DEVELOPMENT OF THE GRADE APPROACH FOR PATIENT VALUES AND PREFERENCES EVIDENCE

DEVELOPMENT OF THE GRADE APPROACH FOR PATIENT VALUES AND PREFERENCES EVIDENCE

By Yuan Zhang, M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University Doctor of Philosophy (2017)

Hamilton, Ontario (Health Research Methodology)

TITLE: Development of the GRADE approach for patient values and preferences evidence

AUTHOR: Yuan Zhang, M.Sc. (Peking University Health Science Center)

SUPERVISOR: Professor Holger J Schünemann

NUMBER OF PAGES: xx, 218

ABSTRACT

Background and objectives:

Incorporating patient values and preferences as an essential input for decision-making has its potential merits in respecting the autonomy of patients, improving adherence and clinical outcomes. The Grading of Recommendations Assessment, Development and Evaluation (short GRADE) working group conceptualizes patient values and preferences as "the relative importance patient place on the main outcomes". The objectives of this thesis include: 1) to provide an overview of a process for systematically incorporating values and preferences in guideline development; 2) to conduct a systematic review on outcome importance studies, using chronic obstructive pulmonary disease (COPD) as an example; 3) to provide guidance on how to assess certainty of evidence describing outcome importance using the GRADE criteria.

Methods:

We performed systematic reviews, asked clinical experts to provide feedback according to their clinical experience, and consulted patient representatives to obtain information about relative importance of outcomes in a new national guideline program. We conducted a systematic review to summarize the COPD related relative importance of outcome studies. We used a multi-pronged approach to develop the guidance for assessing certainty of evidence about relative importance of outcome and values and preferences. We applied the developed GRADE approach to relative importance of outcome systematic review

examples and consulted the stakeholders in the GRADE working group for feedback.

Results and conclusion: We provided an empirical strategy to find and incorporate values and preferences in guidelines by performing systematic reviews and eliciting information from guideline panel members and patient representatives. However, we identified the need for researches on how to assess the certainty of this evidence, and best summarize and present the findings. In our comprehensive systematic review project on COPD patient values and preferences we demonstrated the utility of rating evidence in systematic reviews of outcome importance.

We describe the rationale for considering GRADE domains for the evidence about the importance of outcomes. We propose the assessment of the body of evidence starts at "high certainty", and rate down for serious problems in GRADE domains including risk of bias, indirectness, inconsistency, imprecision and publication bias. Specific to risk of bias domain, we propose a preliminary consideration for risk of bias. The sources of indirectness for relative importance of outcome evidence include indirectness from PICO (population, intervention, comparison, and outcome) elements, and methodological indirectness. As meta-analyses are uncommon when summarizing the evidence about relative importance of outcome, inconsistency and imprecision assessments are challenging. Inconsistency arises from PICO and methodological elements that should be explored. The width of the confidence interval and sample size should inform judgments about imprecision. We also provide suggestions on how to detect publication bias based

on empirical information. Finally, we also discuss the applicability of domains to rate up the certainty.

We develop the GRADE approach for rating risk of bias, indirectness, inconsistency, imprecision and other domains when evaluating a body of evidence describing the relative importance of outcomes. Our examples should guide users and provide a basis for discussion and further development of the GRADE system.

Preface

This thesis has been conducted as a "sandwich thesis" and consists of three individual manuscripts/papers submitted to journal for publications. These are:

1) Chapter 1: Introduction of the thesis
2) Chapter 2: Using Patient Values and Preferences to inform the importance of health outcomes in Practice Guideline Development: Experiences following the GRADE approach
3) Chapter 3: Development of GRADE guidance for assessing the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 1. risk of bias and indirectness
4) Chapter 4: Development of GRADE guidance for assessing the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 2. Inconsistency, Imprecision and other issues
5) Chapter 5: Relative importance of outcomes (values and preferences) for Chronic Obstructive Pulmonary Patients: A systematic review
6) Chapter 6: Conclusion

At the time of writing, Chapter 2 has been accepted for publication, Chapter 3 and Chapter 4 have been approved by the GRADE working group. Chapter 5 has been ready for submission.

Copyright permission:

Mr. Zhang has provided written permission to include copyright material in this Ph.D. thesis from the copyright holder. The permission includes a grant of an irrevocable, non-exclusive license to McMaster University and to Library and Archives Canada to reproduce the material as part of the thesis.

Acknowledgements:

There are so many people to thank for helping me during the last few years. The thesis work would have been impossible without enormous physical and spiritual supports I received from these wonderful people.

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Holger Schünemann for the continuous guidance and motivation during the time of conducting research and writing of this thesis, as well as support at and outside of work. His support is essential for the accomplishment of this thesis work. He helps me lay the foundation for my career and his inspiration would be my assets in the future work. Besides, I also would like to thank my committee, Dr. Gordon Guyatt, Dr. Pablo Alonso-Coello, and Dr. Amiram Gafni, for their endless support, and sometimes challenging questions to guide me through my academic work. What I learnt from all my supervisors is not only how to conduct research, but also a broad perspective, and a collaborative attitude. Without their support, it would be impossible to finish this work!

I would also like to thank all the friends and colleagues for their extremely hard working and firm support in my thesis projects. I am grateful for all the laborious work, and all the major and minor suggestions for my projects.

My family, they always love me and support me, even when we are apart.

Of course, I also have the supports from the friends and staff in the department of Health Research Methods, Evidence and Impact. It is wonderful experience to know these people, who speak different languages, are from different origins, and play different roles in my life. But I really appreciate that I have the chance to know these wonderful people in this journey.

TABLE OF CONTENTS

List of Tables

LIST OF FIGURES

LIST OF APPENDIEXS

List of Abbreviations

AUR: acute urinary retention

CERQUAL: The Confidence in the Evidence from Reviews of Qualitative research

CINAHL: Cumulative Index to Nursing and Allied Health Literature

CI: confidence interval

COPD: Chronic Obstructive Pulmonary Disease

DCE: discrete choice experiment

DECIDE: Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence

EBM: evidence-based medicine

EQ-5D: EuroQual-5-dimension (a quality of life measurement tool)

EtD: Evidence-to-Decision

FEV: Forced expiratory volume

GIN: Guidelines International Network

GRADE: Grading of Recommendations Assessment, Development and Evaluation

GOLD: Global Initiative for Chronic Obstructive Lung Disease

HRQoL: health-related quality of life

HUI: health utility index

IQR: interquartile range

MeSH: Medical Subject Headings

NICE: National Institute for Health and Clinical Excellence

PICO: population, intervention, comparison and outcome

PRISMA: Preferred Reporting in Systematic Reviews and Meta-Analyses

QWB: quality of wellbeing

RCT: randomized controlled trial

SF-6D: Short form-6-dimension (a quality of life measurement tool)

SG: standard gamble

SGRQ: St. George's Respiratory Questionnaire

SLIT: sublingual immunotherapy

TTO: time trade off

VAS: visual analogue scales

VTE: venous thromboembolism

WHO: World Health Organization

Declaration of Academic Achievement

My supervisor, Prof. Schünemann, and myself played the primary role in the conception of the study, with the aid from Prof. Gordon Guyatt and Dr. Alonso Pablo-Coello. This dissertation represents the original research that I conducted. I am the principle contributor and first author of all the manuscripts contained in this thesis. I also made following contribution in all projects included in the dissertation: drafting protocols, designing search strategies for the systematic reviews, designing screening and data abstraction forms, pilot testing, calibration and management the team of co-authors, conducting analyses, designing figures and tables, organizing the mythological discussion meetings. I wrote the manuscript with editorial advice and supervision of Prof. Schünemann, Dr. Pablo-Coello, and Prof. Guyatt. Prof. Gafni provided valuable comments for editing the manuscripts. The co-authors contributed in selecting and acquiring the data, providing inputs on the GRADE approach proposed, commenting manuscripts in preparation for submission for publication. Details of the contributions of co-authors are described at the end of each manuscript. For all the three manuscripts, earlier drafts of parts of this research have been presented at international academic conferences as part of the manuscript's development.

Chapter 1. Introduction

1.1. Incorporation of patient values and preferences in decision making about health

Evidence-based medicine (EBM) as the "conscientious and judicious use of current best evidence from clinical care research in the management of individual patients" is founded in the integration of the best evidence with individual clinical expertise and patients' choices based on their values and preferences (Figure 1.1).[1-3] EBM has seen many efforts of describing on how to incorporate patient values and preferences into the clinical decision making process. Historically, there is no clear definition of "values and preferences". The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group is defining values and preferences as "relative importance people place on the outcomes of interest", which we would further discuss in details.

Despite these efforts to incorporate values and preferences into clinical decision making, Charles et al argued that the components of evidence-based medicine models were not well defined and justified.[4] In addition, how those components interact with each other and contribute to the decision making process was initially unclear. Further evolution of the EBM model incorporated additional factors to make informed health decisions. In the fully evolved model of

evidence-based decision making (Figure 1.2), it is recognized that health care decision makers, i.e. professionals and patients, incorporate evidence that goes beyond knowledge about intervention effects to arrive at the best possible decision. This expanded model to illustrate evidence based decision making includes the detailed consideration of evidence about implementation and integration strategies, values and preferences, prognosis and context for a given individual.[5] Incorporating patient values and preferences in health care decision-making has the potential to increase the likelihood that patients' feelings are respected and patients' are more satisfied with their decisions.[6-9] These decision-making scenarios include decisions made in the individual patient-physician level, or in the group level, as guideline recommendations developed by guideline panel members.

This thesis focuses on the exploration of the best possible integration of patient values and preferences in health care recommendations that ultimately should lead to appropriate decisions on the individual and population level. Indeed, guideline developers and guideline frameworks have recognized the importance of integrating values and preferences in guideline development for some time.[10] It has become a critical criterion in the GRADE Working Group's Evidence to Decision (EtD) Frameworks, probably the most advanced approach for the development of health recommendations.[11-13] Current guidelines consider and obtain information about values and preferences either by consulting patient representative or patient proxies, or collecting existing evidence on patient preferences. [1 6 7 9 14 15] Despite advances there are a number of shortcomings in the

way that guideline developers currently deal with values and preferences. These shortfalls relate to a clear definition of values and preferences in the guideline development context, how they should be obtained and then considered by guideline panels.

1.2. Defining patient values and preferences

When referring to "patient values and preferences", the focus on "patient" is obvious as patients are the most crucial stakeholders of a healthcare decision. However, the focus on and use of the terms "values and preferences" requires further discussion. According to the Merriam-Webster Dictionary (Table 1.1), the definition of preference includes "something that is liked or wanted more than another thing" while the definition of value includes "usefulness or importance".[16]

In the health research and educational setting, "value" and "preference" have been used synchronously as an umbrella term, even though they could be different in spoken language. In this setting, more than one definition for values and preferences exist. Table 1.2 presents some of these alternative descriptions. [17] It is evident that "values and preferences" are ambiguous rather than self-explanatory terms and challenges surround the multitude of the terms' meaning.[18] Firstly, when health researchers and guideline developers use the terms "patient values and preferences", they often relate to concepts of a patient's choice, expectation, experience, need, perspective, and view. Often "consumer", "public", "recipient", "social" "societal" and "user" are interchangeably used with "patient". Moreover, the measurement or elicitation strategies of "values and preferences"

vary and researchers have failed to agree on the best and most appropriate methods until now (Table 1.3).[19-22] Thus, there is a fair degree of confusion of the terms and concepts although they require an exact definition and methodological approach for the purpose of developing health recommendations.

Secondly, the scope of values and preferences is not clearly defined. Health researchers use the concept of "patient values" as "ethical value" of an intervention or program, and "social value" of research. The former relates to the ethical aspects physicians and other healthcare decision makers may consider, while the latter indicates the implication of research.[18]

Thirdly, whose perspective is considered should be made clear. The interchangeable use of "consumer", "public", "recipient", "social" "societal" and "user" with "patient" is inappropriate because they not necessarily have the same meaning to different stakeholders. The implications of using "values and preferences" may differ or even oppose each other when we compare the societal perspective or general population perspective with that of patients. And, it is not unusual that patients do not share the same view with their health professionals.

1.3. Outcome importance valuation: the GRADE Definition

The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group operationalizes the terms "values and preferences" in this context as the "relative importance people place on the outcomes of interest". That is, how much [health] value patients would put on each of those outcomes.

When we consider values and preferences as the relative importance of outcomes (or consequences and health states), we are assuming that individuals would weigh alternative management options on the basis of the importance of the outcomes the options incur. Thus, the choice of an option, e.g. alternative treatments or tests, and the preference for one or the other will be determined by the importance individuals place on the outcomes that the options will incur. For example, when choosing between a medication and surgery for management of a health condition, the choice between these alternatives is determined not by the intervention itself but by the considerations of the subsequent outcomes such as perioperative pain and complications, the burden of taking a pill and anticipated short and long term health outcomes. This implies individuals would not make decisions according to the importance of a single outcome but the decision making process will be influenced by the importance of all anticipated health benefits and harms of one option in relation to one or more alternatives.[23] And, different alternatives will incur different sets of outcomes.

Conceptually, the relative importance of outcomes (RIO) could be measured under a real decision making scenario, which asks individuals to weigh the health states incurred; or it could be measured by asking individuals about the importance he or she places on a certain health state without facing a specific decision. Throughout the following chapters, we will use the meaning of the terms as described here and in Table 1.4. [24]

It is necessary to explain why the concept of the importance people place on outcomes is preferred over the importance of the interventions and management

strategies. Addressing the outcomes an intervention incurs is explicit and ensures that, when measuring preferences for interventions, differences in the results are not influenced by alternative interpretations of what the intervention entails but be based on a clear description of the anticipated outcomes. What follows is that focusing on outcomes rather than interventions allows for explicit considerations of all outcomes rather than a less defined Gestalt of anticipated outcomes and associated consequence of an intervention that differs across individuals. This approach is consistent with the Grossman Health Capital Model, which described that the demand for healthcare (interventions) is a derived demand, and it is derived from the demand for health (outcomes) itself. [25 26]

1.4. Measurements

In an ideal decision making scenario, guideline panels as one group of decision makers would know the distribution of the relative desirability through direct choice studies recruiting a large optimally informed sample with knowledge of all expected outcomes. Optimally informed in this context means providing information about the exact probabilities and nature of the outcomes. At present, such information is very rarely available and an alternative approach, albeit less direct, is that of providing decision makers with a clear description of the anticipated outcomes and their frequency. To achieve this this, decision makers are eliciting or inferring the relative importance patients place on outcomes based on a variety of approaches. Table 1.5 provides an overview of these quantitative and qualitative approaches to elicit the relative importance of outcomes.

