regarding “physical deficiencies”. “Capacity of reaction” and “independence in everyday life” are very much dependent on the previously defined criteria and would therefore not be added in further assessments. Additional criteria regarding “physical deficiencies” were not specified and could therefore not be taken into consideration.

Moreover, the questionnaire showed that some participants were satisfied with the procedure and found it user-friendly, whereas others had problems and found the concept and the procedure hard to understand. These different perceptions could be grounded in the different age as well as unequal levels of education. During the discussions some participants mentioned their former jobs and educational background and a correlation of these factors and the understanding of the MACBETH procedure could be possible. However, additional research would be needed to investigate this relationship. In addition, it is likely that the AHP procedure would have received analogue feedback, as a result of the pairwise comparisons and a similar questioning procedure.

With respect to the semantic categories, as mentioned before, participants would have preferred a smaller scale. Only one participant said the scale would be “just right”. Due to the fact that the linear program behind MACBETH is very complex and cannot easily be replicated, it would be an advantage if M-MACBETH would incorporate an option to choose a preferred scale for the judgements. Salomon [88] also criticised the fixed scale. It would hence be likely that a modification of this factor would make the M-MACBETH software and hence the MACBETH procedure more attractive for healthcare-related group decision making in the future.

It should be mentioned that the pilot panel session for the assessment of frailty criteria took four hours and twenty minutes, including the questionnaire. This can be assessed as rather time-consuming. In particular, the necessity to fill in individual judgements manually into M-MACBETH if no consensus can be found and aggregate the results separately is intricate. Although it is expected that an evaluation with experts would take less time and participants in the pilot panel session approved of the expenditure of time, this could be a reason for decision makers not to choose MACBETH.

4.2 Strengths

This work provides a comprehensive explanation of the AHP and MACBETH and compares both approaches with much detail. To the authors' knowledge, until this point, no comparable analysis about similarities and differences between these approaches had been performed. The results contribute to a thorough understanding of both methods and help identifying the preferred procedure for different applications. Furthermore, the mathematical foundations of the linear program used in MACBETH were described by means of a simple example, which will help decision makers to understand the concept behind MACBETH without the need of a broad understanding of mathematics.

MCDA has shown to be very useful in complex decision problems. Despite its criticism, the AHP has been widely applied in the field of healthcare. This work and the developed guidelines in particular, will ease and most likely enhance the use of MACBETH for future group decision making. The application of MACBETH in healthcare-related multi-criteria group decision making has been absolutely unique and will act as a reference for future assessments. Moreover, the collected data concerning the weights of frailty criteria will be of value for Nutricia as well as other researchers in the field of gerontology. The information gained in this research can be used for future panel sessions regarding clinical outcome measures and criteria related to the prevention of frailty in the elderly.

This was the first instance in which the arithmetic mean was employed for the determination of aggregated weights in group decision making with MACBETH. It has shown to be successful and extremely beneficial for heterogeneous groups where no consensus can be reached. Until now, a consensus between decision makers was essential for making group judgements and the new insights could enhance the applicability of MACBETH for group decision making tremendously. Nonetheless, future research is necessary to validate the results of this aggregation.

4.3 Limitations

In spite of the undeniable value of this research, this work has some limitations. There is a possibility that even though the literature study was performed with much care, important information might have been missed due to missing keywords in the search. E.g. the long versions of AHP and MACBETH ("Analytic Hierarchy Process" and "Measuring Attractiveness by a Categorical Based Evaluation Technique") were not specifically searched, which could have resulted in additional articles.

The aim of the pilot panel session was to test the guidelines for group decision making with MACBETH in the development of new medical products. For this reason, clinical outcome measures were chosen for each frailty criterion. The choice of clinical outcome measures was made with respect to their quality of representing the frailty criteria as well as the ease to explain the measures to elderly individuals. The appointed clinical outcome measures might have influenced the outcomes of the assessment and probably a different selection would have resulted in differing weights. This was also supported by one of the participants during the panel session, who was a retired doctor. The individual at hand also mentioned differences in the patient profiles, namely the improvement of higher significance in the Geriatric Depression Scale and the General Practitioner Assessment of Cognition including the Clock Drawing Test compared to the improvement in Gait Speed. However, this could have various reasons. Either the selection of moderately improved values in the patient profiles could be responsible for this difference, or the perception of the participant was subjective and dependent on personal preferences. Nonetheless, to prevent these difficulties, both clinical outcome measures and the values representing moderate improvements in the patient profiles should be chosen and validated by experts. The former should be achieved with MCDA, the latter should be discussed thoroughly and supported by information from recent, significant literature.

The structure of the hard copies with separate boxes for each individual judgement could have been responsible for biased results. There is a possibility that panel members did not feel comfortable to choose a range of semantic categories due to the presentation of the necessary judgements. It could have also partially been partially responsible for a higher amount of inconsistency. Nonetheless, the possibility to choose a range was mentioned several times during the introduction and the scoring step and changes would be unlikely to deviate strongly from the current weights.

Another limitation of this research could be the subjective evaluation of frailty criteria by the panel members due to the impact of specific illnesses or problems. In the discussions several participants mentioned the feeling that if they were struggling with a specific issue, the corresponding frailty criterion got a higher weight. If the conditions of the panel members would change, it could be likely that weights for frailty criteria would be influenced and vary accordingly. This would cause unstable results.

Finally, the low amount of elderly panel members that took part in the assessment could be a constraint. With only six participants, the criteria weights might not be able to represent the importance of frailty criteria on a broader scale. The perception of the guidelines and the MACBETH approach could have also resulted in different findings, if more elderly individuals would have participated in the panel session. Further research is therefore necessary to verify the results.

4.4 Recommendations

Several recommendations can be proposed to enhance the use of MACBETH in healthcare-related multi-criteria group decision making. First of all, it would be of high value if future research would perform a comparison of AHP and MACBETH in a healthcare application. The same experts or elderly individuals should assess criteria with both the AHP and MACBETH. Subsequently the validity of the data should be analysed and a preferred method should be determined. This could finally settle the ongoing discussion between supporters of the AHP and MACBETH.

The pilot panel session has shown the importance of sufficient time being provided for the introduction. Without the ability to describe the procedure, the frailty criteria, the clinical outcome measures and the patient profiles in detail, elderly individuals in particular would be unable to make an informed decision during the pairwise comparisons. Participants should be provided with all the information necessary as hardcopies to act as guidance through the questioning procedure. In case of expert groups, it would be beneficial to use an audience response system such as “QuestionPress” for the collection of judgements. In addition, the recommendation to first prepare a ranking of the elements that need to be assessed can be very helpful, most notably for individuals of higher age.

The questionnaire and the discussions showed that elderly participants had mixed feelings about the MACBETH-methodology. A thorough introduction was essential for the understanding and the cognitive load was high. It can be assumed that younger individuals would have experienced less problems. However, it has shown to be of major importance to incorporate the patients’ perspective into medical decision making [2] and with nowadays ageing society, the elderly are the main stakeholders in this field. Therefore future assessments with elderly individuals should focus on simplifying the questioning procedure.

Another recommendation concerns the scale used in M-MACBETH. To the authors’ knowledge, no article could be found in which the founders of MACBETH commented on the choice of the semantic categories. This was also criticised by Salomon [88]. As mentioned by Ishizaka and Nemery [17], “other verbal scales can be imagined”. For this reason, an explanation of the selection of the semantic categories by the authors of MACBETH would be proposed. The six currently existing semantic categories for the judgements have been criticised during the discussions of the pilot panel session and in the questionnaire. Most participants would have preferred less categories to choose from. This could be related to the high age of the panel members, but for future assessments it would be beneficial if the developers of M-MACBETH would incorporate a flexible amount of semantic categories for the judgements.

Further suggestions regarding group decision making with MACBETH can be proposed. Until now, a consensus between decision makers was essential for making group judgements. This research has shown that the arithmetic mean is useful for gaining aggregated weights in the assessment of criteria related to frailty in the elderly. It is hence recommended to use this measure for groups which are unlikely to reach a consensus. For this reason, it is highly advocated that developers of M-MACBETH generate a possibility for group decision making in this software. The possibility of connecting an audience response system or keypads to M-MACBETH would be presumably also beneficial. These modifications can be expected to reduce the expenditure of time highly and hence increase the number of applications using MACBETH for group decision making.

Finally, in future assessments for Nutricia concerning the development of nutritional compounds for the prevention of frailty in the elderly, the developed modified guidelines will be useful to support decision makers in the planning and conduction of panel sessions with MACBETH.

5. Conclusion

The aim of this research was the development of guidelines for the use of MACBETH as a method for multi-criteria group decision making in the assessment of criteria related to the development of new medical products. On this account, a thorough literature study was performed to investigate the frequently used AHP and the less commonly applied MACBETH approach. Both methods were first explained in detail, then compared and finally best practices of healthcare-related group decision making with AHP were used for the preparation of the MACBETH-guidelines. As the title of this thesis states (“Ending the war in multi-criteria decision analysis: Taking the best from two worlds”), the best attributes of both approaches were identified and combined in the guidelines.

The next step of this work was to test the developed guidelines in a pilot panel session with elderly individuals concerning the assessment of criteria related to frailty in the elderly. This evaluation resulted in valuable data regarding the weights of different frailty criteria. In addition, discussions and a subsequent questionnaire gave information about the perception of the methodology and the results by elderly individuals. Minor flaws in the guidelines were detected, modifications were arranged and recommendations for an enhanced use of MACBETH were determined.

The ongoing discussion about the AHP is mainly focused on the missing adherence to axioms based on utility theory, but Saaty [78] described the AHP as a method different to utility theory and therefore being independent of the accompanying axioms. Several applications

have shown the suitability of this method for group decision making (e.g. [102-104]. In addition, there is an abundance of software accessible using the AHP. Nonetheless, the high amount of criticism should be taken into consideration when choosing an MCDA method. MACBETH has shown to be methodologically sound and is overall accepted in the MCDA community. Both approaches have their advantages, but it is the belief of the author that MACBETH could become as popular as or more popular than the AHP, if the software provided for this method would be improved. Decision makers usually have a limited amount of time for panel sessions and group decision making with M-MACBETH has shown to be very time consuming. With respect to the title of this thesis, if these modifications in M-MACBETH would be executed or if the acceptance of successful AHP applications would be increased, the “war” in multi-criteria decision analysis could be decided in favour of both MCDA methods.

The assessment showed that the developed guidelines for MACBETH are of high value for multi-criteria group decision making in healthcare. The results of the pilot panel session presented “Cognition” as the criterion with the highest relative importance, whereas “Hypercholesterolemia” came last in the ranking of frailty criteria. These findings will be helpful for the further assessments by Nutricia in September 2014. Two panel sessions are planned for the evaluation of clinical outcome measures and criteria related to the prevention of frailty in the elderly. The first panel session will be conducted with clinicians and payers, the second panel session will take part with patients and carers. It is recommended that the previously developed modified guidelines will be followed in these assessments. A provisional plan of these panel sessions is given in Appendix G.