1.4.1. Relative importance of outcomes based on direct utility measurements and related instruments

From a strict health economic perspective, only measurements made under uncertainty, e.g. through the standard gamble, generate true utilities; otherwise, health economists refer to health state values.[27 28] The standard gamble is based on the notion that people would make rational choices when they deal with uncertainty, that is, the axioms of Von Neumann–Morgenstern utility theory (or expected utility theory) are fulfilled.[29] The standard gamble involves a trade-off between two alternatives: a health state that is certain and a gamble of a better (immediate full health with a probability of p) and a worse outcome (immediate death with a probability of 1-p). The probabilities are altered systematically. When respondents become unable to decide between the two alternatives (the probabilities of immediate full health and immediate death) the probability p will be translated into the utility that is placed on the health state.[**28-30**] The major criticism of the standard gamble is its complexity. Few people are capable of understanding or dealing with probabilities and, for this and other reasons, this approach is difficult to execute.[31 32] Thus, although standard gamble utilities, in theory, are by many considered the reference standard, they are rarely available and not without challenges. [33] Two other widely used techniques that measure the value of health outcomes directly are the rating scale and its variants (e.g., visual analogue scale (VAS) or feeling thermometer) and the time trade-off (TTO).[27 28] The VAS receives criticism because it does not fulfill the Von-Neumann Morgenstern axiom as it does not deal with uncertainty or choices and, therefore,

does not provide true utilities. TTO methods are also not incorporating uncertainty.

There are additional approaches. Discrete choice exercise can also produce utility to indicate the relative importance, although the utility is not anchored to a 0 to 1 scale with 0 suggesting death and 1 suggesting perfect health. With discrete choice exercise, the researcher would systematically change the level of at least one attribute or outcome, and ask participants to choose based on the head-to-head comparison. This category includes methodology such as willingness to pay, paired comparison, ranking, or probability trade-off.[34]

The willingness to pay approach measures how important different attributes or components are in terms of monetary values. Applying the willingness to pay, researchers could measure how much money an individual is willing to sacrifice to get one desirable outcome or avoid one undesirable outcome. However, none of the aforementioned approaches provides clear arguments for their use in the context of considering values in guideline development. In this work we utilize the term values in an overarching manner to include these true utilities and other types of health state values derived with the described approaches.

1.4.2. Relative importance of outcomes based on indirect utility measurements

Other alternatives such as multi-attribute utilities or mapping results based on a health-related quality of life measurement suffer from similar concerns of not incorporating uncertainty.[35] Additionally, multi-attribute utilities cannot directly reflect the values patients place on the outcomes. Rather, multi-attribute utilities

are derived from rating a set of attributes or a single health state with multiple dimensions.[32] The health related quality of life (HRQoL) instruments include the generic instruments such as Quality of Well-Being Scale (QWB), the EQ-5D, the SF-6D and the Health Utility Index (HUI), or disease specific HRQoL instruments such as the St. George's Respiratory Questionnaire (SGRQ) for respiratory disease.[36] To derive values, researchers construct an algorithm or tariff to link the answers on the five questions to results from a time trade-off or VAS based on a norm.[37] Following this strategy, scores of these instruments can be converted to health values based on mathematical and statistical models.

In practice, these instruments can also be combined with the standard gamble, TTO, and VAS.[35] Taking EQ-5D as an example, respondents answer five generic questions on five dimensions: mobility, self-care, usual activities, pain/discomfort and anxiety/depression; these questions may be accompanied by time trade-off questions or a VAS to elicit health state values.

1.4.3. Other quantitative measurements

Other structured instruments or questionnaires respondents answer how desirable or aversive a certain outcome could also suggest importance of outcomes.

1.4.4. Health values from qualitative values

Qualitative studies can provide information on patient values and explore the variation and reasons for variation in the decision making process. In an example of qualitative narrative preference, Borres et al. reported that patients failed to use nasal sprays daily because they felt it as "inconvenient and embarrassing".[38]

1.5. Summary of background for this thesis

Considering the importance people place on outcomes is critical for decision-making, including for healthcare recommendations. Multiple methods exist to obtain information about the relative importance of outcomes, which relate to the general concept of values and preferences. These methods have strength and limitations and so do the studies that are conducted employing these methods to obtain the evidence about values and preferences. In addition, individuals may and often will have different views on relative importance of outcomes.

Utilizing this information to provide health recommendations has its challenges, which are not well explored. The GRADE working group's EtD frameworks require explicit consideration of criteria including the evidence on effectiveness and adverse effects, importance on outcomes, cost, equity, feasibility and acceptability, and certainty of evidence to aid guideline panels with formulating the recommendation.[11 12] Balancing the relevant evidence that informs these criteria is meaningless without information about the underlying values and preferences to which we will refer to as "relative importance of [health] outcomes". This information comes from the approaches that we described above. This thesis explores a series of critical issues that will help with making the process of including relative importance of outcomes in the EtD frameworks more transparent.

1.6. Objectives and outlines

This series of work is based on the assumption that we can address the umbrella term of patient values and preferences by understanding the importance patients place on the health outcomes. Additionally, this series of work will not discuss the theoretical basis of how to reach decision based on all the criteria in the EtD framework. Relevant information but not the importance patients place on outcomes would fall into the scope of feasibility and acceptability.

On this basis, the objectives of this thesis include:

1. To provide a practical example for systematically incorporating the relative importance of outcomes in guideline development;
2. To develop the domains that are relevant for assessing the certainty of evidence in the relative importance of outcomes using the GRADE criteria;
3. To provide guidance on how to operationalize these GRADE domains;
4. To conduct a systematic review on the relative importance of outcomes and apply the GRADE domains using a common chronic disease as an example

The thesis is composed of four articles (Chapters 2 to 5 of the thesis). In Chapter 2, we performed systematic reviews, asked clinical experts to provide feedback according to their clinical experience, and consulted patient representatives to obtain information about relative importance of outcomes in a new national guideline program. Chapter 3 explores how the GRADE domains risk of bias (limitations in the detailed study design and execution), inconsistency, imprecision, indirectness and publication bias apply to evidence dealing with

relative importance of outcomes. It describes the approach and the operationalization for two of these five domains. Chapter 4 provides an operationalization of the remaining domains and explores the challenging issues surrounding imprecision and true variability of relative importance of outcomes. Chapter 5 reports on the application of the approach in a comprehensive systematic review of relative importance of outcomes in patients with chronic obstructive pulmonary disease (COPD). While addressing the application of GRADE as a case example, it also deals with how to systematically identify research evidence about the relative importance of outcomes.

Thus, the work presented here aims to provide guidance in acquiring, appraising and applying evidence about relative importance of outcomes, with an emphasis on using the GRADE approach for determining the certainty of evidence of relative importance of outcomes in developing health recommendations.

Table 1.1. Definitions in the dictionary[16]

Term	Definition in the dictionary
Preference	• a feeling of liking or wanting one person or thing more than another person or thing • an advantage that is given to some people or things and not to others • something that is liked or wanted more than another thing: something that is preferred
Value	• the amount of money that something is worth: the price or cost of something • something that can be bought for a low or fair price • usefulness or importance

Table 1.2. Comparison of several values and preferences definitions

Term	Definition	Source
Patient values and preferences	*"We use values and preferences as an overarching term that includes patients' perspectives, priorities, beliefs, expectations, values and goals for health and life.* *We also use this phrase, more precisely, to mean the process that individuals use in considering the potential benefits, harms, costs, and inconveniences of the management options in relation to one another."*	*Users' Guide to the Medical Literature: A manual for Evidence-Based Clinical Practice*[23]
Patient values	*"The unique preferences, concerns and expectations each patient brings to a clinical encounter and which must be integrated into clinical decisions if they are to serve the patient."*	*KT Clearinghouse*[17]

Table 1.3. Terms used in guideline development manuals and methodological papers

Category	Term used
Choice	*patient choice; personal choice*
Expectation	*consumer expectation; patient expectation*
Experience	*experience of recipients; patient experience*
Involvement	*consumer involvement; patient involvement*
Need	*patient need*
Preference	*health state preference; patient preference; personal preference; preference of recipient; public value*
Utility	*average utilities of the population; disutility; health state utility value; patient utility; utility or utility values;*
Value	*consumer value; health state value; local value; moral value; patient value; public value; value judgement or social value judgment; value of recipient*
View	*consumer view; patient view*

Table 1.4. Terminology used

Terminology	Scope or definition
Outcome	The term outcome includes "***health state***" and non-health states. This includes a broad set of the outcomes directly and indirectly related to a disease or health, an intervention, or non-health consequences. Outcomes can be more or less health-related. For example, from mostly health related to least, patients will have their views regard on the importance of the following outcomes: breathlessness, treatment burden of warfarin or insulin injection, ease of reaching a clinic to undergo blood tests and other monitoring.
Relative importance of outcomes	The ***relative importance of outcomes*** is interchangeably used with ***values and preferences***, ***outcome importance***, or ***outcome valuation***: GRADE defines values and preferences as the relative importance of outcomes.[20]
Methodology (for determining the relative importance of outcomes)	This term, when used referring to measuring relative importance of outcomes, refers to "***measurement tool***" "***measurement methods***" or "***measurement instruments***".
Certainty of evidence	This term is interchangeably used with "***quality of evidence***", ***"strength of evidence",*** and "***confidence in estimate***". ***Certainty of evidence*** has different meanings for systematic reviews and guideline development. For systematic reviews the definition is: The extent of our confidence that the relative importance of the outcomes (and variability) lie in a particular range; for guidelines the definition is: The extent of our confidence that the estimate of the relative importance of the outcomes (and variability) are adequate to support a particular recommendation.[24]

Table 1.5. Different measurements of patient values and preferences

QUANTITATIVE RESEARCH		
	Explanation	**Examples**
Utilities (direct methods)		
Matching methods	Respondents are asked to provide a number (or numbers that will make them indifferent to the good outcome to be valued).	• Time trade off • Willingness to pay • Standard gamble • Allocation game • Visual analogue scale*
Discrete choice exercises and/or conjoint analysis△		• Binary choice experiments • Multinomial choice experiment • Best-worst choice experiment • Full ranking exercise • Probability trade off • Conjoint analysis
Utilities (indirect methods)		
Multi-attribute utility instruments	Respondents describe their health state; value is calculated using a formula that considers the general population preferences.	• For example, EuroQoL-5D, Health Utilities Index 2, or Short form-6D.
Use of health related quality of life tools results	Calculation of utilities using techniques.	
Non-Utilities outcome importance		

Other methods	Because the concept of risk is not incorporated, strictly these methods not provide utilities in the economic sense. We consider them valid for developing recommendations since they still suggest the outcome importance.	• Adaptive questioning • Rating (e.g. numerical rating scales) • Ranking • Direct choice
QUALITATIVE RESEARCH		
	Text-based and interpretative.	• Focus groups • Open, structured or semi-structured interviews • Observation

* Visual analogue scale does not provide utility in economic perspective, but the health state value it provides can be a replacement for utility.

△For all the methods, there should be "at least one attribute of the alternatives systematically varied across respondents in such a way that information related to preference parameters of an indirect utility function can be inferred"

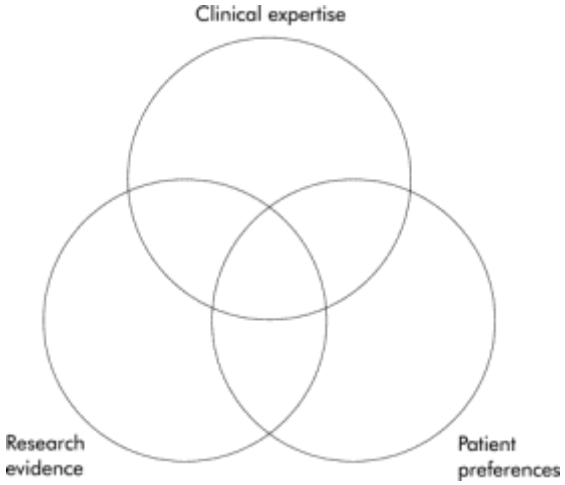


Figure 1.1. Early model of the key elements for evidence-based clinical decisions

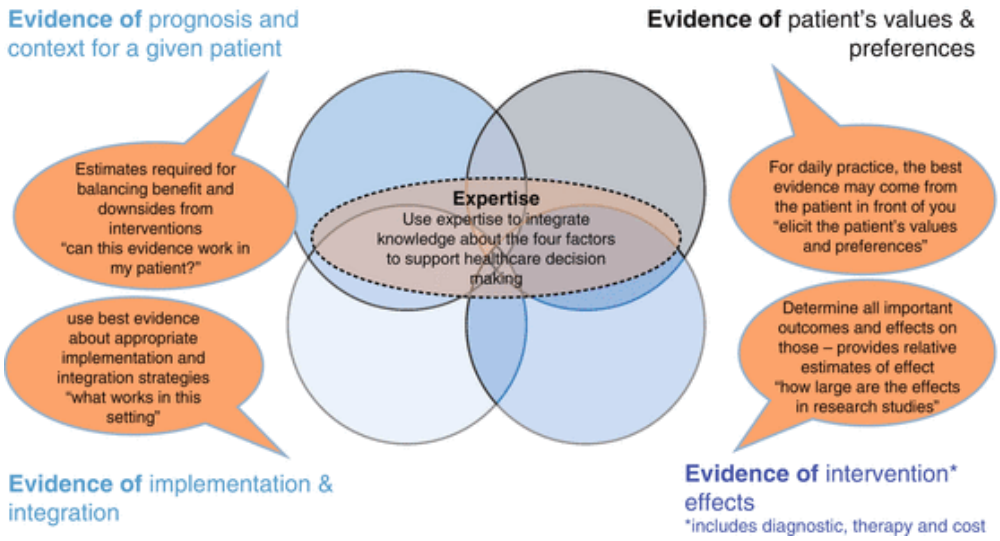


Figure 1.2. Fully developed model of evidence-based decision making[5]

1.7. Reference

1. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed)* 1996;312(7023):71-2. [published Online First: 1996/01/13]
2. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *ACP journal club* 2002;136(2):A11-4. [published Online First: 2002/03/05]
3. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *Evidence Based Medicine* 2002;7(2):36-38. doi: 10.1136/ebm.7.2.36
4. Charles C, Gafni A, Freeman E. The evidence-based medicine model of clinical practice: scientific teaching or belief-based preaching? *Journal of evaluation in clinical practice* 2011;17(4):597-605. doi: 10.1111/j.1365-2753.2010.01562.x [published Online First: 2010/11/23]
5. Schünemann H, Guyatt G. Clinical Epidemiology and Evidence-Based Health Care. In: Ahrens W, Pigeot I, eds. Handbook of Epidemiology: Springer New York 2014:1813-73.
6. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA : the journal of the American Medical Association* 2008;300(4):436-8. doi: 10.1001/jama.300.4.436 [published Online First: 2008/07/24]
7. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e1S-23S. doi: 10.1378/chest.11-2290 [published Online First: 2012/02/15]
8. Stiggelbout AM, Van der Weijden T, De Wit MP, et al. Shared decision making: really putting patients at the centre of healthcare. *BMJ (Clinical research ed)* 2012;344:e256. doi: 10.1136/bmj.e256 [published Online First: 2012/01/31]
9. van der Weijden T, Legare F, Boivin A, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. *Implementation science : IS* 2010;5:10. doi: 10.1186/1748-5908-5-10 [published Online First: 2010/03/09]
10. Schunemann HJ, Munger H, Brower S, et al. Methodology for guideline development for the Seventh American College of Chest Physicians Conference on Antithrombotic and Thrombolytic Therapy: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. *Chest* 2004;126(3 Suppl):174s-78s. doi: 10.1378/chest.126.3_suppl.174S [published Online First: 2004/09/24]
11. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical research ed)* 2016;353:i2016. doi: 10.1136/bmj.i2016

12. Alonso-Coello P, Oxman AD, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. *BMJ (Clinical research ed)* 2016;353:i2089. doi: 10.1136/bmj.i2089
13. Schunemann HJ, Hill SR, Kakad M, et al. Transparent development of the WHO rapid advice guidelines. *PLoS medicine* 2007;4(5):e119. doi: 10.1371/journal.pmed.0040119 [published Online First: 2007/05/31]
14. Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *JAMA : the journal of the American Medical Association* 2008;300(21):2483; author reply 83-4. doi: 10.1001/jama.2008.730 [published Online First: 2008/12/04]
15. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of clinical epidemiology* 2013;66(7):726-35. doi: 10.1016/j.jclinepi.2013.02.003 [published Online First: 2013/04/11]
16. http://www.merriam-webster.com/. The Merriam-Webster Dictionary
17. Clearinghouse K. What is EBM? http://ktclearinghouse.ca/cebm/intro/whatisebm Accessed on Jan 5th, 2016.
18. Giacomini MK, Cook DJ, Streiner DL, et al. Using practice guidelines to allocate medical technologies. An ethics framework. *International journal of technology assessment in health care* 2000;16(4):987-1002. [published Online First: 2001/01/13]
19. National Institute for Health and Clinical Excellence. The guidelines manual (Nov 2012). London: National Institute for Health and Clinical Excellence. Available from: http://www.nice.org.uk/.
20. Schunemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014;186(3):E123-42. doi: 10.1503/cmaj.131237 [published Online First: 2013/12/18]
21. Treweek S, Oxman AD, Alderson P, et al. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. *Implementation science : IS* 2013;8:6. doi: 10.1186/1748-5908-8-6 [published Online First: 2013/01/11]
22. World Health Organization. WHO Handbook for Guideline Development. 2nd ed. Switzerland: World Health Organization, 2014.
23. Gordon Guyatt DR, Maureen O. Meade, and Deborah J. Cook. Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice. Second ed: McGraw-Hill Education 2008.
24. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]
25. Hurley JE. Health Economics: McGraw-Hill Ryerson, Limited 2010.

26. Grossman M. On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy* 1972;80(2):223-55.
27. Morimoto T, Fukui T. Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. *Journal of epidemiology / Japan Epidemiological Association* 2002;12(2):160-78. [published Online First: 2002/05/30]
28. Birch S, Ismail AI. Patient preferences and the measurement of utilities in the evaluation of dental technologies. *Journal of dental research* 2002;81(7):446-50. [published Online First: 2002/08/06]
29. Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health services research* 1994;29(2):207-24. [published Online First: 1994/06/01]
30. Gandjour A, Gafni A. The additive utility assumption of the QALY model revisited. *Journal of health economics* 2010;29(2):325-8; author reply 29-31. doi: 10.1016/j.jhealeco.2009.11.001 [published Online First: 2009/12/17]
31. Streiner DL, Norman GR. Health Measurement Scales: A practical guide to their development and use. 4th ed: Oxford University Press 2008.
32. Drummond MF, Schulpher MJ, Torrance G, et al. Methods for the Economic Evaluation of Health Care Programmes: Oxford University Press 2005.
33. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical research ed)* 2008;336(7650):924-6. doi: 10.1136/bmj.39489.470347.AD [published Online First: 2008/04/26]
34. Weernink MGM, Janus SIM, van Til JA, et al. A Systematic Review to Identify the Use of Preference Elicitation Methods in Healthcare Decision Making. *Pharmaceutical Medicine* 2014;28(4):175-85. doi: 10.1007/s40290-014-0059-1
35. Torrance GW. Utility approach to measuring health-related quality of life. *Journal of chronic diseases* 1987;40(6):593-603. [published Online First: 1987/01/01]
36. Felce D, Perry J. Quality of life: its definition and measurement. *Research in developmental disabilities* 1995;16(1):51-74. [published Online First: 1995/01/01]
37. Horsman J, Furlong W, Feeny D, et al. The Health Utilities Index (HUI): concepts, measurement properties and applications. *Health and quality of life outcomes* 2003;1:54. doi: 10.1186/1477-7525-1-54 [published Online First: 2003/11/14]
38. Borres MP, Brakenhielm G, Irander K. How many teenagers think they have allergic rhinoconjunctivitis and what they do about it. *Annals of Allergy, Asthma and Immunology* 1997;78(1):29-34.

Chapter 2. Using Patient Values and Preferences to inform the importance of health outcomes in Practice Guideline Development following the GRADE approach

Authors

Yuan Zhang[1], Pablo Alonso Coello[1,2], Jan Brożek[1,3], Wojtek Wiercioch[1], Itziar Etxeandia-Ikobaltzeta[1] Elie A. Akl[1,4], Joerg J. Meerpohl[5,6], Waleed Alhazzani[1,3], Alonso Carrasco-Labra[1,7], Rebecca L. Morgan[1], Reem A. Mustafa[1,8], John J. Riva[1,9], Ainsley Moore[1,9], Juan José Yepes-Nuñez[1,10], Carlos Cuello-Garcia[1,11], Zulfa AlRayees[12], Veena Manja[13,14], Maicon Falavigna[15,16], Ignacio Neumann[1,17], Romina Brignardello-Petersen[1,7], Nancy Santesso[1], Bram Rochwerg[1,3], Andrea Darzi[4], Maria Ximena Rojas[18], Yaser Adi[19], Claudia Bollig[5], Reem Waziry[4,20], Holger J. Schünemann[1]

Author Affiliations:

1. Department of Clinical Epidemiology and Biostatistics, McMaster University, 1280 Main Street West Hamilton, ON, L8N 4K1, Canada
2. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
3. Department of Medicine, McMaster University, Hamilton, Canada
4. Department of Internal Medicine, Faculty of Health Sciences, American University of Beirut
5. Cochrane Germany, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Germany
6. Centre de Recherche Épidémiologie et Statistique Sorbonne Paris Cité – U1153, Inserm / Université Paris Descartes, Cochrane France, Hôpital Hôtel-Dieu, 1 place du Parvis Notre Dame, 75181 Paris Cedex 04, France
7. Evidence-Based Dentistry Unit, Faculty of Dentistry, Universidad de Chile, Santiago, Chile
8. Departments of Internal Medicine/Nephrology and Biomedical and Health Informatics, University of Missouri-Kansas City, Kansas City, Missouri, United States
9. Department of Family Medicine, McMaster University
10. School of Medicine, University of Antioquia. Medellín, Colombia
11. Tecnologico de Monterrey School of Medicine, Monterrey, Mexico
12. Saudi Centre for Evidence Based Health Care, Ministry of Health, Saudi Arabia
13. Division of Cardiology, Department of Medicine, Veterans Affairs Medical Center, Buffalo, New York

14. Department of Internal Medicine, University at Buffalo, the State University of New York, Buffalo
15. Hospital Moinhos de Vento, Porto Alegre, Brazil
16. National Institute of Science and Technology for Health Technology Assessment, Federal University of Rio Grande do Sul, Porto Alegre, Brazil
17. Department of Internal Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile
18. Department of Clinical Epidemiology and Biostatistics, Pontificia Universidad Javeriana. Bogotá. Colombia
19. King Faisal Specialist Hospital and Research Centre
20. The Kirby institute, University of New South Wales, New South Wales, Australia

Corresponding author:

Holger J Schünemann, MD, PhD
Chair, Department of Clinical Epidemiology & Biostatistics
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON, L8N 4K1, Canada
Email: schuneh@mcmaster.ca
Tel: +1 905 525 9140 x 24931

Abstract

Background

There are diverse opinions and confusion about defining and including patient values and preferences (i.e. the importance people place on the health outcomes) in the guideline development processes. This article aims to provide an overview of a process for systematically incorporating values and preferences in guideline development.

Methods

In 2013 and 2014, we followed the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to adopt, adapt and develop 226 recommendations in 22 guidelines for the Ministry of Health of the Kingdom of Saudi Arabia. To collect context-specific values and preferences for each recommendation, we performed systematic reviews, asked clinical experts to provide feedback according to their clinical experience, and consulted patient representatives.

Results

We found several types of studies addressing the importance of outcomes, including those reporting utilities, non-utility measures of health states based on structured questionnaires or scales, and qualitative studies. Guideline panels used the relative importance of outcomes based on values and preferences to weigh the balance of desirable and undesirable consequences of alternative intervention options. However, we found few studies addressing local values and preferences.

Conclusions

Currently there are different but no firmly established processes for integrating patient values and preferences in healthcare decision-making of practice guideline development. With GRADE Evidence-to-Decision (EtD) frameworks, we provide an empirical strategy to find and incorporate values and preferences in guidelines by performing systematic reviews and eliciting information from guideline panel members and patient representatives. However, more research and practical guidance are needed on how to search for relevant studies and grey literature, assess the certainty of this evidence, and best summarize and present the findings.

Keywords

Patient values, patient preferences, outcome importance, systematic review, guideline development, evidence to decision

2.1. Introduction

According to the World Health Organization (WHO), "a recommendation [in a practice guideline] tells the intended end-user of the guideline what he or she can or should do in specific situations to achieve the best health outcomes possible, individually or collectively…".[1] A recommendation does not only depend on the magnitude of an intervention effect, but should incorporate other considerations and criteria that determine the direction and strength of a recommendation, such as the importance or weight of the health outcomes.[2] Recommendations are the deliberate product of inclusively considering these criteria that influence decision-making by a multidisciplinary group through a structured process.[3-6] This multidisciplinary group typically includes content experts, patients, methodologists and other stakeholders.[7-9] These different individuals may choose different treatment options when they are presented with the same evidence. When full understanding of the information is ensured, different choices for recommendations are often the result of disparate values and preferences. Although infrequently practiced, ideally this information should be based on evidence from thoroughly conducted research, which is collected through a systematic approach.[10] The main reason for incorporating values and preferences in guideline development process is that recommendations aligned with patient values and preferences may be more easily accepted, implemented and adhered to by those intended to benefit from the guidelines. Additionally, in the individual physician-patient encounters, recommendations with consideration of patient's preferences, can better inform the decision-making process.[10-15] Further

motivations for incorporating patient values and preferences in guideline include ethical and moral imperatives, accountability and legitimacy of the guideline developers.

The Grading of Recommendation, Assessment, Development and Evaluation (GRADE) working group developed the Evidence-to-Decision (EtD) framework to facilitate the process of integrating the criteria considered necessary in guideline development and documenting such process for different audiences.[3 5] With this framework, to formulate a recommendation, these criteria include: balance between desirable and undesirable effects, certainty in the evidence informing the recommendation, resource utilization, and impact on health system equity, feasibility of the recommendation, stakeholder acceptability, patient values and preferences. A number of tools and initiatives explicitly describe the factors that should be considered when developing recommendations with different stakeholders. These include the development of the Guidelines International Network (GIN)-McMaster Guideline Development Checklist,[4] the presentation methods developed in GRADE's Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE) Project[16] as well as collaborative guideline development activities with professional and governmental organizations. However, we still recognize paucity in practical strategies to incorporate patient values and preferences in the guideline development process.