This work has answered the main research question as well as all five sub-questions notably through extensive research and therefore made a great contribution in the field of MCDA. It is expected that this research will help towards an improved decision making in healthcare by gaining new information and assisting to increase transparency and rationality.

Besides the recommendations mentioned in section 4.4, future research should investigate the validity of calculated overall weights using the arithmetic mean with MACBETH. A focus on the development and verification of different group decision making approaches using MACBETH is of high importance to increase the application of this method in healthcare, due to many multidisciplinary decision makers and complex structures in this field.

Finally, a remarkable development should be mentioned. In a conversation with the founders of MACBETH it was stated that Bana Consulting is currently working on a major project in cooperation with the European Medicines Agency for the assessment of treatment options by patients. The study will involve 3000 participants in Europe and will be conducted with an online questionnaire. The new web application used for this project is called WISED.

It uses the MACBETH, but without the need for neither finding a consensus nor gathering all decision makers in one location. In addition, different levels of a hierarchy are assessed separately in WISED, which means that sub-criteria are assessed with respect to the intermediate goal instead of the overall goal. This is comparable to the AHP. Although more research is needed, it is expected that these developments will result in an enhanced use of MACBETH in healthcare-related multi-criteria group decision making.

Bibliography

1. Keeney, R.L. and H. Raiffa, *Decisions with multiple objectives: Preferences and value tradeoffs*. 1976, New York: J. Wiley.
2. IQWIG, *Analytic Hierarchy Process (AHP) - Pilotprojekt zur Erhebung von Patientenpräferenzen in der Indikation Depression*, in *IQWIG-Berichte*. 2013, Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen: Cologne, Germany.
3. Ishizaka, A. and P. Nemery, *General Introduction*, in *Multi-criteria Decision Analysis: Methods and Software*, A. Ishizaka and P. Nemery, Editors. 2013, John Wiley & Sons: Chichester, UK. p. 1-9.
4. Belton, V. and T. Stewart, *The Process of MCDA*, in *Multiple criteria decision analysis: an integrated approach*, V. Belton and T. Stewart, Editors. 2002, Springer. p. 5-8.
5. Fülöp, J. *Introduction to decision making methods*. in *BDEI-3 Workshop, Washington*. 2005.
6. Guitouni, A. and J.-M. Martel, *Tentative guidelines to help choosing an appropriate MCDA method.* European Journal of Operational Research, 1998. **109**(2): p. 501-521.
7. Diaby, V., K. Campbell, and R. Goeree, *Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis.* Operations Research for Health Care, 2013. **2**(1): p. 20-24.
8. Marsh, K., et al., *Assessing the Value of Healthcare Interventions Using Multi-Criteria Decision Analysis: A Review of the Literature.* PharmacoEconomics, 2014. **32**(4): p. 345-365.
9. Baltussen, R. and L. Niessen, *Priority setting of health interventions: the need for multi-criteria decision analysis.* Cost Effectiveness and Resource Allocation, 2006. **4**(1): p. 14.
10. Marold, J., et al., *A field study of group decision-making in health care.* Cahiers de la Sécurité Industrielle 2013(Number 2013-11).
11. Kee, F., T. Owen, and R. Leathem, *Decision Making in a Multidisciplinary Cancer Team: Does Team Discussion Result in Better Quality Decisions?* Medical Decision Making, 2004. **24**(6): p. 602-613.
12. Ho, D., et al., *Communication and sense-making in intensive care: an observation study of multi-disciplinary rounds to design computerized supporting tools.* AMIA Annu Symp Proc, 2007: p. 329-33.
13. Belton, V. and J. Pictet, *A Framework for Group Decision Using a MCDA Model: Sharing, Aggregating or Comparing Individual Information?* Journal of Decision Systems, 1997. **6**(3): p. 283-303.
14. Saaty, T.L., *Decision making with the analytic hierarchy process.* Int. J. Services Sciences, 2008. **1(1)**: p. 83-98.
15. Liberatore, M.J. and R.L. Nydick, *The analytic hierarchy process in medical and health care decision making: A literature review.* European Journal of Operational Research, 2008. **189**(1): p. 194-207.
16. National Institutes of Health. *Introduction to Expert Choice*. Training 2005 25/07/2014]; Available from: http://www.ors.od.nih.gov/OD/OQM/Documents/expertChoiceTraining.ppt.
17. Ishizaka, A. and P. Nemery, *MACBETH*, in *Multi-criteria Decision Analysis: Methods and Software*, A. Ishizaka and P. Nemery, Editors. 2013, John Wiley & Sons: Chichester, UK. p. 114-133.

18. Sanchez-Lopez, R., C.A. Bana e Costa, and B. Baets, *The MACBETH approach for multi-criteria evaluation of development projects on cross-cutting issues.* Annals of Operations Research, 2012. **199**(1): p. 393-408.
19. Tangerini, A., J. Pictet, and N. Soguel, *Using A Multiple criteria Decision Analysis approach for Landscape Quality Assessment.* 2006.
20. Bana e Costa, C.A., J.-M. De Corte, and J.-C. Vansnick, *On the mathematical foundation of MACBETH*, in *Multiple Criteria Decision Analysis: state of the art surveys*, J. Figueira, S. Greco, and M. Ehrgott, Editors. 2005, Springer. p. 409-442.
21. Nutricia. *Who we are*. About Nutricia 2014 04/03/2014]; Available from: http://www.nutricia.com/about-nutricia/who-we-are.
22. Brinkmeyer, D. and R.A.E. Müller, *Entscheidungsunterstützung mit dem AHP.* Zeitschrift für Agrarinformatik, 1994: p. 82-92.
23. Bahurmoz, A.M., *The Analytic Hierarchy Process: A Methodology for Win-Win Management.* JKAU: Econ. & Adm, 2006. **20**(1): p. 3-16.
24. Miller, G.A., *The magical number seven, plus or minus two: some limits on our capacity for processing information.* Psychological review, 1956. **63**(2): p. 81.
25. Saaty, T.L. and M.S. Ozdemir, *Why the magic number seven plus or minus two.* Mathematical and Computer Modelling, 2003. **38**(3–4): p. 233-244.
26. Hummel, J.M., J.F. Bridges, and M.J. Ijzerman, *Group Decision Making with the Analytic Hierarchy Process in Benefit-Risk Assessment: A Tutorial.* Patient, 2014.
27. Saaty, T.L., *A scaling method for priorities in hierarchical structures.* Journal of Mathematical Psychology, 1977. **15**(3): p. 234-281.
28. Ishizaka, A., D. Balkenborg, and T. Kaplan, *Influence of aggregation and measurement scale on ranking a compromise alternative in AHP.* Journal of the Operational Research Society, 2011. **62**(4): p. 700-710.
29. Harker, P.T. and L.G. Vargas, *The theory of ratio scale estimation: Saaty's analytic hierarchy process.* Management science, 1987. **33**(11): p. 1383-1403.
30. Lootsma, F., *Conflict resolution via pairwise comparison of concessions.* European Journal of Operational Research, 1989. **40**(1): p. 109-116.
31. Ishizaka, A., D. Balkenborg, and T. Kaplan, *Influence of aggregation and preference scale on ranking a compromise alternative in AHP.* 2006.
32. Ma, D. and X. Zheng, *9/9-9/1 Scale Method of AHP*, in *Proceedings of the 2nd International Symposium on the Analytic Hierarchy Process*, J.M. Katz and U.o. Pittsburgh, Editors. 1991, University of Pittsburgh. p. 197-202.
33. Salo, A.A. and R.P. Hämäläinen, *On the measurement of preferences in the analytic hierarchy process.* Journal of Multi-Criteria Decision Analysis, 1997. **6**(6): p. 309-319.
34. Ishizaka, A. and P. Nemery, *Analytic Hierarchy Process*, in *Multi-criteria Decision Analysis: Methods and Software*, A. Ishizaka and P. Nemery, Editors. 2013, John Wiley & Sons: Chichester, UK. p. 13-54.
35. Ishizaka, A. and M. Lusti, *How to derive priorities in AHP: a comparative study.* Central European Journal of Operations Research, 2006. **14**(4): p. 387-400.
36. Golany, B. and M. Kress, *A multicriteria evaluation of methods for obtaining weights from ratio-scale matrices.* European Journal of Operational Research, 1993. **69**(2): p. 210-220.
37. Srdjevic, B., *Combining different prioritization methods in the analytic hierarchy process synthesis.* Computers & Operations Research, 2005. **32**(7): p. 1897-1919.
38. Johnson, C.R., W.B. Beine, and T.J. Wang, *Right-left asymmetry in an eigenvector ranking procedure.* Journal of Mathematical Psychology, 1979. **19**(1): p. 61-64.

39. Cogger, K. and P. Yu, *Eigenweight vectors and least-distance approximation for revealed preference in pairwise weight ratios.* Journal of Optimization Theory and Applications, 1985. **46**(4): p. 483-491.
40. Saaty, T.L., *Calculating the priorities using an approximation method*, in *Decision making for leaders: the analytical hierarchy process for decisions in a complex world*. 1982, Lifetime Learning Publications. p. n.s.
41. Crawford, G. and C. Williams, *A note on the analysis of subjective judgment matrices.* Journal of mathematical psychology, 1985. **29**(4): p. 387-405.
42. Jensen, R.E., *An alternative scaling method for priorities in hierarchical structures.* Journal of mathematical psychology, 1984. **28**(3): p. 317-332.
43. Chu, A., R. Kalaba, and K. Spingarn, *A comparison of two methods for determining the weights of belonging to fuzzy sets.* Journal of Optimization theory and applications, 1979. **27**(4): p. 531-538.
44. Blankmeyer, E., *Approaches to consistency adjustment.* Journal of Optimization Theory and Applications, 1987. **54**(3): p. 479-488.
45. Cook, W.D. and M. Kress, *Deriving weights from pairwise comparison ratio matrices: An axiomatic approach.* European Journal of Operational Research, 1988. **37**(3): p. 355-362.
46. Belton, V. and T. Gear, *On a short-coming of Saaty's method of analytic hierarchies.* Omega, 1983. **11**(3): p. 228-230.
47. Holder, R., *Some comments on the analytic hierarchy process.* Journal of the Operational Research Society, 1990: p. 1073-1076.
48. Holder, R., *Response to Holder's comments on the Analytic Hierarchy Process: response to the response.* Journal of the Operational Research Society, 1991: p. 914-918.
49. Lootsma, F., *Scale sensitivity in the multiplicative AHP and SMART.* Journal of Multi-Criteria Decision Analysis, 1993. **2**(2): p. 87-110.
50. Barzilai, J. and F. Lootsma, *Power relations and group aggregation in the multiplicative AHP and SMART.* Journal of Multi-Criteria Decision Analysis, 1997. **6**(3): p. 155-165.
51. Saaty, T., *Group Decision Making and the AHP*, in *The Analytic Hierarchy Process*, B. Golden, E. Wasil, and P. Harker, Editors. 1989, Springer Berlin Heidelberg. p. 59-67.
52. Mikhailov, L., *Group prioritization in the AHP by fuzzy preference programming method.* Computers & operations research, 2004. **31**(2): p. 293-301.
53. Altuzarra, A., J.M. Moreno-Jimenez, and M. Salvador, *A Bayesian priorization procedure for AHP-group decision making.* European Journal of Operational Research, 2007. **182**(1): p. 367-382.
54. Cho, Y.-G. and K.-T. Cho, *A loss function approach to group preference aggregation in the AHP.* Computers & Operations Research, 2008. **35**(3): p. 884-892.
55. Dolan, J.G., B.J. Isselhardt, and J.D. Cappuccio, *The Analytic Hierarchy Process in Medical Decision Making A Tutorial.* Medical Decision Making, 1989. **9**(1): p. 40-50.
56. Vaidya, O.S. and S. Kumar, *Analytic hierarchy process: An overview of applications.* European Journal of Operational Research, 2006. **169**(1): p. 1-29.
57. Dyer, J.S., *Remarks on the analytic hierarchy process.* Management science, 1990. **36**(3): p. 249-258.
58. Saaty, T.L. and L.G. Vargas, *The legitimacy of rank reversal.* Omega, 1984. **12**(5): p. 513-516.