In fact, the definition of and strategies for determining values and preferences are still under debate. The GRADE approach includes the consideration of patient

values and preferences as the relative importance of outcomes or health states of interest.[2 4 17 18] Similarly, in health economics, preference is a general term that includes health utilities elicited under uncertainty (e.g. results from standard gamble), as well as the values elicited under certainty (e.g. time trade off or visual analogue scale).[19-21] With this GRADE definition, the preference for or against an intervention is conceptually equivalent to the importance placed on outcomes that follow from the decision to undergo an intervention. That is, the preference for or against an intervention is a result of indirectly weighing the health outcomes it causes (e.g. the outcome burden when taking a medication or the consequences of undergoing surgery such as the outcome postoperative pain).[4] Thus, the preference for or against a treatment intervention is an implicit result of the relative importance of the health outcomes an individual connects to the intervention. However, although values and preferences directly relate to the relative importance of health outcomes in practice guidelines, they also implicitly relate to achieving better health outcomes when judging other aspects that are relevant for a decision. These other aspects such as attitudes, expectations and beliefs are also considered under this umbrella term.[22 23] In the GRADE EtD, these aspects often fall within other criteria of the EtD framework (e.g., equity, feasibility or acceptability considerations). For example, if a society places low value on avoiding resource expenditure for wide implementation of a new intervention, it may be considered feasible. Patients may find an intervention administered by a health worker other than a physician not acceptable, if they expect the latter to administer it. Thus, feasibility and acceptability are

considerations related to values and preferences but not as directly related to the importance patients place on the health outcomes.

Box 2.1. Relevant criteria in Evidence-to-Decision Framework

People values and Preferences: the relative importance people place on the health outcomes; since we consider an intervention in the context of the consequences it incurs, the preferences for or against an intervention is a consequence of the relative importance people place on the expected or definite health outcomes it incurs.
Acceptability and feasibility: views or perspectives or importance of health outcomes placed by stakeholders beyond the target population of the recommendation

Despite the increasing importance of practice guidelines in the management of health problems, there is a lack of evidence informing about initiatives using values and preferences in the guideline development process. Therefore, we addressed the challenges of integrating values and preferences in practice guidelines. Generally, we utilized the GRADE system for guideline development that is endorsed by over 100 organizations and applied worldwide.[10] Specifically, we first developed an approach for systematically identifying information on values and preferences. Second, we conducted case studies on how to consider local values and preferences evidence in the guideline development process. Our case studies were based on 22 guidelines with 226 recommendations covering diverse clinical areas in a new national guideline program for the Ministry of Health of Saudi Arabia.

2.2. Methods

For these guidelines, we were specifically interested in identifying values and preferences relevant to the context of the Saudi society. Methodological details of the guideline development process for the Saudi Ministry of Health are described elsewhere.[24 25] The Ministry of Health of Saudi Arabia had embarked on standardizing and coordinating guideline development nationally to promote the awareness and practice of evidence-based medicine. [24 25] In this project, we used the definition of "relative importance of outcomes" for patient values and preferences. We undertook several steps to obtain information about patient values and preferences. We performed a systematic review to summarize relevant studies of values and preferences in populations of interest. In addition, we sought input from clinical experts and consulted patient representatives (see the Figure 2.1. Process of Integrating Values and Preferences). To assess the feasibility of our approach, we also monitored the workload resulting from conducting systematic reviews on values and preferences during guideline development.

2.2.1. Systematic review

Our approach to comprehensively identifying and understanding existing evidence about values and preferences started with a systematic review summarizing the relevant research evidence.[26] Similar to any systematic review process this included formulation of research questions, literature search, screening according to eligibility criteria, as well as appraisal and summary of the available evidence.[8 14 27]

1. Formulation of research question and GRADE definition of values and preferences

We defined the values and preferences as the relative importance of outcomes and formulated the research question for the systematic review of values and preferences as: "what is the relative importance that a population of interest places on the main outcomes?" With this research question, we considered both the studies on the relative importance of outcomes and studies on the preferences for or against an intervention eligible in the 22 guidelines and the detailed recommendations therein.

2. Eligibility Criteria

Studies reporting the "relative importance of outcomes" relevant to the guideline disease topics were included. We included studies that elicited utilities of outcomes through direct measurement techniques including standard gamble, time trade off, visual analogue scales (VAS), and indirect measurement techniques based on generic tools such as EuroQol five dimensions questionnaire (EQ-5D), HUI (health utility index), QWB (quality of wellbeing), as well as utility or health status values transformed (mapping) from quality of life measurement.[28-30] We recognize that not all scientists consider VAS a utility instrument because it does not include a choice under uncertainty. While acknowledging this, we consider VAS measures as eligible to indicate the relative importance of outcomes. Direct choice refers to the technique of asking participants to choose from a set of options. We included studies that expressed the preferences through willingness to pay, probability trade off, discrete choice exercise, ranking, and paired

comparison. We also included studies that used other questionnaires and scales, sometimes self-developed to ask preference for outcomes. We also included studies that measured the importance of outcomes in qualitative studies (See Table 1).[23 31] Eligible studies included either participants who were experiencing the relevant health states or participants who did not experience the health state of interest but were provided with descriptions of scenarios of the health state.[32-34]

3. Literature Search

We conducted 22 systematic reviews on information suggesting the importance of outcomes; one for each guideline. We developed a broad search filter for values and preferences studies for Ovid Medline, EMBASE and PsychInfo, informed by a search strategy utilized in a previous guideline development processes.[14] This search filter included keywords for the following concepts: *health state values, preference, utility, attitude to health, patient decision, patient participation, patient satisfaction, patient view, patient perception* and their variant formats so as to be as inclusive as possible and capture all potential relevant studies (see Supplementary Material). The development of the search strategy is another ongoing project and the detailed development process will be reported in another publication.[35]

In order to address local values and preferences and enhance contextual information, we also added a geographic search filter that restricted the search to the Kingdom of Saudi Arabia and more broadly to the Middle East. Thus, we developed a complex search strategy based on three search filters: a broad values

and preferences filter, the disease specific filters for each guideline, and a geographic filter. These filters were combined using a Boolean "AND".

4. Screening and Data abstraction

We systematically screened titles and abstracts and retrieved studies for full text screening if they were deemed eligible or if the abstract lacked the detail to determine eligibility by at least one of the screeners. We reviewed the full text articles and summarized the findings stratified according to Table 2.1 and incorporated them into the GRADE EtD frameworks for each of the 22 guideline areas. We a priori broadened our inclusion criteria and included indirect evidence from other settings when we did not identify information specific for the Saudi Arabia setting

2.2.2. Input from panel members

Furthermore, we asked guideline panel members (including patient representatives with and without previous experience in the condition of interest) to provide their views on the relative importance of the main outcomes, and their experience related to the disease of interest. We specifically asked clinicians to reflect on patients' views based on their previous clinical interactions with patients. However, we did not conduct *de novo* studies on eliciting values and preferences for these guidelines.

2.3. Results

2.3.1. Findings of the systematic reviews

We identified a wide variety of eligible studies using utility elicitation, non-utility estimates from questionnaires or scales, as well as qualitative research. Due to heterogeneity of designs and outcomes, we did not pool results and thus provided narrative summaries of the results for each topic. We summarized the information in EtD frameworks for each panel to consider and allow for them to provide feedback. Here, we present guideline-specific examples of the identified studies to illustrate our findings. They are based on a description by utility tools that were used in the original studies.

1. **Utility based estimates**

For the antithrombotic guidelines that we produced, utilities for severe, moderate and mild nonfatal intracranial bleeds were identified ranging from 0.10 to 0.51, 0.29 to 0.77 and 0.47 to 0.94, respectively.[36 37] The utility was 0.63 for nonfatal pulmonary embolism, and 0.44 to 0.84 for major bleed. A systematic review on breast lump-related values and preferences reported the following utilities: 0.96 for disease-free survival, 0.76 to local-regional recurrence, 0.72 to contralateral breast cancer and 0.64 to distant metastasis.[36]

2. **Non-utility measurements**

For the guideline on management of breast lump and primary breast cancer, the systematic review identified one study reporting an additional year in life

expectancy or 3% in survival rates was sufficient to make adjuvant chemotherapy worthwhile for 68–84% of women.[36]

3. Qualitative findings

For the guideline addressing the screening and treatment of precancerous lesions for cervical cancer prevention, we identified one qualitative research study suggesting that women fear screening and may have a high level of anxiety related to colposcopy or treatment.[36]

2.3.2. Input from panel members

Our consultations with panel members suggested that they were not aware of any studies that were missed by our systematic review process. We also asked them to indicate if indirect evidence from other settings is applicable to the Saudi Arabia setting. Generally, the panelist did not believe there were significant differences except in a few cases. For example, for breast cancer screening, the panel members suggested that in the Saudi Arabia setting, patients place a lower value for any psychological effect of false positive results and frequency of screening compared to the perceived benefits of screening strategies on mortality. In the venous thromboembolism (VTE) treatment guideline development, panelists reflected that oral anticoagulation requires frequent testing and monitoring, diet and medication restrictions, stoppage for procedures. However, anticoagulation would be given for a relatively limited period of time and patients would view potential reduction in mortality and symptomatic VTE favourably.[36]

In the allergic rhinitis guideline, panel members suggested that some patients in Saudi Arabia would not accept sublingual immunotherapy with some allergens of

animal origin. The panel evaluating hemodialysis options described that: "the preference to delay dialysis may be stronger in Saudi patients compared to non-Saudi patients (i.e. Saudi patients are more hesitant/resistant to start dialysis)".[36]

2.3.3. Use of the information as part of decision-making process

The importance patients place on outcomes influences the balance of benefits and harms thereby impacting on the direction and strength of a health recommendation. Thus, being explicit about the relative importance requires a transparent description of how they influenced the recommendation. The panels were made aware that, following the GRADE approach, high variability or uncertainty about the values and preferences typically lead to weak or conditional recommendations.[10]

Table 2.2 summarizes some examples showing how the guideline panels used the information when formulating recommendations. Panels were instructed to use the information provided about the relative importance of the main outcomes and balance of the desirable and undesirable consequences. Panelists also made judgments about the variability and uncertainty about the values and preferences information.

For example, for the antithrombotic guideline, the systematic review on utilities suggested that major bleeding was equivalent to nonfatal pulmonary embolism; while intracranial bleed overall was 2 to 3 times worse than major bleed or pulmonary embolism.[37] In the Breast Lump guidelines we found that recurrence and metastasis are the most important outcomes for women, and were considered as such by the panel.[36]

2.3.4. How consideration of local values and preferences influenced recommendations

The presumption that local values and preferences differ from those obtained in other settings, questions the usefulness of using the latter. In several cases, local values and preferences contributed significantly to the formulation of recommendations. For example, the allergic rhinitis management guideline stated since "there is important variability about how much people value its ([sublingual immunotherapy, SLIT]] effectiveness because there is a concern that some patients in Saudi Arabia would not accept SLIT with some allergens of animal origin". Consequently, the recommendation was a weak recommendation suggesting sublingual immunotherapy for treatment of adults with seasonal or intermittent allergic rhinitis based on moderate quality evidence.[36] Although the recommendation was not different from the source guideline,[38] one of the main reasons for this weak recommendation was the expression of local patient values and preferences described above.

The recommendation comparing ultrasonography versus mammography, as part of the triple assessment of palpable breast masses in women aged 30 - 40 years, was associated with very low certainty in the evidence of effects. The panel suggested "patients would likely favour the use of ultrasonography" because mammography can be more painful and uncomfortable for patients. In the panels' view this consideration of values and preferences justified a strong recommendation because ultrasonography showed better diagnostic accuracy

(sensitivity and specificity) compared with mammography despite very low certainty in the evidence.[36]

2.3.5. Workload related to values and preferences

Incorporating values and preferences in guideline development required resources on the following levels: literature searches, screening and synthesis, preparation of the GRADE EtD frameworks and consideration of values and preferences in decision-making. During development of the search strategy, we noted that many relevant studies were difficult to identify because of the lack of a validated filter or of standardized keywords (Medical Subject Headings: MeSH terms) being used to tag eligible studies. With the definition, measurement and methodology of values and preferences for guidelines still under debate, our aim to not miss relevant information was time and resource consuming. We managed this burden by limiting our search strategy through the stepwise use of a geographic search filter when required. For example, in the Migraine Headache guidelines, we first applied a geographic filter. After identifying no eligible studies, we felt it was necessary to spend additional time and resources to do a larger search for indirect evidence outside of the local context.

Panels recognized the importance of explicitly incorporating the information in the process and considered it in all of the 226 recommendations. The structured summary and presentation of the values and preferences information for each question in the GRADE EtD framework facilitated the process of considering this type of evidence.

2.4. Discussion

We describe an approach for the incorporation of the relative importance of health outcomes in healthcare recommendations. We applied a multi-faceted approach utilizing a systematic review strategy complemented by other information sources. We use illustrative examples to show the usefulness of identifying relevant studies and using their findings in drafting the recommendations.

2.4.1. Strengths and limitations

The systematic and transparent approach to identify and summarize published literature on values and preferences is the strength of the proposed strategy. The feedback from experienced panel members suggested that we did not miss important relevant studies. A second strength is our pre-conceived and structured approach to incorporate both published and elicited local values and preferences in the decision making process. Guideline developers can assume an international or national, or, alternatively, a localized or specific perspective. By considering the appropriate setting the recommendations could potentially be more acceptable to stakeholders. While the former strategy would be helpful for international organizations such as WHO, those adapting recommendations to a specific setting should consider locally relevant evidence, as was the case in this project.[1 39]

This study has some weaknesses. While the study is based on the development of over 20 guidelines and over 200 recommendations, it is restricted to one geographic setting. Also, limited local information was identified for patient values and preferences. The one related advantage is identifying the necessity of conducting more research on local values and preferences. Second, our definition

and eligibility criteria for values and preferences were broad. The inclusion of a variety of study designs resulted in challenges with determining the eligibility of individual studies and the category they belong to. The time and resources spent on systematic reviews of values and preferences varied across guideline topics. We also did not formally assess the certainty or the quality of the evidence in the values and preferences from published studies. As for information about values and preferences from panel members, the collected information was unsystematic, potentially biased, and sometimes difficult to use. Furthermore, we were not able to assess publication bias due to the nature of the study question, study design and the geographic filter we used. While we identified studies with a variety of designs providing relevant evidence, the lack of standardized methods for reporting and identifying the evidence places additional limitations on current guideline development but not on our work.

2.4.2. How to interpret and present information about values and preferences in guidelines

Although the integration of values and preferences is considered standard for trustworthy guideline development processes, using systematic reviews to identify values and preferences in a structured approach is still uncommon.[1 8 12 40 41] The Saudi Arabian panels weighted the relative importance of outcomes using information from literature reviews, the panel members themselves, and patient representatives. This facilitated adoption, adaptation and creating new recommendations according to local values. The GRADE EtD framework helped facilitate the use of values and preferences information in the decision making

process by explicitly calling attention to the criterion when balancing benefits and harms. The approach we used has face validity because the panel members did not identify missing studies on local values and preferences. As guideline methodology is refined, how to define, measure, and incorporate patient values and preferences will evolve.

There are other guideline efforts that consider patient values and preferences in the process of developing recommendations. For example, the National Institute for Health and Clinical Excellence (NICE) also considers the impact of values and preferences on the strength of recommendation. The process includes asking patient representatives to reveal their experience in addition to reviews of qualitative research evidence and cross-sectional surveys. However, NICE does not operationalize values and preferences as the importance of outcomes.[40]

Thus, despite recently increasing numbers of available primary studies and systematic reviews on values and preferences,[42-45] they are still rarely used in guidelines. This is likely also a result of poor guidance and definitions for how to incorporate this information appropriately. Our study provides a feasible approach to consider patient values and preferences in guideline development. However, other challenges in using this information remain. This includes accepted approaches to assessing the quality or certainty of evidence which is recognized by the GRADE working group and work is ongoing to develop an approach.[31 46-48] Furthermore, existing systematic reviews seldom have a clear definition, valid search strategy, or transparent synthesis methods to identify evidence about the relative importance of outcomes. Our experience of using GRADE EtD

frameworks, that do not yet routinely include modeling based on preferences, need to be seen in the context of other approaches that routinely include modeling.[10 49]

2.5. Conclusions

Although considering the relative importance of health outcomes is essential in informing healthcare decision-making, use of this type of information remains a complex area to integrate. Our experience shows that guidelines in general and GRADE EtD frameworks in particular, lend themselves to the incorporation of this aspect in clinical and public health recommendations. To further facilitate this process a methodologically rigorous and consistent approach for reporting, summarizing and interpreting the information is needed due to the great heterogeneity on the definition, perspective and measurement of values and preferences. We provide an empirical approach to address this concern through systematic reviews and panel members' input.

Abbreviations

DECIDE: Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence

EtD: Evidence-to-Decision

GIN: Guidelines International Network

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HUI: health utility index

MeSH: Medical Subject Headings

NICE: National Institute for Health and Clinical Excellence

QWB: quality of wellbeing

SLIT: sublingual immunotherapy

VAS: visual analogue scales

VTE: venous thromboembolism

WHO: World Health Organization

Declarations

Ethics approval and consent to participate

Not applicable. This study does not involve de novo patient data collection and describes the process of using preferences and values in guideline development. Information presented here is based on literature reviews and extracts from already published material in the GRADE EtD frameworks, that were approved by all panel members

(http://www.moh.gov.sa/endepts/Proofs/Pages/Guidelines.aspx). No patient informed consent and Institutional Review Board approval have been sought.

Consent for publication

Not applicable.

Availability of Data and Materials

The datasets supporting the conclusions of this article are included within the article and its additional file.

Author's contributions

YZ, PA, JB, HS designed the systematic review methodology for patient values and preferences and drafted the manuscript; HS and WW designed the methodology for Saudi Arabian healthcare guideline development. YZ, JB, WW, IE, EA, JM, WA, AC, RLM, RAM, JJR, JJY, CC, ZA, VM, MF, IN, RB, NS, BR, AD, MX, YA, CB, RW, and HS conducted the literature search, screening, and data abstraction, and consulted the panelists in the panel meetings. All authors read and approved the final manuscript.

Competing interests

Several authors are members of the GRADE working group and have helped developing the Evidence to Decision Frameworks.

Acknowledgements

We are grateful to McMaster University research librarians Jennifer Lawson and Tamara Navarro for their assistance in developing literature search strategies.

Funding

The Ministry of Health, Saudi Arabia and McMaster University, Hamilton, Canada.

The sponsor had no role in the design of the study or interpretation of the results except through the employed authors of this study.

Table 2.1. Eligibility criteria for the systematic review of Patient Values and Preferences

Category	Measurement
Utility/Health Status Value	Standard Gamble Time Trade Off Visual Analogue Scale Multi-attribute instruments (i.e. EQ-5D utility, HUI utility) Utility or health status values transformed (mapping) from quality of life measurements (both generic or disease specific tools)
Non-utility, quantitative information	Direct/Forced Choice exercise: choice from a set of options
	Non-utility measurement of health states: other self-developed questionnaires and scales
Qualitative information	Qualitative research

[a] Referring to transforming scores from quality of life measurement into a utility or health status value based on transformation equations

Table 2.2. Sources of information and how it was used by panels

Source of Information	What is the information?	How can it be used?
Update of prior systematic review	<u>*Utility estimate*</u> Nonfatal Intracranial Bleed (severe): 0.1 to 0.51 Nonfatal Intracranial Bleed (moderate): 0.29 to 0.77 Nonfatal Intracranial Bleed (mild): 0.47 to 0.94 Nonfatal Pulmonary Embolism: 0.63 Major Bleed: 0.44 to 0.84 *"This result suggested intracranial bleed overall was 2 to 3 times worse than major bleed or pulmonary embolism."*	To help guideline panelists weigh the benefits (absolute reduction in pulmonary embolism) and harms (absolute increase in bleedings).
Systematic review	<u>*Non-utility estimate*</u> For the guideline on management of breast lump and primary breast cancer, the systematic review identified one study reporting an additional year in life expectancy or 3% in survival rates were sufficient to make adjuvant chemotherapy worthwhile by 68–84% of women.	To judge to what extent women are willing to accept the burden of adjuvant chemotherapy to benefit from a specific amount of increased survival
Systematic review	<u>Qualitative finding</u> *"Evidence from qualitative studies suggested women may fear screening and may have a high level of anxiety related to colposcopy or treatment."*	To suggest what are the views of local women on cervical cancer screening tests in relation to its psychological impact
Panel members (either physicians or patients)	<u>*Panelists experience*</u> In some guideline topics, *patient inputs corroborated the panel's perception*.	To serve as complementary sources in addition to the information from systematic review.

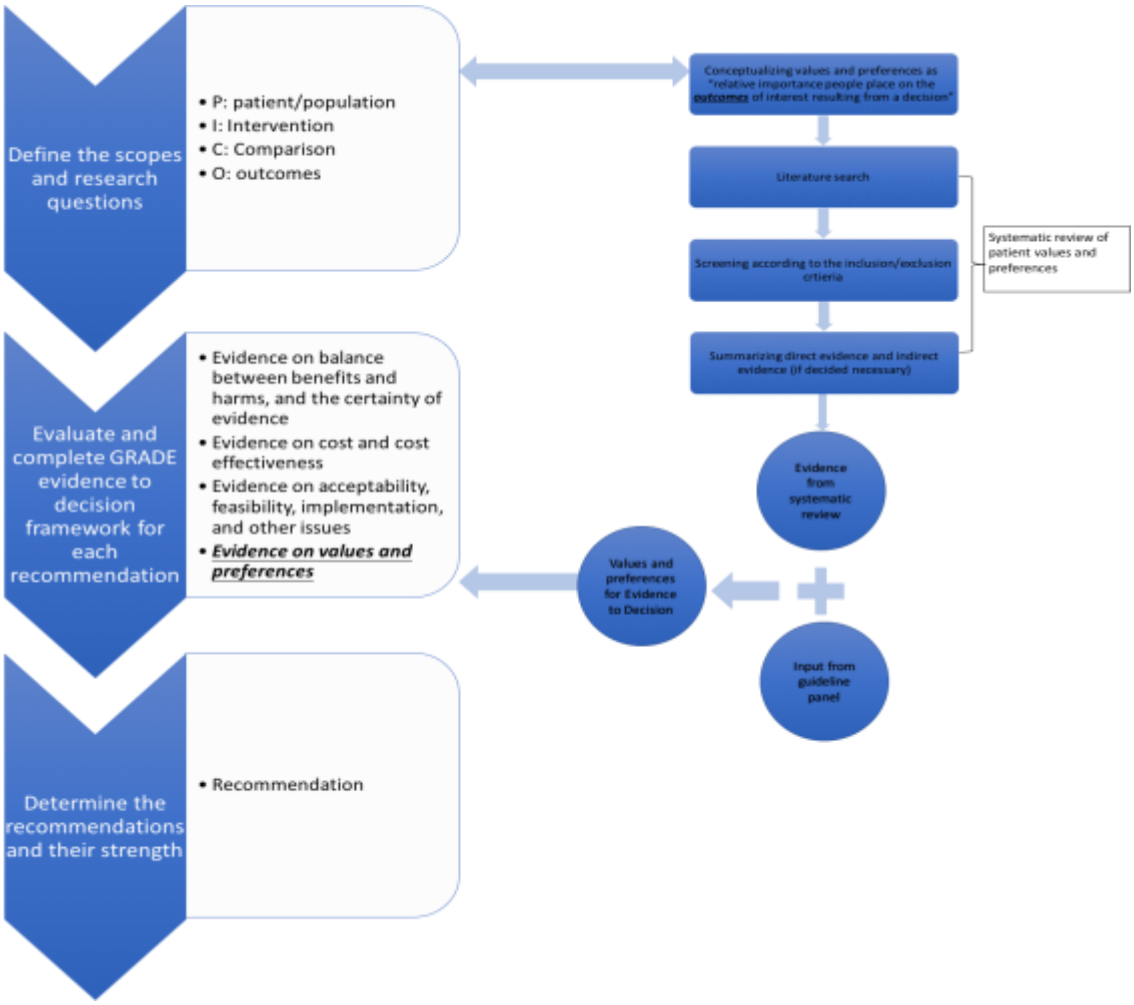


Figure 2.1. Process of Integrating Values and Preferences.

The steps on the left show the process of integrating values and preferences in guideline development. The guideline panel formulated the recommendations based on evidence on values and preferences, together with other evidence, e.g., evidence on the balance between benefits and harms and cost.

2.6. References

1. World Health Organization. WHO Handbook for Guideline Development. 2nd ed. Switzerland: World Health Organization, 2014.
2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical research ed)* 2008;336(7650):924-6. doi: 10.1136/bmj.39489.470347.AD [published Online First: 2008/04/26]
3. Schunemann HJ, Mustafa R, Brozek J, et al. Development of the GRADE Evidence to Decision (EtD) frameworks for tests in clinical practice and public health. *Journal of clinical epidemiology* 2016;76:89-98. doi: 10.1016/j.jclinepi.2016.01.032 [published Online First: 2016/03/05]
4. Schunemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014;186(3):E123-42. doi: 10.1503/cmaj.131237 [published Online First: 2013/12/18]
5. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical research ed)* 2016;353:i2016. doi: 10.1136/bmj.i2016
6. Alonso-Coello P, Oxman AD, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. *BMJ (Clinical research ed)* 2016;353:i2089. doi: 10.1136/bmj.i2089
7. Cluzeau F, Wedzicha JA, Kelson M, et al. Stakeholder involvement: how to do it right: article 9 in Integrating and coordinating efforts in COPD guideline development. An official ATS/ERS workshop report. *Proc Am Thorac Soc* 2012;9(5):269-73. doi: 10.1513/pats.201208-062ST [published Online First: 2012/12/21]
8. Kelson M, Akl EA, Bastian H, et al. Integrating Values and Consumer Involvement in Guidelines with the Patient at the Center: Article 8 in Integrating and Coordinating Efforts in COPD Guideline Development. An Official ATS/ERS Workshop Report. *Proc Am Thorac Soc* 2012;9(5):262-68. doi: 10.1513/pats.201208-061ST
9. Schunemann H, Fretheim A, Oxman A. Improving the use of research evidence in guideline development: 10. Integrating values and consumer involvement. *Health Research Policy and Systems* 2006;4(1):22.
10. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of clinical epidemiology* 2013;66(7):726-35. doi: 10.1016/j.jclinepi.2013.02.003 [published Online First: 2013/04/11]
11. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed)* 1996;312(7023):71-2. [published Online First: 1996/01/13]

12. van der Weijden T, Legare F, Boivin A, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. *Implementation science : IS* 2010;5:10. doi: 10.1186/1748-5908-5-10 [published Online First: 2010/03/09]
13. Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *JAMA : the journal of the American Medical Association* 2008;300(21):2483; author reply 83-4. doi: 10.1001/jama.2008.730 [published Online First: 2008/12/04]
14. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e1S-23S. doi: 10.1378/chest.11-2290 [published Online First: 2012/02/15]
15. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA : the journal of the American Medical Association* 2008;300(4):436-8. doi: 10.1001/jama.300.4.436 [published Online First: 2008/07/24]
16. Treweek S, Oxman AD, Alderson P, et al. Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence (DECIDE): protocol and preliminary results. *Implementation science : IS* 2013;8:6. doi: 10.1186/1748-5908-8-6 [published Online First: 2013/01/11]
17. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. *BMJ (Clinical research ed)* 2004;328(7454)
18. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 2011;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026 [published Online First: 2011/01/05]
19. Dolan P, Gudex C, Kind P, et al. Valuing health states: a comparison of methods. *Journal of health economics* 1996;15(2):209-31. [published Online First: 1996/03/08]
20. Gafni A, Birch S. Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Social science & medicine (1982)* 1995;40(6):767-76. [published Online First: 1995/03/01]
21. Torrance GW. Utility measurement in healthcare: the things I never got to. *PharmacoEconomics* 2006;24(11):1069-78. [published Online First: 2006/10/28]
22. Giacomini MK, Cook DJ, Streiner DL, et al. Using practice guidelines to allocate medical technologies. An ethics framework. *International journal of technology assessment in health care* 2000;16(4):987-1002. [published Online First: 2001/01/13]
23. Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. *International journal of technology assessment in health care* 2005;21(3):312-8. [published Online First: 2005/08/23]