59. Saaty, T.L., *Fundamentals of decision making and priority theory with the analytic hierarchy process*. 1994, PA: RWS Publications: Pittsburgh.
60. Donegan, H., F. Dodd, and T. McMaster, *A new approach to AHP decision-making.* The Statistician, 1992: p. 295-302.
61. Barzilai, J., *Measurement and preference function modelling.* International Transactions in Operational Research, 2005. **12**(2): p. 173-183.
62. Saaty, T.L., *Highlights and critical points in the theory and application of the analytic hierarchy process.* European Journal of Operational Research, 1994. **74**(3): p. 426-447.
63. Saaty, T.L., *What is relative measurement? The ratio scale phantom.* Mathematical and Computer Modelling, 1993. **17**(4–5): p. 1-12.
64. Ishizaka, A. and A. Labib, *Review of the main developments in the analytic hierarchy process.* Expert Systems with Applications, 2011. **38**(11): p. 14336-14345.
65. Bana e Costa, C.A. and J.-C. Vansnick, *A critical analysis of the eigenvalue method used to derive priorities in AHP.* European Journal of Operational Research, 2008. **187**(3): p. 1422-1428.
66. Wang, Y.-M., K.-S. Chin, and Y. Luo, *Aggregation of direct and indirect judgments in pairwise comparison matrices with a re-examination of the criticisms by Bana e Costa and Vansnick.* Information Sciences, 2009. **179**(3): p. 329-337.
67. Korhonen, P., *Setting "Condition of Order Preservation" Requirements for the Priority Vector Estimate in AHP is not Justified*. 2008.
68. Anderlini, C.G., V.P. Salomon, and I.S. González, *A systemic rebuttal to the criticism of using the eigenvector for priority assessment in The Analytic Hierarchy Process for decision making.* Computación y Sistemas, 2008. **12**(2): p. 192-207.
69. Karapetrovic, S. and E. Rosenbloom, *A quality control approach to consistency paradoxes in AHP.* European Journal of Operational Research, 1999. **119**(3): p. 704-718.
70. Ishizaka, A. and M. Lusti, *An expert module to improve the consistency of AHP matrices.* International Transactions in Operational Research, 2004. **11**(1): p. 97-105.
71. Peláez, J. and M. Lamata, *A new measure of consistency for positive reciprocal matrices.* Computers & Mathematics with Applications, 2003. **46**(12): p. 1839-1845.
72. Aguaron, J. and J.M.a. Moreno-Jiménez, *The geometric consistency index: Approximated thresholds.* European Journal of Operational Research, 2003. **147**(1): p. 137-145.
73. Ji, P. and R. Jiang, *Scale transitivity in the AHP.* Journal of the Operational Research Society, 2003. **54**(8): p. 896-905.
74. Alonso, J.A. and M.T. Lamata, *Consistency in the analytic hierarchy process: a new approach.* International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2006. **14**(04): p. 445-459.
75. Stein, W.E. and P.J. Mizzi, *The harmonic consistency index for the analytic hierarchy process.* European journal of operational research, 2007. **177**(1): p. 488-497.
76. Vargas, L.G., *Comments on Barzilai and Lootsma: Why the multiplicative AHP is invalid: A practical counterexample.* Journal of Multi-Criteria Decision Analysis, 1997. **6**(3): p. 169-170.
77. Fishburn, P.C., *Utility theory for decision making*. 1970, DTIC Document.
78. Saaty, T.L., *An Exposition on the AHP in Reply to the Paper" Remarks on the Analytic Hierarchy Process".* Management science, 1990. **36**(3): p. 259-268.

79. Bana e Costa, C.A., de Corte, Jean-Marie and Vansnick, Jean-Claude, *MACBETH*, in *Operational Research working papers, LSEOR 03.56*. 2003, Department of Operational Research: London, UK.
80. Bana e Costa, C.A. and J.-C. Vansnick, *General overview of the MACBETH approach*, in *Advances in multicriteria analysis*, P.M. Pardalos, Y. Siskos, and C. Zopounidis, Editors. 1995, Springer. p. 93-100.
81. Bana e Costa, C.A. and J.-C. Vansnick, *A theoretical framework for measuring attractiveness by a categorical based evaluation technique (MACBETH)*, in *Multicriteria Analysis*, J. Climaco, Editor. 1997, Springer. p. 15-24.
82. Bana e Costa, C.A. and J.-C. Vansnick, *Cardinal Value Measurement with Macbeth*, in *Decision Making: Recent Developments and Worldwide Applications*, S. Zanakis, G. Doukidis, and C. Zopounidis, Editors. 2000, Springer US. p. 317-329.
83. Bana e Costa, C.A. and J.-C. Vasnick, *The MACBETH Approach: Basic Ideas, Software, and an Application*, in *Advances in Decision Analysis*, N. Meskens and M. Roubens, Editors. 1999. p. 131-157.
84. Bana e Costa, C.A. and M.P. Chagas, *A career choice problem: An example of how to use MACBETH to build a quantitative value model based on qualitative value judgments.* European journal of operational research, 2004. **153**(2): p. 323-331.
85. Bana e Costa, C.A., *MACBETH.* International journal of information technology & decision making, 2012. **11**(2): p. 359-387.
86. Bana e Costa, C.A., et al., *A Socio-technical Approach for Group Decision Support in Public Strategic Planning: The Pernambuco PPA Case.* Group Decision and Negotiation, 2014. **23**(1): p. 5-29.
87. Ertay, T., C. Kahraman, and I. Kaya, *Evaluation of renewable energy alternatives using MACBETH and fuzzy AHP multicriteria methods: the case of Turkey.* Technological and Economic Development of Economy, 2013. **19**(1): p. 38-62.
88. Salomon, V.A. *An example on the unreliability of MACBETH applications*. in *4th International Conference on Production Research*. 2008.
89. Saaty, R.W., *The analytic hierarchy process—what it is and how it is used.* Mathematical Modelling, 1987. **9**(3): p. 161-176.
90. Wedley, W.C. *AHP answers to problems with known composite values*. in *ISAHP* 2001. Berne, Switzerland.
91. Bana e Costa, C.A., J.-M. De Corte, and J.-C. Vansnick, *MACBETH.* International Journal of Information Technology & Decision Making, 2012. **11**(2): p. 359-387.
92. Choo, E.U., B. Schoner, and W.C. Wedley, *Interpretation of criteria weights in multicriteria decision making.* Computers & Industrial Engineering, 1999. **37**(3): p. 527-541.
93. Pinheiro, P.R., A. Castro, and M. Pinheiro. *A multicriteria model applied in the diagnosis of Alzheimer's disease: a Bayesian Network*. in *Computational Science and Engineering, 2008. CSE'08. 11th IEEE International Conference on*. 2008. IEEE.
94. Cox, R., J. Sanchez, and C.W. Revie, *Multi-Criteria Decision Analysis Tools for Prioritising Emerging or Re-Emerging Infectious Diseases Associated with Climate Change in Canada.* Plos One, 2013. **8**(8): p. 16.
95. Bana e Costa, C.A., M.C. Carnero, and M.D. Oliveira, *A multi-criteria model for auditing a Predictive Maintenance Programme.* European Journal of Operational Research, 2012. **217**(2): p. 381-393.
96. Santos, F.A., A.E. Margotti, and R. Garcia, *Multi-Criteria Decision Aid (MCDA) as a Tool to Support Health Technology Incorporation Process*, in *World Congress on Medical*

Physics and Biomedical Engineering May 26-31, 2012, Beijing, China, M. Long, Editor. 2013, Springer Berlin Heidelberg. p. 692-695.

97. Bana e Costa, C.A., T.G. Fernandes, and P.V. Correia, *Prioritisation of public investments in social infrastructures using multicriteria value analysis and decision conferencing: a case study.* International Transactions in Operational Research, 2006. **13**(4): p. 279-297.
98. Bana e Costa, C.A., et al., *Facilitating bid evaluation in public call for tenders: a socio-technical approach.* Omega, 2002. **30**(3): p. 227-242.
99. Gooskens, P.J., *Prioritization of medical needs in frail elderly*, in *Health Services and Management*. 2013, University of Twente: Enschede, the Netherlands.
100. QuestionPress. *Home*. 2014 28/04/2014]; Available from: http://www.questionpress.com/home.
101. Kurverwaltung Bad Mergentheim, *Volkskrankheit Diabetes: Viele unterschätzen ihr Risiko*, K.B. Mergentheim, Editor. 2010: Bad Mergentheim.
102. Srdevic, Z., B. Blagojevic, and B. Srdevic, *AHP based group decision making in ranking loan applicants for purchasing irrigation equipment: A case study.* Bulgarian Journal of Agricultural Science, 2011. **17**(4): p. 531-543.
103. Lai, V.S., B.K. Wong, and W. Cheung, *Group decision making in a multiple criteria environment: A case using the AHP in software selection.* European Journal of Operational Research, 2002. **137**(1): p. 134-144.
104. Al-Harbi, K.M.A.-S., *Application of the AHP in project management.* International Journal of Project Management, 2001. **19**(1): p. 19-27.
105. Bana e Costa, C.A. and J.-C. Vansnick, *Preference relations and MCDM*, in *Multicriteria Decision Making*, T. Gal, T.J. Stewart, and T. Hanne, Editors. 1999, Springer. p. 4-1 - 4-23.
106. de Castro, A.K.A., P.R. Pinheiro, and M.C.D. Pinheiro, *A Hybrid Model for Aiding in Decision Making for the Neuropsychological Diagnosis of Alzheimer's Disease*, in *Rough Sets and Current Trends in Computing, Proceedings*, C.C. Chan, J.W. GrzymalaBusse, and W.P. Ziarko, Editors. 2008, Springer-Verlag Berlin: Berlin. p. 495-504.
107. Oliveira, M.D., et al., *Prioritizing health care interventions: A multicriteria resource allocation model to inform the choice of community care programmes*, in *Advanced Decision Making Methods Applied to Health Care*. 2012, Springer. p. 141-154.
108. Pictet, J. and D. Bollinger, *The silent negotiation or how to elicit collective information for group MCDA without excessive discussion.* Journal of Multi-Criteria Decision Analysis, 2005. **13**(5-6): p. 199-211.
109. Saaty, T., *Making and validating complex decisions with the AHP/ANP.* Journal of Systems Science and Systems Engineering, 2005. **14**(1): p. 1-36.
110. Hummel, M.J., et al., *Assessing medical technologies in development.* International journal of technology assessment in health care, 2000. **16**(04): p. 1214-1219.
111. Hummel, M.M., et al., *Using the Analytic Hierarchy Process to Elicit Patient Preferences.* The Patient: Patient-Centered Outcomes Research, 2012. **5**(4): p. 225-237.
112. Hummel, J.M., J.F.P. Bridges, and M.J. Ijzerman, *Group Decision Making with the Analytic Hierarchy Process in Benefit-Risk Assessment: A Tutorial.* Patient, 2014: p. 1-12.
113. Srdjevic, B., et al., *A two-phase algorithm for consensus building in AHP-group decision making.* Applied Mathematical Modelling, 2013. **37**(10-11): p. 6670-6682.

114. Belton, V. and T. Gear, *The legitimacy of rank reversal—a comment.* Omega, 1985. **13**(3): p. 143-144.
115. Salomon, V.A. and J.A.B. Montevechi. *A compilation of comparisons on the analytic hierarchy process and others multiple criteria decision making methods: some cases developed in Brazil*. in *6th International Symposium on the Analytic Hierarchy Process, Bern*. 2001.
116. Barzilai, J., *On the decomposition of value functions.* Operations Research Letters, 1998. **22**(4): p. 159-170.
117. Saaty, R.W. *Why barzilai's criticisms of the ahp are incorrect*. in *Proceedings of the International MCDM Meeting*. 2004.
118. Garuti Anderlini, C., V. Pamplona Salomon, and I. Spencer González, *A Systemic Rebuttal to the Criticism of Using the Eigenvector for Priority Assessment in the Analytic Hierarchy Process for Decision Making.* Computación y Sistemas (CyS), 2008. **12**(2).
119. Bui, T.X., *Design issues for the GDSS model component*, in *Co-oP: A Group Decision Support System for Cooperative Multiple Criteria Group Decision Making*, T.X. Bui, Editor. 1987, Springer Berlin Heidelberg. p. 51-89.
120. Thokala, P. and A. Duenas, *Multiple criteria decision analysis for health technology assessment.* Value in Health, 2012. **15**(8): p. 1172-1181.
121. Cesari, M., et al., *Prognostic value of usual gait speed in well-functioning older people--results from the Health, Aging and Body Composition Study.* J Am Geriatr Soc, 2005. **53**(10): p. 1675-80.
122. Purser, J.L., et al., *Identifying frailty in hospitalized older adults with significant coronary artery disease.* J Am Geriatr Soc, 2006. **54**(11): p. 1674-81.
123. Afilalo, J., et al., *Gait Speed as an Incremental Predictor of Mortality and Major Morbidity in Elderly Patients Undergoing Cardiac Surgery.* Journal of the American College of Cardiology, 2010. **56**(20): p. 1668-1676.
124. Öberg, T., A. Karsznia, and K. Öberg, *Basic gait parameters: reference data for normal subjects, 10-79 years of age.* Journal of rehabilitation research and development, 1993. **30**: p. 210-210.
125. Podsiadlo, D. and S. Richardson, *The timed "Up & Go": a test of basic functional mobility for frail elderly persons.* J Am Geriatr Soc, 1991. **39**(2): p. 142-8.
126. Verghese, J. and X. Xue, *Identifying frailty in high functioning older adults with normal mobility.* Age and ageing, 2010. **39**(3): p. 382-385.
127. da Câmara, S.M.A., et al., *Using the Short Physical Performance Battery to screen for frailty in young-old adults with distinct socioeconomic conditions.* Geriatrics & gerontology international, 2013. **13**(2): p. 421-428.
128. Guralnik, J.M., et al., *Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery.* J Gerontol A Biol Sci Med Sci, 2000. **55**(4): p. M221-31.
129. Steffen, T.M., T.A. Hacker, and L. Mollinger, *Age-and gender-related test performance in community-dwelling elderly people: Six-Minute Walk Test, Berg Balance Scale, Timed Up & Go Test, and gait speeds.* Physical therapy, 2002. **82**(2): p. 128-137.
130. Prost, E. *Berg Balance Scale*. Examination Tool Kit 2013 16/05/2014]; Available from: geriatrictoolkit.missouri.edu/Berg-Balance-Scale.doc.

131. Dite, W. and V.A. Temple, *A clinical test of stepping and change of direction to identify multiple falling older adults.* Archives of physical medicine and rehabilitation, 2002. **83**(11): p. 1566-1571.
132. Whitney, S.L., et al., *The Reliability and Validity of the Four Square Step Test for People With Balance Deficits Secondary to a Vestibular Disorder.* Archives of Physical Medicine and Rehabilitation, 2007. **88**(1): p. 99-104.
133. Desai, A., et al., *Relationship Between Dynamic Balance Measures and Functional Performance in Community-Dwelling Elderly People.* Physical Therapy, 2010. **90**(5): p. 748-760.
134. Jones, C.J., R.E. Rikli, and W.C. Beam, *A 30-s chair-stand test as a measure of lower body strength in community-residing older adults.* Research quarterly for exercise and sport, 1999. **70**(2): p. 113-119.
135. American Thoracic Society, *ATS statement: guidelines for the six-minute walk test.* Am J Respir Crit Care Med, 2002. **166**(1): p. 111-7.
136. Krause, M.P., et al., *Influence of the level of physical activity over the cardiorespiratory capacity in older women.* Revista Brasileira de Medicina do Esporte, 2007. **13**(2): p. 97-102.
137. Yoshitake, H., *Three Characteristic Patterns of Subjective Fatigue Symptoms.* Ergonomics, 1978. **21**(3): p. 231-233.
138. Massy-Westropp, N.M., et al., *Hand Grip Strength: age and gender stratified normative data in a population-based study.* BMC research notes, 2011. **4**(1): p. 127.
139. Alley, D.E., et al., *Grip strength cutpoints for the identification of clinically relevant weakness.* J Gerontol A Biol Sci Med Sci, 2014. **69**(5): p. 559-66.
140. Bohannon, R.W., et al., *Reference values for adult grip strength measured with a Jamar dynamometer: a descriptive meta-analysis.* Physiotherapy, 2006. **92**(1): p. 11-15.
141. Topend Sports. *Leg Extension Strength Test*. Fitness Testing 2014 16/05/2014]; Available from: http://www.topendsports.com/testing/tests/leg-isometric-extension.htm.
142. Alzheimer's Association, *Cognitive assessment toolkit. A guide to detect cognitive impairment quickly and efficiently during the Medicare Annual Wellness Visit*, A.s. Association, Editor. 2014: Chicago, USA.
143. Malloy-Diniz, L.F., et al., *The Rey auditory-verbal learning test: applicability for the Brazilian elderly population.* Revista Brasileira de Psiquiatria, 2007. **29**(4): p. 324-329.
144. Folstein, M.F., S.E. Folstein, and P.R. McHugh, *"Mini-mental state": a practical method for grading the cognitive state of patients for the clinician.* Journal of psychiatric research, 1975. **12**(3): p. 189-198.
145. Kurlowicz, L. and S.A. Greenberg, *The geriatric depression scale (GDS).* AJN The American Journal of Nursing, 2007. **107**(10): p. 67-68.
146. Hamilton, M., *A rating scale for depression.* Journal of neurology, neurosurgery, and psychiatry, 1960. **23**(1): p. 56.
147. American Heart Association. *Understanding Blood Pressure Readings*. High Blood Pressure 2012 16/05/2014]; Available from: http://www.heart.org/HEARTORG/Conditions/HighBloodPressure/AboutHighBloodPressure/Understanding-Blood-Pressure-Readings_UCM_301764_Article.jsp#.
148. Mayo Clinic. *Cholesterol levels: What numbers should you aim for?* Diseases and Conditions. High cholesterol. 2012 16/05/2014]; Available from:

http://www.mayoclinic.org/diseases-conditions/high-blood-cholesterol/in-depth/cholesterol-levels/art-20048245.

149. Abbott Diabetes Care Deutschland. *Kontinuierliche Glukosemessung*. Diabetes Management 2014 17/05/2014]; Available from: https://www.abbott-diabetes-care.de/de-de/insulinanwender-und-angehoerige/diabetes-management/kontinuierliche-glukosemessung/.
150. Diabetes Ratgeber. *Blutzucker (Glukose)*. Laborwerte 2011 17/05/2014]; Available from: http://www.diabetes-ratgeber.net/laborwerte/blutzucker.
151. Studenski, S., et al., *Gait speed and survival in older adults.* JAMA, 2011. **305**(1): p. 50-58.
152. Braden, H., *Self-selected gait speed: A critical clinical outcome*, in *Lower Extremity Review Magazine*. 2012, Richard Dubin: New York, USA.
153. Bean, J.F., et al., *The relationship between leg power and physical performance in mobility-limited older people.* J Am Geriatr Soc, 2002. **50**(3): p. 461-7.
154. Lange, T., et al., *Six-minute walk distance target in elderly patients with idiopathic pulmonary arterial hypertension–consideration of predicted values.* International journal of clinical practice, 2014. **68**(5): p. 543-550.
155. Frisoni, G.B., et al., *Principal lifetime occupation and MMSE score in elderly persons.* Journal of gerontology, 1993. **48**(6): p. S310-S314.
156. Kompetenz Centrum Geriatrie. *Geriatrische Depressionsskala*. n.s. 15/05/2014]; Available from: http://www.kcgeriatrie.de/instrumente/gds.pdf.