24. Schunemann H, Mustafa R, Brozek J, et al. Saudi Arabian Handbook for Healthcare Guideline Development. http://wwwmohgovsa/endepts/Proofs/Pages/GuidelineAdaptationaspx *Accessed on Nov 11th, 2015* 2015
25. Schünemann HJ, Wiercioch W, Brozek J, et al. GRADE Evidence to Decision (EtD) frameworks for adoption, adaptation, and de novo development of trustworthy recommendations: GRADE-ADOLOPMENT. *Journal of clinical epidemiology* 2016;Article in Progress(81):101-10. doi: http://dx.doi.org/10.1016/j.jclinepi.2016.09.009
26. Garg AX, Hackam D, Tonelli M. Systematic review and meta-analysis: when one study is just not enough. *Clinical journal of the American Society of Nephrology : CJASN* 2008;3(1):253-60. doi: 10.2215/cjn.01430307 [published Online First: 2008/01/08]
27. Higgins JPT, Green S, (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. : The Cochrane Collaboration, 2011. Available from http://handbook.cochrane.org/.
28. Longworth L, Yang Y, Young T, et al. Use of generic and condition-specific measures of health-related quality of life in NICE decision-making: A systematic review, statistical modelling and survey. *Health Technology Assessment* 2014;18(9):1-224. doi: http://dx.doi.org/10.3310/hta18090
29. Furlong WJ, Feeny DH, Torrance GW, et al. The Health Utilities Index (HUI) system for assessing health-related quality of life in clinical studies. *Annals of medicine* 2001;33(5):375-84. [published Online First: 2001/08/09]
30. Starkie HJ, Briggs AH, Chambers MG, et al. Predicting EQ-5D values using the SGRQ. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2011;14(2):354-60. doi: 10.1016/j.jval.2010.09.011 [published Online First: 2011/03/16]
31. DeJean D, Giacomini M, Vanstone M, et al. Patient experiences of depression and anxiety with chronic disease: a systematic review and qualitative meta-synthesis. *Ontario health technology assessment series* 2013;13(16):1-33. [published Online First: 2013/11/15]
32. Torrance GW. Preferences for health states: a review of measurement methods. *Mead Johnson Symposium on Perinatal and Developmental Medicine* 1982(20):37-45. [published Online First: 1982/01/01]
33. Longworth L, Rowen D. NICE DSU Technical Support Document 10: The use of mapping methods to estimate health state utility values. 2011. Available from http://www.nicedsu.org.uk/.
34. Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health services research* 1994;29(2):207-24. [published Online First: 1994/06/01]
35. Selva A SI, Y Z, AJ S, et al. Development of a search strategy for studies about patients' values and preferences (submitted for publication).
36. The Saudi Center for Evidence Based Healthcare (EBHC): Clinical Practice Guidelines 2015 [Available from: Available from: http://www.moh.gov.sa/endepts/Proofs/Pages/Guidelines.aspx.

37. Al-Hameed F, Al-Dorzi HM, Shamy A, et al. The Saudi clinical practice guideline for the diagnosis of the first deep venous thrombosis of the lower extremity. *Annals of thoracic medicine* 2015;10(1):3-15. doi: 10.4103/1817-1737.146849 [published Online First: 2015/01/17]
38. Brozek JL, Bousquet J, Baena-Cagnani CE, et al. Allergic Rhinitis and its Impact on Asthma (ARIA) guidelines: 2010 revision. *The Journal of allergy and clinical immunology* 2010;126(3):466-76. doi: 10.1016/j.jaci.2010.06.047 [published Online First: 2010/09/08]
39. World Health Organization. Estonian Handbook for Guidelines Development. Switzerland World Health Organization, 2011.
40. National Institute for Health and Clinical Excellence. The guidelines manual (Nov 2012). London: National Institute for Health and Clinical Excellence. Available from: http://www.nice.org.uk/.
41. Graham R, Mancher M, Miller Wolman D, et al. Clinical Practice Guidelines We Can Trust. Washington DC: the National Academy of Sciences 2011.
42. Lin OS, Kozarek RA, Gluck M, et al. Preference for colonoscopy versus computerized tomographic colonography: A systematic review and meta-analysis of observational studies. *Journal of General Internal Medicine* 2012;27(10):1349-60. doi: http://dx.doi.org/10.1007/s11606-012-2115-4
43. Mohiuddin S, Payne K. Utility values for adults with unipolar depression: Systematic review and meta-analysis. *Medical Decision Making* 2014;34(5):666-85. doi: http://dx.doi.org/10.1177/0272989X14524990
44. Peeters Y, Stiggelbout AM. Health state valuations of patients and the general public analytically compared: A meta-analytical comparison of patient and population health state utilities. *Value in Health* 2010;13(2):306-09. doi: http://dx.doi.org/10.1111/j.1524-4733.2009.00610.x
45. Sadique MZ, Legood R. Women's preferences regarding options for management of atypical, borderline or low-grade cervical cytological abnormalities: A review of the evidence. *Cytopathology* 2012;23(3):161-66. doi: http://dx.doi.org/10.1111/j.1365-2303.2011.00873.x
46. Jackson LJ, Auguste P, Low N, et al. Valuing the health states associated with chlamydia trachomatis infections and their sequelae: A systematic review of economic evaluations and primary studies. *Value in Health* 2014;17(1):116-30. doi: http://dx.doi.org/10.1016/j.jval.2013.10.005
47. Bremner KE, Chong CA, Tomlinson G, et al. A Review and meta-analysis of prostate cancer utilities. *Medical Decision Making* 2007;27(3):288-98. doi: http://dx.doi.org/10.1177/0272989X07300604
48. Cronin M, Meaney S, Jepson NJ, et al. A qualitative study of trends in patient preferences for the management of the partially dentate state. *Gerodontology* 2009;26(2):137-42. doi: 10.1111/j.1741-2358.2008.00239.x [published Online First: 2009/06/06]
49. Nutt DJ, King LA, Phillips LD. Drug harms in the UK: a multicriteria decision analysis. *Lancet (London, England)* 2010;376(9752):1558-65. doi: 10.1016/s0140-6736(10)61462-6 [published Online First: 2010/11/03]

Chapter 3. Development of GRADE guidance for assessing the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 1. Risk of bias and indirectness

Yuan Zhang[1], Pablo Alonso Coello[1,2], Gordon Guyatt[1], Juan Jose Yepes-Nunez[1], Elie A. Akl[1,3], Glen Hazlewood[4], Hector Pardo-Hernandez[2], Itziar Etxeandia-Ikobaltzeta[1], Amir Qaseem[5], John W. Williams Jr. [6], Peter Tugwell[7], Yaping Chang[1], Yuqing Zhang[1], Reem A. Mustafa [1, 8] , Holger Schünemann[1, 9]
1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada
2. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
3. Department of Internal Medicine, Faculty of Medicine, American University of Beirut, Lebanon
4. Department of Medicine and Department of Community Health Sciences, University of Calgary, Canada
5. American College of Physicians, Philadelphia, Pennsylvania, USA
6. Center of Innovation for Health Services Research in Primary Care at the Durham Veterans Affairs Medical Center and Duke University, Durham, NC 27701, USA.
7. Department of Medicine, University of Ottawa, Canada
8. Department of Internal Medicine, Division of Nephrology and Hypertension, University of Kansas Medical Center, Kansas City, Kansas, USA
9. Department of Medicine, McMaster University, Canada

Corresponding author:
Holger J Schünemann, MD, PhD
Chair, Department of Health Research Methods, Evidence, and Impact
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON, L8N 4K1, Canada
Email: schuneh@mcmaster.ca
Tel: +1 905 525 9140 x 24931

Although this manuscript is final for the doctoral thesis of Yuan Zhang, this is a GRADE working group project, and the final manuscript will require approval by the GRADE Working Group. Members of the GRADE working group making appropriate contributions to this manuscript at that stage may become authors on the byline.

Abstract

The GRADE working group defines patient values and preferences as the relative importance patients place on the main health outcomes. Although the GRADE working group has developed guidance for treatment, diagnosis, resource and prognosis questions, similar guidance has been lacking for the relative importance of outcomes of alternative management strategies.

We applied the GRADE domains to rate the certainty of evidence in the importance of outcomes to several systematic reviews, conducted consensus meetings and consulted stakeholders in the GRADE working group for feedback. This is the first of two articles that provide guidance on how users can assess the certainty of relative importance of outcome evidence. A body of evidence addressing the importance of outcomes starts at "high certainty"; concerns with risk of bias, indirectness, inconsistency, imprecision and publication bias lead to rating this evidence down to moderate, low or very low certainty in the evidence. Given the lack of accepted risk of bias assessment tools for this type of studies, we propose subdomains of risk of bias as selection of the study population, missing data, the type of measurement instrument, and confounding; we have developed items for each subdomain. The population, intervention, comparison and outcome (PICO) elements associated with the evidence determine the degree of indirectness. For rating population indirectness, we suggest a gradient of optimal populations from which to elicit the relative importance of outcome.

In conclusion, this article provides guidance and examples for rating risk of bias and indirectness for a body of evidence summarizing the importance of outcomes.

3.1. Introduction

The development of appropriate methods for assessing evidence regarding intervention effects has increased the credibility of health care recommendations. Decisions in health care, however, require not only knowledge of the effects of interventions (e.g. the absolute risk reduction or increase for an outcome in a particular population resulting from a specific intervention when compared to an alternative) but also knowledge of the relative importance of the outcomes that interventions prevent or cause. The balance of desirable and undesirable effects is a major factor in determining the preference for alternative management, screening or diagnostic options. This balance is determined not only by the absolute risk differences for the outcomes of interest, but also by the relative importance of those outcomes (see Box 3.1 for a hypothetical example).

Box 3.1. A hypothetical example for considering the importance of outcomes

The evidence comparing a new intervention to standard care shows an absolute risk reduction of 10 per 1000 for harmful outcome 'A', and an absolute risk increase of 10 per 1000 for harmful outcome 'B'. If outcomes A and B are judged as equally important (e.g., thrombosis and bleeding respectively), then the balance of benefits and harms does not favor or disfavor the new intervention. If outcome A is judged as relatively more important than outcome B (e.g., mortality and bleeding respectively), then the balance of benefits and harms is in favor of the new intervention.

Incorporating these considerations in health care decision-making has gained attention in the evidence-based medicine (EBM) community, often under the umbrella of *values and preferences*. [1-6] In the context of decision-making, values and preferences can be conceptualized as the relative importance people place on the outcomes of interest resulting from a decision (e.g. about accepting a treatment or undergoing a test).[2]

In individual physician-patient encounters, consideration of the importance patients place on outcomes is essential for shared decision-making; in the context of developing guideline recommendations, values and preferences represent the typical relative importance that those affected by the recommendations place on the outcomes of interest.[1 3 4 7-9] Knowledge of that relative importance allows guideline panel members to balance the anticipated desirable and undesirable health outcomes.

The methods that investigators use to ascertain the relative importance of outcomes (RIO) include: a) directly measurement of the utility or value of outcomes, e.g. with the standard gamble,[10-12] time trade off,[13 14] or rating scales; [11 15 16] b) indirect measurement of utility: results from instruments such as the EQ-5D utility, or SF-6D utility, which would transform the measurement results across several domains, i.e., pain, mobility, into the utility; [17 18] c) conjoint analysis including discrete choice experiments,[19 20] contingent valuation and willingness to pay,[21] probability trade off,[22 23] paired comparison, or d) other quantitative surveys and questionnaires.[24 25] In addition, qualitative studies can provide evidence about the relative importance of outcomes.[26 27]

Given healthcare decisions will be influenced by both the health effects of interventions, and the relative importance of outcomes of interest, they both require appropriate methods of certainty assessment. The GRADE working group has developed approaches to assess certainty of evidence addressing intervention effects, test accuracy, resources and prognosis.[28] There is now a need to develop a transparent and structured approach to assessing the certainty of evidence for the relative importance of outcomes.