Appendices

Appendix A: Literature Overview with the Most Important Articles

MACBETH

Theory

Bana e Costa, De Corte [20]
Bana e Costa [79]
Bana e Costa and Vansnick [80]
Bana e Costa and Vansnick [81]
Bana e Costa and Vansnick [82]
Bana e Costa and Vasnick [83]
Bana e Costa [85]
Bana e Costa and Vansnick [105]

Applications (particularly in Healthcare)

Bana e Costa and Chagas [84]
Cox, Sanchez [94]
Bana e Costa, Carnero [95]
de Castro, Pinheiro [106]
Oliveira, Rodrigues [107]

Group Decision Making

Bana e Costa, Lourenço [86]
Pictet and Bollinger [108]

AHP

Theory

Saaty [14]
Brinkmeyer and Müller [22]
Saaty [27]
Ishizaka, Balkenborg [28]
Saaty [89]
Saaty [109]

Healthcare

IQWIG [2]
Liberatore and Nydick [15]
Dolan, Isselhardt [55]
Hummel, van Rossum [110]
Hummel, Volz [111]

Group Decision Making

Hummel, Bridges [112]
Srdjevic, Srdjevic [113]

Discussion AHP and MACBETH

Belton and Gear [46]
Saaty and Vargas [58]
Saaty [59]
Ishizaka and Labib [64]
Bana e Costa and Vansnick [65]
Wang, Chin [66]
Bana e Costa and Vansnick [80]
Salomon [88]
Korhonen [67]
Belton and Gear [114]
Salomon and Montevechi [115]
Barzilai [116]
Saaty [117]
Garuti Anderlini, Pamplona Salomon [118]

MCDA Group Decision Making

Belton and Pictet [13]
Pictet and Bollinger [108]
Bui [119]

MCDA in Healthcare

Diaby, Campbell [7]
Marsh, Lanitis [8]
Baltussen and Niessen [9]
Thokala and Duenas [120]

Appendix B: Overview Clinical Outcome Measures

Table 8: Explanations of clinical outcome measures.

MOBILITY		
COM	**Definition**	**Scale/Interpretation**
Gait Speed [121-124]	Accepted as a measure of frailty **Test:** • Patient has to walk 5 metres at a comfortable pace • Time is recorded • Repeated 3 times • Calculation of average speed • Measurement: metres/seconds	**Healthy Men, Age 70-79:** Slow: 0.8 m/s Normal: 1.18 m/s Fast: 1.59 m/s **Healthy Women, Age 70-79:** Slow: 0.74 m/s Normal: 1.11 m/s Fast: 1.42 m/s • Improvement of 0.1 m/s should be the goal
Timed Up and Go [125]	Test of basic mobility skills **Test:** • Patient is sitting on arm chair • Stands up from chair • Walks 3 metres at normal pace • Turn, walk back • Sit back on chair • Time is recorded • Measurement: seconds	**Healthy adults, Age 70-79:** Average: 8.2-10.2 s **Groups:** <10 s: No constraints 11-19 s: Low constraints 20-29 s: Relevant constraints >30 s: Severe constraints
Flight of Stairs Questionnaire	No sufficient information found	No sufficient information found
Short Physical Performance Battery [126-128]	Good for detecting early stages of frailty **Test:** 1. Repeated Chair Stands (5 repetitions) 2. Balance Testing - Semitandem - Side-by-Side - Tandem 3. 8' Walk (2.44 metres) Max. points per category: 4 Max. points summary: 12	**Scores:** 0-6: Poor performance 7-9: Intermediate performance 10-12: High performance

BALANCE		
COM	**Definition**	**Scale/Interpretation**
Berg Balance Scale [129, 130]	Developed as a performance-oriented measure of balance in elderly **Test:** • 14 items, mobility tasks • scored from 0 (unable to fulfil task) – 4 (fully able to fulfil task) • Maximum total score: 56	**Scores:** 0-20: high fall risk 21-40: medium fall risk 41-56: low fall risk • A change of 8 points: relevant
Four Square Step Test [131, 132]	Validated as a balance test for identifying fall risk in older adults **Test:** • The FSST involves stepping over 4 canes that are laid on the ground at 90° angles to each other • Patient has to rotate clockwise around the "plus sign" by moving forward, to the right, backward, to the left • 1 repetition • Best score counts	**Scores:** ≤ 15 s: low risk for multiple falls >15 s: high risk for multiple falls
Accelerometric Measurements on a Mat with Sensors [133]	The Mat can be used to record the foot's center of pressure (COP), the sway path length and the loss of balance on different surfaces during specific activities and with eyes open or closed	**Loss of balance frequency:** Non-fallers: 2.4 (median) Fallers: 3.5 (median) **Sway path length (sponge surface):** Non-fallers: 19.1 cm (median) Fallers: 24.1 cm (median) **Center of pressure excursion:** Non-fallers: 17.2 cm (median) Fallers: 23.5 cm (median)
Chair Stand Test [134]	**Test:** • Patient sits on chair • On signal, patient rises to full stand • Then returns to a fully seated position • As many times as possible in 30 seconds • Measurement: number of full stands	**Scores Men (per Age):** 75-79: 11-17 stands 80-84: 10-15 stands 85-90: 8-14 stands **Scores Women (per Age):** 75-79: 10-15 stands 80-84: 9-14 stands 85-90: 8-13 stands

ENERGY		
COM	**Definition**	**Scale/Interpretation**
Six-Minute Walk Distance [135, 136]	This test measures the distance that a patient can quickly walk on a flat, hard surface in a period of 6 minutes (the 6MWD) • Measurement: metres	**Average scores (per Age):** 70-74: 483.2 ± 71.1 metres 75-79: 461.2 ± 70.8 metres >80 : 426.1 ± 73.2 metres • Relevant improvement: 70 metres
Subjective Fatigue Questionnaire [137]	Questionnaire consists of 30 questions which can be answered with "Yes" or "No" • 10 questions: Drowsiness and dullness • 10 questions: Difficulty of concentration • 10 questions: Projection of physical impairment	No sufficient information found
GRIP STRENGTH		
COM	**Definition**	**Scale/Interpretation**
Jamar Dynamometer [138-140]	Test to measure upper body strength **Test:** • Tool: Jamar Analogue Hand Dynamometer • Participants are seated • Elbow by their side and flexed to right angles • Neutral wrist position • Dynamometer handle position II • Repetition: 3 times on each hand • Measurement: Grip strength in kg	**Average Men (per Hand & Age):** Right hand 70-74: 38.2 kg 75+ : 28.0 kg Left hand 70-74: 36.2 kg 75+ : 29.8 kg • <26: considered weak **Average Women (per Hand & Age):** Right hand 70-74: 24.2 kg 75+ : 18.0 kg Left hand 70-74: 22.5 kg 75+ : 16.4 kg • <16: considered weak
Leg Extension Strength [141]	Test to measure lower body strength **Test:** • Participant sits on table, legs hang down, knee angle 90° • A shin guard attached to a resistance is fastened around the right lower leg of the participant • Participant must try to extend the right leg by raising the lower leg with maximum strength and hold that position for 3 seconds • Measurement: kg of force • Repetition: 3 times • Resting period in between: 30s	No sufficient information found

COGNITION		
COM	**Definition**	**Scale/Interpretation**
General Practitioner Assessment of Cognition incl. Clock Drawing Test [142]	Test was designed as a GP screening tool for dementia **Test:** • Name and address for subsequent recall test • Time Orientation • Clock Drawing • Information news • Recall • Additional informant section Maximum Score: 9	**Scores:** • Score 9: no significant cognitive impairment and further testing not necessary • Score 5-8: more information required, proceeding with informant section • Score 0-4: cognitive impairment is indicated, conduct standard investigations
Rey Auditory Verbal Learning Test combined with Clock Drawing Test [143]	The test measures recent memory, verbal learning, susceptibility to interference, retention of information and recognition memory **Test:** • 15 substantives are read aloud to the participant five times (list A) • Test of retrieval after each attempt (list A) (Attempt A1 – A5) • 15 new substantives are read aloud to the participant (list B) • Test of retrieval (list B) (Attempt B1) • Recall of words from list A (Attempt A6) • 20 minute interval • Test of retrieval (list A) (Attempt A7) • Test of memory recognition: words from list A plus words from list B plus 20 distracting words: Individual is asked to indicate if words belong to list A Calculation of • ITP (proactive interference): (B1/A1) • ITR (retroactive interference): (A6/A5) • VE (forgetting speed): (A7/A6) • REC (recognition): sum of correct answers • LOT (Rate of learning): Sum(A1-A5) – 5*(A1)	**Average Scores Men (per Age):** MEN WOMEN 70-74: ITP: 0.9 0.8 ITR: 0.9 0.8 VE: 0.9 1.0 REC: 7.2 9.0 LOT: 15.3 14.5 75-79: ITP: 0.8 0.8 ITR: 0.8 0.8 VE: 0.9 0.9 REC: 7.5 6.2 LOT: 12.5 12.0 80-84: ITP: 0.8 0.6 ITR: 0.9 0.9 VE: 0.9 0.8 REC: 5.5 6.1 LOT: 16.0 11.4
Standardized Mini-Mental State Examination [144]	The SMMSE is a 30-point questionnaire test that is used to screen for cognitive impairment **Test:** • Orientation to time (5 Points) • Orientation to place (5 Points) • Registration (3 Points) • Attention and calculation (5 Points) • Recall (3 Points) • Language (2 Points) • Repetition (1 Point) • Complex commands (6 Points) Maximum of 30 Points	**Scores:** 25-30: Questionably significant degree of impairment 20-25: Mild degree of impairment 10-20: Moderate degree of impairment 0-10: Severe degree of impairment

DEPRESSION

COM	Definition	Scale/Interpretation
Geriatric Depression Scale [145]	Brief, 15-item questionnaire in which participants are asked to respond by answering yes or no in reference to how they felt over the past week **Test:** • 10 questions indicate the presence of depression when answered positively • 5 questions indicate the presence of depression when answered negatively	**Scores:** 0-4: normal, depending on age, education and complaints 5-8: indicates mild depression 9-11: indicates moderate depression 12-15: indicates severe depression
Hamilton Rating Scale for Depression [146]	The HRSC is a questionnaire that consists of 17 items, used to provide an indication for depression Test • Depressed Mood • Feelings of Guilt • Suicide • Insomnia • Work and Activities • Retardation • Agitation • Anxiety • Somatic Symptoms • Genital Symptoms • Hypochondriasis • Loss of Weight • Insight Maximum score: 66	**Scores** 0-7: Normal 8-13: Mild Depression 14-18: Moderate Depression 19-22: Severe Depression ≥ 23: Very Severe Depression

HYPERTENSION

COM	Definition	Scale/Interpretation
Blood Pressure [147]	**Systolic Blood Pressure** Measures the pressure in the arteries when the heart beats (when the heart muscle contracts) **Diastolic Blood Pressure** Measures the pressure in the arteries between heartbeats (when the heart muscle is resting between beats and refilling with blood) Measurement: mm Hg	**Categories:** SBP / DBP Normal: < 120 / < 80 Prehypertension: 120-139 / 80-89 Hypertension Stage 1: 140-159 / 90-99 Hypertension Stage 2: ≥ 160 / ≥ 100 Hypertensive Crisis: > 180 / > 110

HYPERCHOLESTEROLEMIA		
COM	**Definition**	**Scale/Interpretation**
Cholesterol Level [148]	The Cholesterol Level is measured in millimoles (mmol) per liter (L) of blood in most European countries and in milligrams (mg) of cholesterol per deciliter (dL) of blood in the United States and some other countries	**Interpretation:** Total Cholesterol: Desirable: < 5.2 mmol/L < 200 mg/dL Borderline high: 5.2-6.2 mmol/L 200-239 mg/dL High: > 6.2 mmol/L > 240 mg/dL LDL Cholesterol: Ideal: <1.8 mmol/L < 70 mg/dL Ideal at risk: < 2.6 mmol/L < 100 mg/dL Near ideal: 2.6-3.3 mmol/L 100-129 mg/dL Borderline: 3.4-4.1 mmol/L 130-159 mg/dL High: 4.1-4.9 mmol/L 160-189 mg/dL Very high: > 4.9 mmol/L > 190 mg/dL HDL Cholesterol: Poor Men: < 1 mmol/L < 40 mg/dL Poor Women: < 1.3 mmol/L < 50 mg/dL Better Men: 1-1.3 mmol/L 40-49 mg/dL Better Women: 1.3-1.5 mmol/L 50-59 mg/dL Best: ≥ 1.6 mmol/L ≥ 60 mg/dL
FLUCTUATING GLUCOSE LEVELS		
COM	**Definition**	**Scale/Interpretation**
Continuous Glucose Monitoring [149, 150]	Continuous glucose monitoring (CGM) systems use a tiny sensor inserted under the skin to check glucose levels in interstitial fluid	**Target blood glucose levels:** Non-diabetic: Before meals: 4-5.9 mmol/L 2 h after meals: < 7.8 mmol/L Type 2 diabetes: Before meals: 4-7 mmol/L 2 h after meals: < 8.5 mmol/L Type 1 diabetes: Before meals: 4-7 mmol/L 2 h after meals: < 9 mmol/L

Appendix C: Patient Profiles

Table 10: Patient profiles for different clinical outcome measures.

Criteria	Clinical Outcome Measure	Patient Profile 1	Patient Profile 2	Range
Mobility	Gait Speed (GS) [123, 151, 128, 124, 152]	0.53 m/s	0.64 m/s	0 - >1.40 m/s
	Timed Up and Go (TUG) [125]	22 s	18 s	<10, 10-19, 20-30, >30
	Flight of Stairs Questionnaire (FOSQ)			
	Short Physical Performance Battery (SPPB) [153, 127, 128]	7	8	Scores: 0-12
Balance	Berg Balance Scale (BBS) [130]	31	39	Scores: <21, 21-40, >40 (0-56)
	Four Square Step Test (FSST) [131]	23 s	18 s	≤15,>15
	Acc. Meas. on Sensor Mat (AMSM)			
	Chair Stand Test (CST) [134]	9 stands/30s	11 stands/30s	0 - >19
Energy	Six-Minute Walk Distance (6MW) [154]	190 m	260 m	0 - >600
	Subjective Fatigue Questionnaire (SFQ)			
Grip Strength	Jamar Dynamometer (JD) [140, 138]	17.1 kg	18.3 kg	0 - > 40 kg
	Leg Extension Strength (LES) [153]	25 kg	28.6 kg	
Cognition	GP Assessment of Cognition (GPCOG+CDT) [142]	5	7	Scores: 0-9
	Rey Auditory Verbal Learning Test (RAVLT) [143]	5	7	Words: max. 15
	Stand. Mini-Mental State Exam. (SMMSE) [155]	20	24	Scores: 0-30
Depression	Geriatric Depression Scale (GDS) [156]	5	3	Score: 15-0
	Hamilton Rating Scale (HRSC/HAM-D17/7) [146]	10	6	Score: 66-0
Hypertension	Blood Pressure (SBP+DBP) [147]	153/92	136/84	
Hypercholesterolemia	Cholesterol Level (CL) [148]	6.2 mmol/L	5.2 mmol/L	
Fluctuating Glucose Levels	Continuous Glucose Monitoring (CGM) [150]	Fasting: 110 mg/dL 2h after food intake: 140 mg/dL	Fasting: 99 mg/dL 2h after food intake: 120 mg/dL	

Appendix D: Clinical Outcome Measures Pilot Panel Session

Table 11: Overview choice of clinical outcome measures for pilot panel session.

Criterion	**Clinical Outcome Measure**
Mobility	Gait Speed
Balance	Chair Stand Test
Energy	Six-Minute Walk Distance
Grip Strength	Jamar Dynamometer
Cognition	The General Practitioner Assessment of Cognition incl. Clock Drawing Test
Depression	Geriatric Depression Scale
Hypertension	Blood Pressure
Hypercholesterolemia	Cholesterol Level
Fluctuating Glucose Levels	Continuous Glucose Monitoring

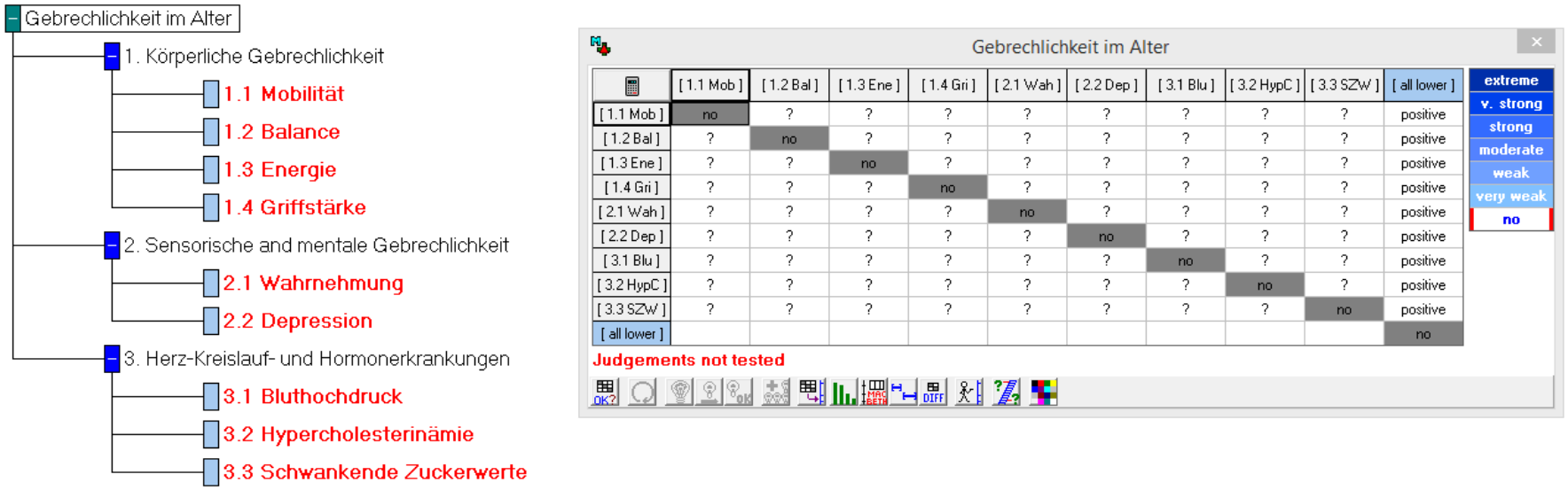


Figure 10: Decision tree and judgement matrix for pilot panel session (German).

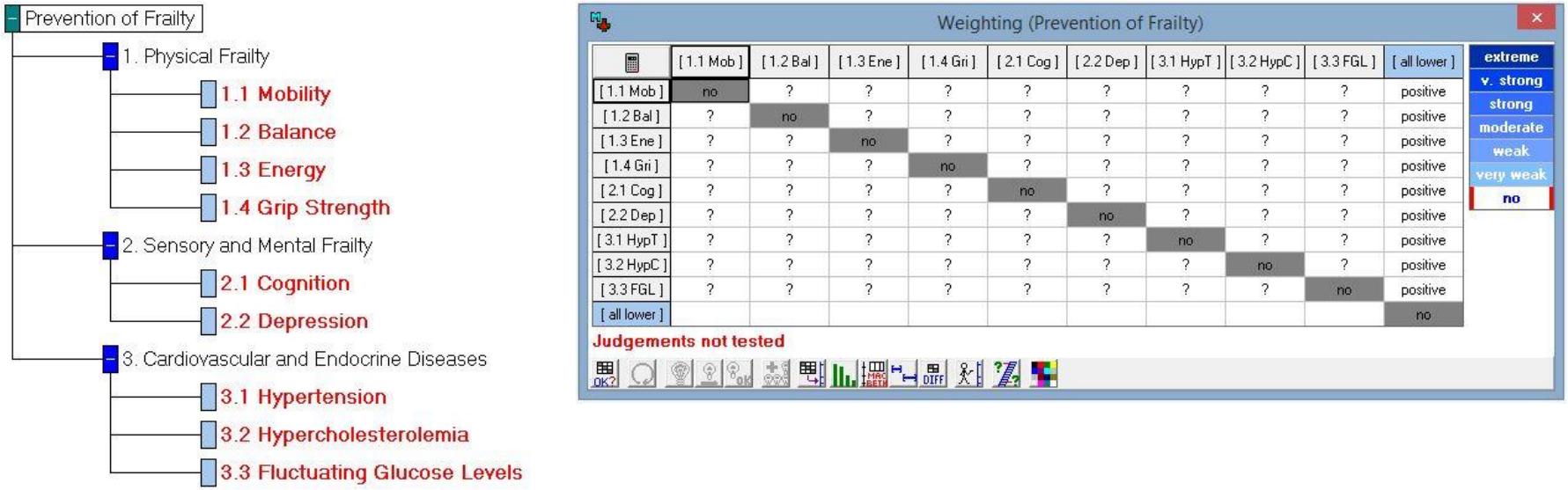


Figure 11: Decision tree and judgement matrix for pilot panel session (English).