The aim of these two articles is to provide guidance on the GRADE approach for assessing the certainty of a body of evidence dealing with the relative importance of outcomes. The first article of this series focuses on the definitions and methods of this project, and the GRADE approach for the domains risk of bias and indirectness. The second article will focus on the GRADE approach for the domains of inconsistency, imprecision and publication bias and rating up the certainty of evidence. The second article will also clarify what variability of values and preferences or the relative importance of outcomes means in this context. We will illustrate the rationale of the GRADE considerations and provide examples for these considerations beginning with a description of our terminology (See Box 3.2). [2 29]

Box 3.2. Terminology

Terminology	Scope or definition
Outcome	The term outcome includes "***health state***" and non-health states that are relevant to the alternative treatment under consideration. This includes a broad set of the outcomes directly and indirectly related to health or a disease, an intervention, or non-health consequences. Outcomes can be more or less health-related. For example, from mostly health related to least, patients

	will have their views regard on the importance of the following outcomes: breathlessness, treatment burden of warfarin or insulin injection, ease of reaching a clinic to undergo blood tests and other monitoring.
Relative importance of outcomes	The ***relative importance of outcomes*** is interchangeably used with ***values and preferences***, ***outcome importance***, or ***outcome valuation***: GRADE defines values and preferences as the relative importance of outcomes.[2]
Instrument (for determining the relative importance of outcomes)	This term, when used referring to measuring relative importance of outcomes, refers to "***measurement tool***" "***measurement methods***" or "***measurement instruments***".
Certainty of evidence	This term is interchangeably used with "***quality of evidence***", ***"strength of evidence",*** and "***confidence in estimate***". ***Certainty of evidence*** has different meanings for systematic reviews and guideline development. For systematic reviews the definition is: The extent of our confidence that the relative importance of the outcomes (and variability) lie in a particular range; for guidelines the definition is: The extent of our confidence that the estimate of the relative importance of the outcomes (and variability) are adequate to support a particular recommendation.[29]

3.2. Methods

We applied a multi-pronged approach to develop this GRADE guidance: a) we summarized information from a database of systematic reviews addressing this topic to explore if the current GRADE domains sufficiently cover aspects of certainty related to the relative importance of outcomes, and developed the draft GRADE approach with specific consideration regarding this type of evidence; b) we applied the results of the prior steps to ten systematic review examples assessing the certainty of evidence; c) we modified initial guidance based on the examples assessed and then presented this work for final guidance development to the GRADE project group on relative importance of outcome through

teleconferences, online video meetings and in-person meetings and through electronic documents to the entire GRADE working group. This work focuses here on quantitative estimates of relative importance of outcomes. For qualitative evidence, we refer readers to the work of the GRADE-CERQUAL project group. [30]

The work began in 2012 with developing a dissertation project for the first author. We subsequently conducted the work with members of the GRADE project group on relative importance of outcomes, co-supervised by the co-chairs of the GRADE working group and another member of the GRADE working group with expertise in values and preferences and shared decision-making studies. The ethics board at McMaster University approved this research.

3.2.1. Summarizing certainty domains and methods for assessing the certainty of evidence and developing the GRADE approach

Based on a previous systematic survey project, [31] we identified systematic reviews addressing relative importance outcomes and qualitatively summarized existing methods utilized to assess the certainty of a body of evidence and other potential quality indicators, i.e. all factors perceived to influence certainty. After discussion, we constructed a list of possible factors and then attempted to match them to the existing GRADE domains. We considered the existing GRADE domains of risk of bias, inconsistency, indirectness, imprecision and publication bias for downgrading; [29] and large effect sizes, the existence of a dose-response gradient or if residual plausible confounding bias would increase our certainty for

upgrading[32]. We planned to record any additional domains that did not fit into existing GRADE domains.

3.2.2. Application of GRADE approach to examples

We selected a sample of 10 systematic reviews from a previous project using a maximum variance sampling strategy ensuring that the selection would allow us to illustrate all GRADE domains and address a diversity of health conditions.[31] We independently assessed the certainty of evidence in pairs by rating risk of bias, inconsistency, imprecision, indirectness and publication bias. We recorded decisions supporting downgrading and resolved disagreements through discussion or feedback from senior GRADE members. We developed GRADE evidence profiles to facilitate the work.[33] For the assessment of risk of bias, we developed signallingsignaling questions and drafted guidance utilizing the approach GRADE took for its prior guidance, e.g. guidance on prognostic evidence. [34]

3.2.3. Consulting for feedback

To ensure a broad perspective, we provided the examples to a group of individuals including guideline developers, systematic review and health technology assessment authors, clinical epidemiologists, biostatisticians, psychologists and social scientists, clinicians, and researchers with experience in relative importance of outcome assessment from Canada, the US and Europe. We collected feedback from this group through six rounds of online meetings, complemented by emails and in-person meetings and telephone calls. We documented the adjustments made and circulated the records for review and comments as part of a GRADE project group.

Following each round of feedback, we iteratively improved the preliminary GRADE guidance for assessing the certainty of evidence in the relative importance of outcomes, and illustrated the rationale with examples. After the work and guidance had been presented and discussed with members of the GRADE working group at two of their regular meetings, we finalized the guidance. This article was then approved by the GRADE working group and, finally, the GRADE guidance group as official guidance.

3.3. Guidance for GRADE domains

We did not identify additional domains beyond what the GRADE working group had suggested previously: risk of bias, inconsistency, indirectness, imprecision, publication bias, and domains to rate up the evidence.[29] For each of the ten systematic review examples we produced an evidence profile accordingly (see table 3.1 for an example). Here we focus on the detailed guidance for the GRADE domains risk of bias and indirectness.

3.4. Risk of bias or limitations in the detailed study design or execution

Risk of bias may be a concern at different stages of an investigation into relative importance of outcomes, including study design, execution, data analysis and reporting. [35] Assessing risk of bias for the relative importance of outcomes is similar to assessment of risk of bias for intervention effects in that it requires assessment of individual studies, but differs in several important ways. First,

unlike studies on treatment effect, there is no accepted or commonly used tool for its assessment.[35] Second, given that the relative importance of an outcome is an estimate that does not represent an effect but is conceptually closer to an estimate of test accuracy or baseline risk, certainty of evidence from non-randomized studies begins as high certainty.[34 36]

We developed the following subdomains and, for each subdomain, signalling questions, for assessing the risk of bias domain (Table 3.2):

1. Risk of bias due to selection of participants into the study: To what extent does the enrolled study population reflect the intended sample? Improper sample selection will lead to biased estimates of relative importance of outcome if differing characteristics are associated peoples' relative importance of outcome.
2. Risk of bias due to missing data: was the attrition sufficiently low to minimize the risk of bias? High attrition, or low response rate for cross-sectional studies may result in participating individuals who differ systematically in their relative importance of outcomes from those who do not participate.[37 38]
3. Risk of bias due to the measurement instrument: Is the instrument chosen to elicit the relative importance of outcomes determine valid? This subdomain includes four items: choice of the instrument, administration of the instrument, outcome presentation, and understanding of the instrument by the study population.

4. Risk of bias due to confounding: Does inappropriate data analysis lead to biased results? Was adjustment, stratification in the analysis and model selection, if any, appropriate to avoid distorted results from confounding?

With the above signalling questions, for each subdomain each study, depending on the likelihood of bias and the magnitude of its impact on the estimates, would be classified as low, moderate, serious, and critical risk of bias (see Box 3.3 and Box 3.4). Across a body of evidence, an assessment of risk of bias should focus on the risk of bias domains across studies: that assessment (each subdomain across studies) would be labeled not serious, serious, and very serious. The decision to rate down for risk of bias would then require looking at the overall pattern of results across domains and across studies. A classification of risk of bias of individual studies (across these subdomains within a study) may be helpful for describing individual studies but is not the determining factor for an assessment across studies - that is, the body of evidence. We encourage raters to attempt making a judgment based on the information available (either in the study report or after obtaining additional information from authors), and including inferences about what is not stated, but is most likely.

Box 3.3. Judgment of risk of bias for risk of bias subdomains

Response option	Criteria
Low risk of bias	The estimate in this relative importance of outcome study is unlikely to be biased with regard to this subdomain.
Moderate risk of bias	The estimate in this relative importance of outcome study is likely to be biased with regard to this subdomain but the influence of the bias is limited.
Serious risk of bias	The estimate in this relative importance of outcome study is probably biased with regard to this subdomain

	and the influence of the bias is substantial.
Critical risk of bias	The estimate in this relative importance of outcome study is certain to be distorted with regard to this subdomain and the estimate is not trustworthy.

We now provide additional guidance for rating for the risk of bias subdomains.

3.4.1. Bias due to selection of participants into the study

Considering risk of bias related to study population, the users should ask the following questions: ***Was an appropriate study sample selected from the sampling frame? (Answer options: yes; probably yes; probably no; no)***

When, as in this situation, there is only one signalling question for a domain, a study will be classified, depending on the likelihood and magnitude of impact, as low or moderate if the response to the signalling question is yes, or probably yes, and high and critical if the response is no, or probably no.

The study population selection is a critical component since it will influence the results through the population the researchers study. When answering this question, users should consider the study's sampling strategy, in particular whether only a subset of the target population were likely selected, and if so whether that subset would lead to biased estimates compared to the entire target population.

There is an inevitable grey area between classifying limitations in selection of the population as an issue of bias or an issue of directness (also known as representativeness, generalizability, external validity, applicability). Here we are addressing selection bias, not whether the result apply to the target population of a recommendation or other health care question, though whether recruitment of a

subpopulation is best classified as selection bias or an applicability issue may be a matter of judgment.

There is no single standard for classifying a study as low, moderate, high or critical risk of bias in the sample selection subdomain. As an example of the judgment for a cross-sectional study, one might consider that a stratified random sampling strategy would minimize the risk of selecting a study population that is not representative of the sampling frame, while a convenience sample might probably be a biased sample for the study population.

Example: Lenert et al. reported the importance of outcomes related to treatment of deep vein thrombosis using a multimedia program.[39] The researchers used a convenience sample: they recruited 30 healthy women "from the communities surrounding our institution by placing flyers in shopping malls and other public areas". Convenience sampling of female only participants could bias the results if there were gender differences and the researchers intended to study both genders (though, with respect to this issue one would get an unbiased sample of female views of the matter, which one might apply to female patients). Unequivocally related to bias, however, is using flyers: the women who choose to enroll are likely to have different preferences than those who decline, and there is no way of knowing to whom exactly the evidence obtained applies. Thus, the answer to the question "was an appropriate study sample selected from the sampling frame" is "no", and our judgment of magnitude of the limitation led us to classify the magnitude of bias as critical.

3.4.2. Bias due to missing data

To consider risk of bias related to missing participant data, the users should ask the following question: ***Was the attrition sufficiently low to minimize the risk of bias? (Answer options: yes; probably yes; probably no; no)***

To answer the question, users need to consider the response rate; if follow-up was involved, the attrition rate; and the characteristics of the participants who responded and those who did not. Thus, this subdomain of missing data includes both the response rate of the study population approached and attrition rate during the follow-up process.

High response rates are clearly preferable, and a high proportion of nonresponse could be problematic. For studies with follow-up planned and completed, the attrition rate is another concern: participants may be lost to follow up. Participants providing responses may very plausibly differ from those who do not; to the extent this is the case, results coming only from those followed may be misleading.

We do not suggest a single cut-off for an "inadequate" response rate. While the judgment about the response rate is subjective, users should report transparently the reason for their risk of bias assessment.

Example: To conduct a decision analysis, investigators invited 180 people meeting study eligibility criteria to derive utility measures for health states. Only 64 of the invitees agreed, of whom 57 completed the study. The low response rate is likely to bias the estimates of utilities and impact the credibility of this and future decision analysis based on these utilities.[40] The answer to the question

"Was the attrition sufficiently low to minimize the risk of bias" was no. We classified the study at serious risk of bias in the missing data subdomain.

3.4.3. Bias due to the measurement instrument

To consider the risk of bias related to measurement instrument, users should ask the following questions:

1. ***Is the chosen the instrument for eliciting relative importance of outcomes valid and reliable? (Answer options: Yes; probably yes; probably no; no)***

Issues relevant to this question include both the reliability and the validity of the instrument. Low reliability or validity can result from intrinsic limitations of the measurement instrument or administration error. Authors may provide information regarding the measurement properties of the instrument they have chosen. Alternatively, they may have chosen an instrument with which assessors are familiar and with widely accepted reliability and validity.

A tentative list of generic instruments with accepted validity and reliability include: standard gamble, time trade off, visual analogue scale (or feeling thermometers), discrete choice, treatment trade-off, and willingness to pay. Use of these instruments does not, however, guarantee that they have been administrated appropriately.

If the authors have neither used an instrument with widely accepted satisfactory measurement properties, or have provided information regarding satisfactory reliability and validity, the risk of bias is likely to be substantial.

Example: Polonsky et al. reported a study examining patient preferences regarding a once-weekly glucose-lowering medication. Patients responded on a Likert-type scale describing their preferences from 1 (strongly disagree) to 5 (strongly agree).[41] Given absence of demonstrated validity and reliability for this preference elicitation, our response to the signalling question was "probably not" and we classified this study as serious risk of bias for the subdomain of measurement instrument. We did not rate it critical risk of bias because five point Likert type scales have proved valid in other contexts.

2. ***Was the instrument administered in the intended way? (Answer options: Yes; probably yes; probably no; no)***

The previous subdomain examines the shortcoming of the instrument employed; this subdomain considers how the instrument was actually administered. For a specific study, the researchers should demonstrate that the instrument has been administered correctly, or in a manner conforming to their rationale to minimize the risk of introducing bias. In addition, the measurement instruments should be administered in a consistent manner across different subpopulations.

Example: Empirically, systematic reviews suggested that the way researchers ask the time trade off questions to elicit preferences may influence the results; this is also true for the standard gamble.[17 42 43] In a study assessing the utility of people with traumatic spinal cord injuries, the researchers used a telephone interview strategy to administer the standard gamble. In this study, the participants were asked to dedicate 30-49 minutes to the telephone interview. During the interview, unlike the usual case for standard gamble in which the instrument administration

includes a visual prompt, the participants were asked to imagine "that they would live in their current health states for an average life expectancy of 25 years". Additionally, the alternate probabilities in the standard gamble process were only verbally described.[44] For this study, the answer to the signalling question is "no"; we classified this as "serious risk of bias" because the measurement instrument was not administered in the intended way.

3. ***Was a valid representation of the outcome (health state) utilized? (Answer options: yes; probably yes; probably no; no)***

The description of outcomes is another possible source of bias. Optimal representation of the outcome includes of a detailed explanation of how the outcome that defines the experience, probability, duration and consequences was developed. Pragmatically, we suggest users classify a study as serious or critical risk of bias only if they have serious doubt regarding the appropriateness of the outcome presentation. This question only applies when the participants are asked to indicate the importance they would like to place on a set of hypothetical or described outcomes, rather than their own health.