Appendix F: Questionnaire Pilot Panel Session

1. **Do you think that all relevant frailty criteria were included in the evaluation? If not, which frailty criterion would you have added? Or which frailty criterion did you find not relevant?**

   - P1: I was missing the frailty criterion “capacity of reaction”
   - P2: Criteria about the independence of the elderly in everyday life were missing
   - P3: I think physical deficiencies should have gotten more criteria
   - P4: Blank
   - P5: Blank
   - P6: Yes

2. **Do you think that participants felt comfortable and free to communicate their point of view and do you think that the discussions took place in an atmosphere of trust?**

   - P1: Yes
   - P2: Yes, no manipulation
   - P3: Yes
   - P4: Yes
   - P5: I believe that participants were able to communicate their point of view without any problems
   - P6: Yes

3. **Do you think that participants were put under pressure and that participants were scared to stand behind their own opinions?**

   - P1: No
   - P2: No
   - P3: No
   - P4: No
   - P5: No
   - P6: No

4. **What do you think about the semantic categories available for your judgement? Was the scale just right, too comprehensive or not sufficient?**

   - P1: The scale was too comprehensive for my taste
   - P2: Too comprehensive
   - P3: Too comprehensive
   - P4: Too comprehensive, but I guess you need these different categories for all those different criteria
   - P5: Too comprehensive
   - P6: Just right

5. **Do you think that you got enough information in the introduction to make your judgements?**
   - P1: Yes
   - P2: Yes
   - P3: It was difficult to identify myself with some of the criteria in the patient profile, because I never experienced fluctuating glucose levels or high blood pressure, but there was enough information
   - P4: Yes
   - P5: Yes
   - P6: Once I understood the clinical outcome measures, it was fine

6. **Do you think, that the moderator has motivated participants equally to take part in the discussions? What else would you have liked from the moderator?**

   - P1: Same attention was given to all participants
   - P2: The moderator did well, all problems were solved, she answered all the questions and provided help if needed
   - P3: Yes, the moderator was good
   - P4: The moderator was very professional and helped motivating participants
   - P5: The moderator motivated participants equally
   - P6: All good, good explanations

7. **What do you think about the procedure of pairwise comparisons? Do you think that it reduced the cognitive load of the evaluation? What changes would you find necessary?**

   - P1: Yes, I think it was good and it reduced the cognitive load, but I was unaccustomed to it
   - P2: I found the pairwise comparison with the scale very difficult
   - P3: No, I did not like the procedure, it was very difficult
   - P4: Blank
   - P5: Blank
   - P6: I liked the procedure

8. **What do you think about the expenditure of time of this panel session? Do you think it was just right, too extensive or too short?**

   - P1: Just right
   - P2: The expenditure of time was fine
   - P3: Just right
   - P4: Fine
   - P5: The expenditure of time is suitable
   - P6: Just right, but should not have been much longer, due to the loss of concentration

9. **Do you think that the procedure was easy to understand and user-friendly?**

- P1: Not really easy to understand, takes some time to get used to it
- P2: The procedure needs a good introduction and understanding of the questions
- P3: Not very user-friendly, at least not for people my age
- P4: Blank
- P5: After the introduction through the moderator the procedure was user-friendly
- P6: Yes, it was user-friendly

10. **Do you think that the results of today's panel session are realistic and that they represent your personal evaluation?**

- P1: Yes
- P2: Not readable
- P3: Yes
- P4: Yes
- P5: My personal weights represent my personal assessment, the overall weights are also still acceptable for me
- P6: Blank

11. **Do you think that the procedure to assess the importance of different criteria is valuable and efficient? Do you believe it could be helpful in other similar assessments in healthcare?**

- P1: Yes, I can imagine it would be valuable in future assessments
- P2: I am not sure
- P3: Yes
- P4: I believe it could be very valuable
- P5: Blank
- P6: Blank

12. **Other comments/thoughts?**

- P1: It was very interesting
- P2: The whole evaluation seems very difficult for people my age or older, and you need so much concentration for the answers to be consistent. It does not seem to be realistic to do these assessments with older people. Furthermore, some criteria are quite dependent on each other, e.g. mobility and balance or depression and everything else. Then it seems difficult to say which is more important. And the clinical outcome measures seem to be both important and critical at once, because in my eyes, mobility would mean a lot more than just walking faster for 5 metres
- P3: Physical and psychological deficits should be further examined and defined
- P4: Blank
- P5: Blank
- P6: Blank

Appendix G: Plan Panel Sessions Nutricia

Appendix G offers a plan for the panel sessions for Nutricia in September 2014. These sessions will focus on the assessment of clinical outcome measures and criteria regarding frailty in the elderly. The panel sessions are designed to include relevant stakeholders in the assessment of frailty criteria and clinical outcome measures. The goal is to gain valid and valuable data for an upcoming research regarding the effectiveness of several nutritional compounds in the prevention of frailty in the elderly.

Panel members (in the following also named participants or decision makers) will be invited by Nutricia and will consist of payers and clinicians in the first panel and patients and carers in the second panel. Prior to the panel sessions, payers and clinicians will receive information about the procedure, the two patient profiles and the clinical outcome measures to provide an introduction of the topic and to clarify potential issues. At the start of the panel session, panel members will be greeted and announced, the researchers will introduce themselves and a short introduction to MCDA and MACBETH will be provided. A tablet will be handed to each participant, the audience response system "QuestionPress" will be explained and a short example of how to use this tool will be given. The next step is going to be the presentation of the decision structure, the semantic categories available for the judgements and the organisation of necessary judgements.

In the first panel session with payers and clinicians, clinical outcome measures will be assessed with respect to the frailty criteria using pairwise comparisons. Assessments are only necessary, if more than one outcome measure is proposed for the evaluation of one criterion.

As an example, the panel members will be asked:

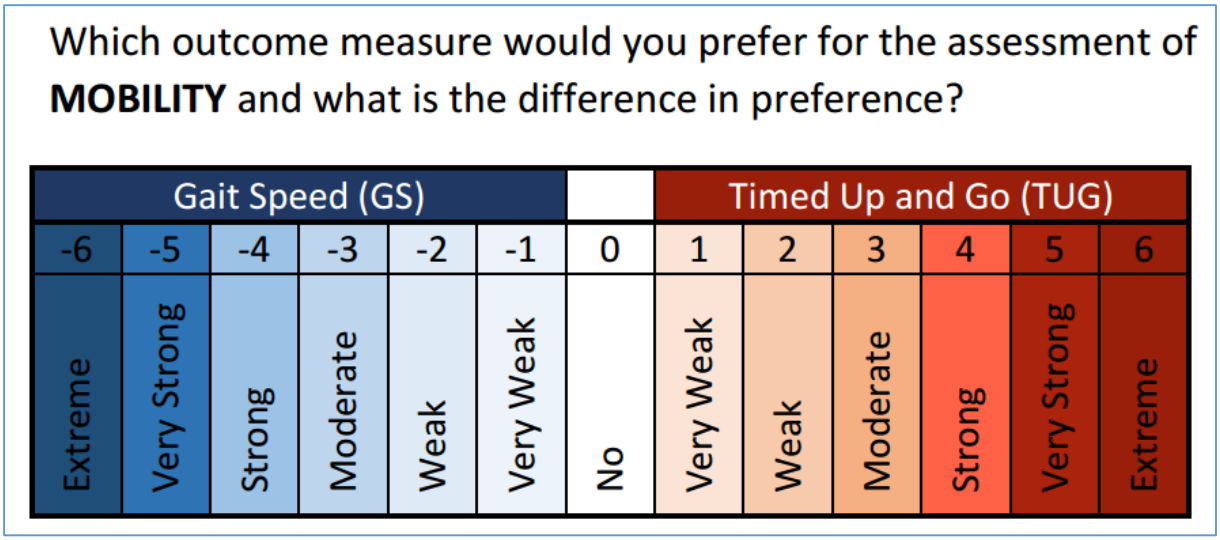


Figure 12: Example of questioning procedure.

Table 12: Semantic categories for the judgements.
According to Bana e Costa and Vansnick [80].

Number	Semantic Categories
C_1	Very weak difference in attractiveness/preference/importance
C_2	Weak difference in attractiveness/preference/importance
C_3	Moderate difference in attractiveness/preference/importance
C_4	Strong difference in attractiveness/preference/importance
C_5	Very strong difference in attractiveness/preference/importance
C_6	Extreme difference in attractiveness/preference/importance

Decision makers will be able to vote for the most attractive outcome measure using "QuestionPress". Results of the votes will be displayed on screen in a bar chart. Decision makers will then have time to discuss their opinions and revise their individual judgements in a second vote, if conflict of opinion is high. The discussions and votings are then repeated. If a consensus is found, the next pairwise comparison can take part. In case of a continuing very strong disagreement after three rounds, the majority rule will be adapted.

The final judgements of the pairwise comparisons are going to be inserted into the M-MACBETH software to check for consistency and to gain priorities for each clinical outcome measure. Once all the necessary pairwise comparisons are gathered, the most important clinical outcome measures of each frailty criterion will be presented on the screen. It will follow the adaption of the most important outcome measures in the patient profiles, eliminated outcome measures will be deleted. The reduced patient profiles will then be displayed on screen. Participants will be asked to evaluate and discuss the profiles with respect to realistic improvements on each outcome measure and make modifications if necessary.

The next step in the MACBETH procedure would usually be the assessment of options with respect to the criteria. However, due to the fact that only two options (patient profiles 1 and 2) are evaluated, the scores 0 and 100 are assigned to the patient profiles 1 and 2, respectively, with respect to each criterion.

Next, the frailty criteria will be weighed using pairwise comparisons. A total of 21 pairwise comparisons will be requested as proposed by Bana e Costa and Chagas [84]. The recommended questioning procedure of Bana e Costa [85] as explained in part 2.2.2 is slightly modified to the following examination:

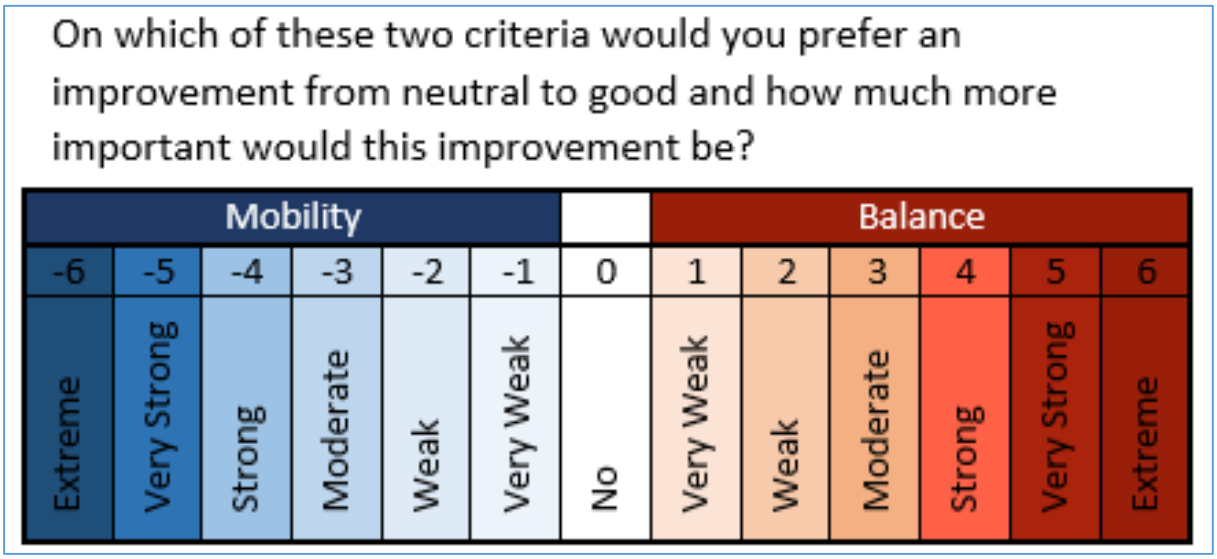


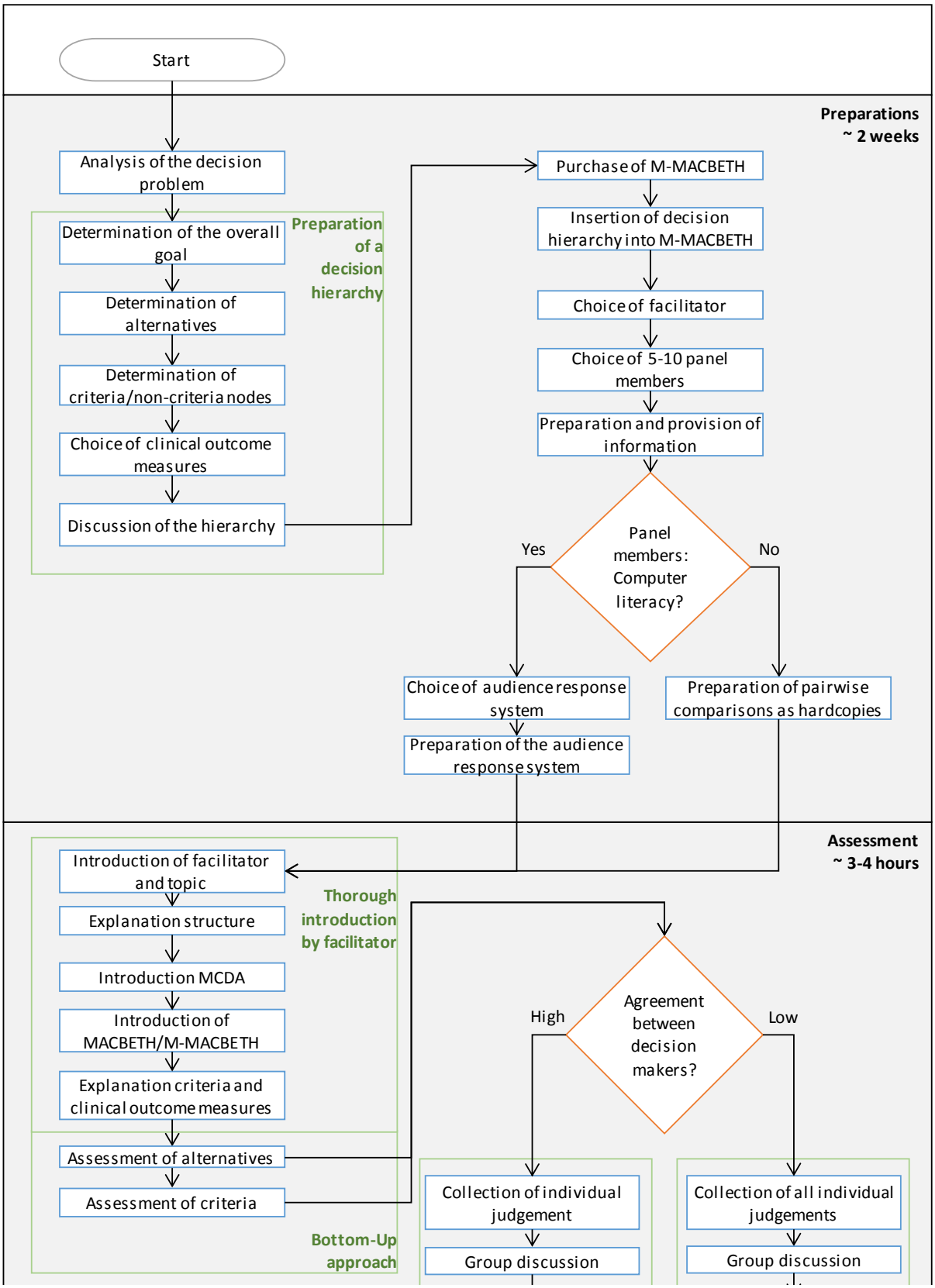
Figure 13: Questioning procedure for the weighting of frailty criteria.

Modifications are the skip of the criteria ranking and the judgement from neutral to good on each criterion, to reduce the number of necessary judgements and to keep an overview in the group decision making process. According to the questioning procedure, all pairwise comparisons will be conducted through "QuestionPress". After this assessment a short break will take part, in which researchers are going to insert the judgements of the decision makers into the M-MACBETH software. After the break the results of the criteria weighting will be shown on screen to the panel members and necessary adjustments will be discussed to reach consistency. Decision makers are then also able to discuss and adjust weights until a consensus is reached. Once all questions are answered and all comments are gathered, the panel members will be dismissed.

The second panel session will basically follow the same procedure. The only differences are that decision makers then consist of patients and carers instead of payers and clinicians and that the assessment of suitable clinical outcome measures for each frailty criterion will be skipped due to the focus on clinical experts in the evaluation of this first part.

Following the panel sessions, sensitivity and robustness analyses will be performed with M-MACBETH to investigate the effects of changes of different variables and uncertainty in the decision making process. In addition, individual judgements of the decision makers will be evaluated and compared to the outcome of the group decision making. Difficulties in the assessment of both clinical outcome measures and frailty criteria with respect to the MACBETH methodology will be analysed by means of participants' comments, a qualitative questionnaire and researchers' perceptions.

Appendix H: Flow Chart of Modified Guidelines



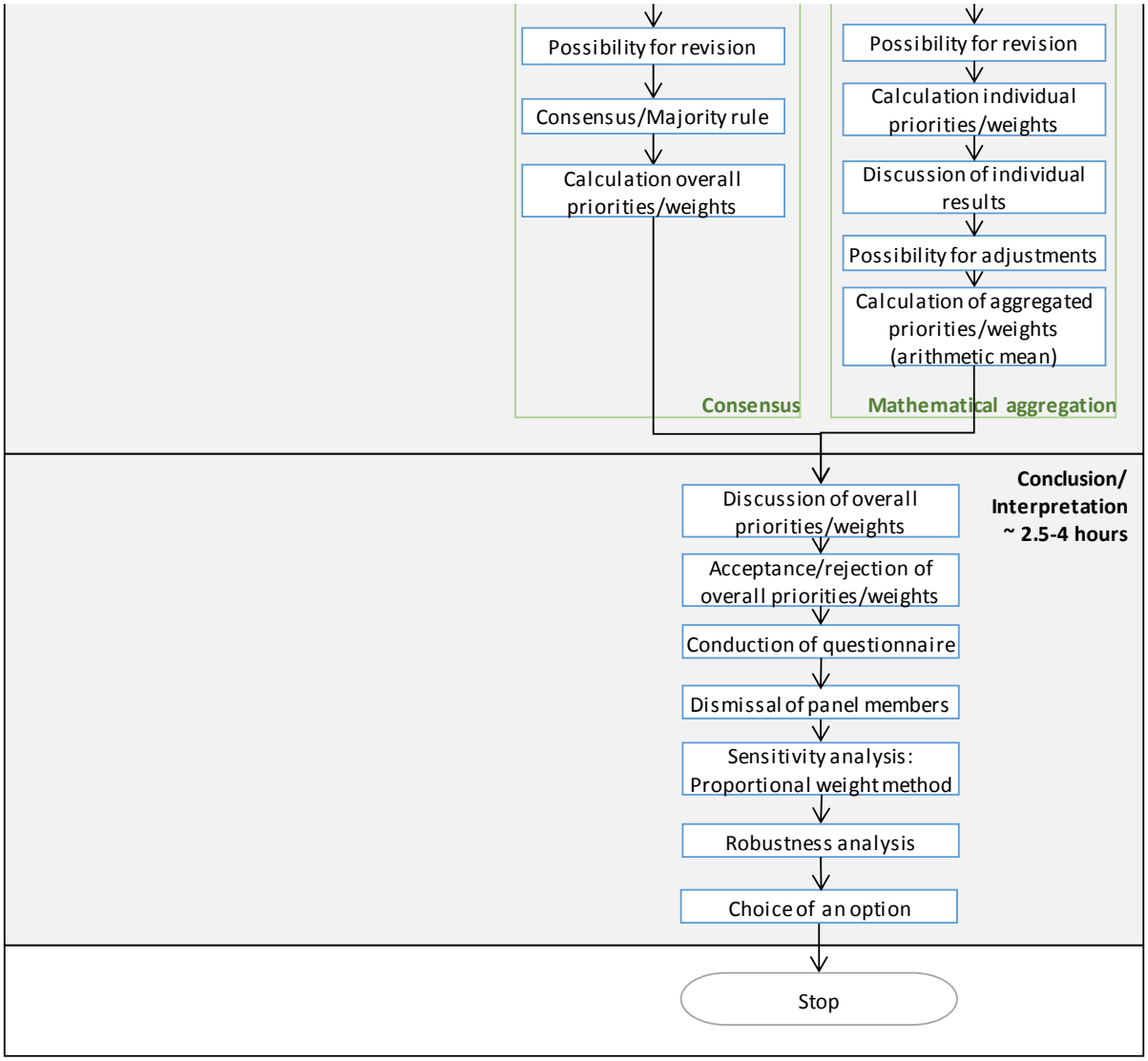


Figure 14: Flow chart of modified guidelines.

Table 13: Distances of individual and aggregated weights.

Distance							
Criterion	**Participant 1**	**Participant 2**	**Participant 3**	**Participant 4**	**Participant 5**	**Participant 6**	**Sum**
Cognition	2.40	2.64	2.70	4.11	3.41	15.28	30.55
Balance	10.60	3.54	6.73	1.40	3.10	4.18	29.55
Energy	4.14	0.24	7.54	6.25	3.48	12.90	34.55
Mobility	1.28	2.85	0.14	0.16	1.70	5.87	12.01
Depression	5.89	2.90	7.67	10.11	9.30	5.51	41.38
Hypertension	1.39	1.29	0.41	2.92	3.96	1.22	11.19
Grip Strength	3.14	2.12	8.04	1.25	0.01	8.24	22.80
Fluctuating Glucose Levels	4.73	2.25	0.69	1.21	1.78	7.09	17.75
Hypercholesterolemia	3.28	3.68	3.73	3.86	5.03	9.50	29.07
Sum	**36.85**	**21.51**	**37.65**	**31.27**	**31.79**	**69.79**	**228.85**
Euclidian distance	**14.71**	**7.80**	**15.73**	**13.66**	**12.92**	**26.31**	