Example: In the study examining patient preferences regarding a once-weekly glucose-lowering medication option, the researchers presented seven potential outcome characteristics to participants, with five positive and two negative outcomes. The researchers did not report their reason for selection these outcomes. Moreover, the descriptions of outcomes lacked detail; for example, "once-a-week medication could improve my quality of life."[41] Thus, the answer to the question "Was a valid representation of the outcome (health state) utilized" is "no". The

vague descriptions likely led to varied understanding of the outcome leading us to classify the subdomain as serious risk of bias.

4. Did the researchers check the understanding of the instrument?

Answer options include:

- The investigator tested the understanding and understanding was adequate;
- The investigators did not formally test the understanding, but there was evidence suggesting adequate understanding;
- The investigator tested the understanding and the understanding was inadequate;
- The investigators did not formally test the understanding and there was evidence suggesting inadequate understanding.

In choosing among these options, reviewers should consider the following: Was the instrument simple enough to assume understanding? Did the researchers pilot the instrument? Did the researchers formally test the understanding and did the results suggest understanding of the tasks?

If participants did not understand the task, it is not possible to obtain accurate results. There is likely to be, however, a gradient in understanding. Evaluating the risk of bias on this subdomain requires checking whether the study authors have provided evidence of adequate understanding. Checking understanding of participants is, however, neither common practice in the execution nor in the reporting of a study. Fortunately, reviewers may be able to deduce that understanding was adequate if the instrument applied was simple enough, or if the authors describe successful piloting of the instrument.

Example: Gage et al. reported decision and cost-effectiveness analyses comparing warfarin versus aspirin for prophylaxis of stroke. The researchers used the time trade-off technique to elicit the utility of outcomes. Of 69 participants who completed the study, 57 reportedly understood the technique.[45] Nearly 20% (12 of 69, 17.4%) of the participants had difficulties with understanding the instrument providing the rationale to classify this study as serious risk of bias in subdomain of measurement instrument.

3.4.4. Bias due to confounding

Users should ask: ***Were the results analyzed appropriately to avoid influence of bias and confounding? (Answer options: yes; probably yes; probably no; no)***

To answer this question, users also need to consider whether the adjustment, stratification, or model selection was appropriate. Studies addressing the importance of outcomes should present results adjusted for important co-variates. For example, if the importance of an outcome is associated with prior experiences and this can be appropriately controlled for, reporting adjusted results is likely to be informative.

For some methodologies, such as discrete choice experiments, researchers need to select appropriate models and adjust potential characteristics that could distort the results and conclusion. Raters need to take data analysis into the risk of bias consideration, with scrutiny on the adjustment, stratifications, model selections and interactions. This domain may not be applicable to all primary studies because not all studies will require controlled data analysis.

Example: In a discrete choice exercise study, the researchers invited 489 screening-naive and 496 screened individuals to determine the preferences of various screening tests and to predict uptake for colorectal cancer screening programs. The researchers applied a multi-nominal logit model to analyze the data. Although they conducted sensitivity analyses to include irrational responses, they did not conduct other sensitivity analyses. And although they reported respondent characteristics, they did not examine the interaction between choice and respondent characteristics and only compared the differences among subgroups with Chi-square and Student's t-tests. Thus, our answer to the signalling question was probably not, and we classified this study as serious risk of bias for data analysis.

3.4.5. Summary of risk of bias

Consistent with the GRADE approach for other types of evidence, the risk of bias assessment is conducted for each outcome. Raters should summarize risk of bias for an outcome across studies, first by subdomain and then, after possible sensitivity analysis that evaluates whether or not risk of bias in individual studies is likely to influence the overall results, across subdomains and studies.
Users need to make an overall judgment regarding the relative weight or contribution of studies classified as low, moderate, high or critical in the subdomains and items we listed above. If most information is from studies at low risk of bias for all subdomains, the overall judgment of risk of bias should be "low risk of bias" and in the GRADE certainty of evidence, raters would not downgrade. However, as the contribution of studies with risk of bias concerns

(studies classified as "moderate" "serious" or even "critical" risk of bias) to the body of evidence increases (see Box 3.4), and accordingly, raters downgrade the certainty of evidence by one or more levels due to risk of bias.[46] Risk of bias assessment on any domain is a continuum, and reviewers must bear this in mind when making their overall judgments.

Box 3.4. Overall risk of bias for a study

Response option	Criteria
Low risk of bias	The study is classified as with low risk of bias across subdomains.
Moderate risk of bias	The study is classified as low or moderate risk of bias across subdomains.
Serious risk of bias	The study is classified as serious risk of bias for at least one subdomain, but not classified as critical risk of bias for any subdomain.
Critical risk of bias	The study is classified as critical risk of bias for at least one subdomain.

Example: One systematic review summarized the utility of severe non-fatal strokes. [4] Two of the seven included studies reported a low response rate, and 17% of participants in a third study had difficulties to understand the instrument; these three studies contributed approximately 35% of all participants who provided information for the estimates. However, because no other concern was raised for other risk of bias subdomains, and the results from studies with risk of bias concerns were similar to those at low risk of bias, we did not downgrade for risk of bias.[4] In a review to assess the patient preferences for type 2 diabetes treatment related outcomes, of all 61 included studies, only six showed that the respondents were similar to non-respondents. [47] Thus, we downgraded the certainty of evidence due to risk of bias resulting from selection of participants into the studies.

3.5. Indirectness

For evidence of treatment effects, evidence can be indirect because of the differences in the population, interventions of interest, outcomes of interest, and indirect comparisons. Similarly, indirectness can be a reason to rate down the certainty of evidence of the relative importance of outcomes.[48] The assessment of indirectness for relative importance of outcomes has its specific features. First, studies usually do not directly compare the intervention options; rather, the focus is on outcomes. Secondly, surrogate outcomes or outcomes that are not patient-important would be a source of indirectness for treatment questions - this may not be the case for the evidence of relative importance of outcomes. In importance studies, the outcomes are indirect only because the outcomes are not representative. Thus, if we are interested in the importance of a surrogate outcome from the patients' perspective, being a "surrogate" does not justify rating down the certainty of evidence. Additionally, there is no indirect comparison in the relative importance of outcomes evidence. Lastly, the methods to elicit the relative importance of outcomes could be a source of indirectness. Here we provide the rationale and examples of these considerations, which we organize into two categories: indirectness due to PICO elements, and indirectness due to methodological elements (Table 3.3).

3.5.1. PICO elements

For a guideline development project, the research question will be defined following the PICO format (P: population, I: intervention; C: comparison; O: outcome). For a systematic review addressing the relative importance of outcomes,

we could define the research question as "what is the relative importance that patients place on the outcomes when they make a decision related to…", for which we still need a clearly defined PICO elements. PICO elements could be sources of indirectness, when the PICO elements in the body of evidence do not represent the PICO elements of interest. To consider PICO elements, users should ask the following signaling questions:

1. ***Is the population studied matching the population of interest?***

   (Answer options: yes; probably yes; probably no; no.)

The certainty of evidence will be lower if the evidence is based on populations differing from those who would face the choice of interest. Ideally, the population would be newly diagnosed patients facing the same choice. But the optimal population should be a case-by-case judgment. Patients newly diagnosed but who have already made the decision of interest may not be appropriate: cognitive dissonance may influence their answer, with an inclination to report relative importance of outcomes consistent with their prior decision. However, it is also argued that the patients who have already made the decision could weigh the outcomes thoughtfully without being overwhelmed by the diagnosis and in a rush to make decisions.

More indirect would be people who are at high risk of the condition of interest and who therefore may face the choice in the near future, but for whom – for the time being – the decision remains hypothetical. Indirectness increases further if a study enrolls proxies (e.g., spouse, other family members or caregivers) to provide indirect evidence of the relative importance of outcomes of the target population.

Similarly, differences in setting differs from the setting (e.g., primary versus secondary care; outpatients versus inpatients; a different country) may constitute indirectness of the population.

2. ***Were the outcomes matching the outcomes of interest? (Answer options: yes; probably yes; probably no; no)***

As mentioned before, indirectness of outcomes differs in value and preference versus treatment studies: surrogate outcomes warrant rating down for indirectness in the latter, but not the former. If, however, the outcomes considered in the available studies are not representative of the outcomes of interest, the confidence placed on the evidence is necessarily lower.

3. ***Are the options studied matching the alternative options of interest? (Answer options: Yes; probably yes; probably no; no)***

Whether the intervention options are a source of indirectness depends on to what extent the outcome considered is different when it is incurred by one intervention versus another. In studies to explore the relative importance of outcomes, the objective is to understand the importance participants place on the outcomes of interest. If we understand the relative importance patients place on the outcomes, and the probability of those outcomes occurring with the alternative management strategies under consideration, we can infer patients' choices (the formal way of doing so would be to conduct a decision analysis). Following this logic, differences between the management options used in the studies and those of interest represent another potential source of indirectness. Interventions may differ in many aspects – surgical skills or approaches, or drug dosages, durations,

or routes of administration route, but only when we are concerned that the differences in interventions would probably cause the difference in outcomes. Thus the difference in options is a signal to the potential differences in outcomes. If the focus is on individual outcomes, and we are confident the outcomes would be the same (preferably with the same description), irrespective of the intervention, we should not rate down for indirectness. For example, to understand the utility of bleeding as an adverse effect of antithrombotic therapy, it is irrelevant if the major bleed occurs as a result of aspirin, warfarin, or direct anticoagulant therapy. Some studies would ask participants to choose from options, for example, to choose stroke prevention strategy between aspirin versus no aspirin, or choose between warfarin and no treatment, then to infer their importance on outcomes according to the choice. In these studies, the outcomes, such as disease burden, or bleeding, may be weighed in differently, and the inference of the importance of outcome may be subsequently influenced.

Example of indirectness due to PICO elements: A systematic review summarized the relative importance patients placed on health states associated with benign prostatic hyperplasia: the assessment of symptom improvement, decreased prostate size, risks of acute urinary retention (AUR), and surgery. [49] It suggested men would wait longer for symptom improvement in exchange for decreased prostate size (13 months) than they would in exchange for an absolute 1% decrease in the risks of AUR (2 months) and surgery (8 months). However, for this valuation, 208 men aged $\geq$ 40 years from the general population were included. We consider the optimal study population in this case as aging male

population who are at the risk of benign prostatic hyperplasia. This is not the optimal study population because the study population (male ≥ 40 years) was generally younger than the population who are facing the decision. We rated the certainty of evidence down for indirectness of the population because the trade-off and valuation of outcomes involve AUR and surgery, which are usually not the decision most men aged ≥ 40 years old from general population would make.[49] In this case, other than probably indirect population, there was no concern for indirectness due to intervention or outcomes.

Meanwhile, although not optimal ageing males from the general population are at risk of prostatic hyperplasia and the presented considerations are not totally irrelevant for them. As this example demonstrates, the merit of GRADE approach is not to eliminate disagreement, but rather to provide a transparent and explicit assessment process.

3.5.2. Methodological aspects

Because, for some methods (e.g. indirect measurement of utilities with multiattribute utility approaches such as the EQ-5D and the health utilities index or Quality of Well-being instrument), relative importance of outcome is based on a linkage or transformation function (and thus the values are not those of the respondents, but of another population),[50] the methods used to elicit relative importance of outcomes also represent a source of indirectness.

Were the participants answering questions to directly value the relative importance of outcomes? (Answer options: yes; probably yes; probably no; no)

This question would be applicable whenever investigators have used an indirect measurement technique (i.e. a multiattribute utility index) to measure the utility of outcomes (utilities from EQ-5D, SF-6D, QWB or HUI) or when a mapping algorithm was used to estimate generic utility based on the estimates from other measurement (i.e., estimating EQ-5D utility from St. George Respiratory Questionnaire). To answer this question, users should also consider the sub-questions:

1. ***Were direct methodologies for outcome utilities rather than indirect methodologies used?***
2. ***Was the utility directly estimated from an instrument to elicit utilities rather than mapped from instrument whose purpose are not eliciting utility?***

When one asks patients to rate the value they place on health states, on can ask the question directly - asking patients to rate their own health state, or a clinical scenario, using the standard gamble, time trade off, and visual analogue scale. Multiattribute utility measurement instruments have used such direct measurement instrument, together with measurements on health domains to develop scoring systems for health state ratings, which is the algorithm to help transform measurements on health domains into utility. And then in application, the users of multiattribute utility measurement instruments would ask respondents to describe their own health state with the health domains. Thus, respondents are not providing their own evaluation of importance, but simply providing

information about their experience. The values come from someone else, and thus the rating of utility – and through utility of relative importance - is indirect. Essentially the same situation exists when researchers convert disease-specific quality of life scores into generic utilities; EQ-5D utility usually serves as the target measurement. In this case, indirect utilities are not estimated, but predicted from research results obtained using an instrument whose purpose was to assess the magnitude of disability, not to estimate the target measurement. Again, the values come from someone else, and thus are indirect.

However, depending on the perspective taken in the health care decision-making process, either in a healthcare policy decision making scenario, a clinical guideline development project, or a decision for an individual patient, the indirectness may not be a reason to rate down. If one accepts that the population completing a multiattribute utility instrument has the same relative importance of outcomes as the individuals who participated in the scenario rating that led to the weighting algorithm in the first place, then one might infer that ratings are those that would be provided by a direct assessment of relative importance of outcome. Making this assumption, one would not rate down due to indirectness.

3.5.3. Different strategies for systematic review authors and guideline panellists

In most cases, systematic review authors would only include studies in which the population, compared interventions, and outcomes meet the eligibility criteria, thus assuring directness.[48] However, in some situations, systematic review authors may include indirect evidence and rate down for indirectness with respect to their

population and outcome of primary interest. In contrast to systematic reviews, use of indirect evidence is very common in the setting of clinical practice guidelines.

These different purposes of utilizing and considering evidence in systematic reviews and guidelines could lead to the different indirectness judgment for the same body of evidence. As the previous example suggested, for a systematic review addressing the utility of bleeding, a major bleeding happens after taking aspirin is no more indirect compared to a major bleeding after taking warfarin. In contrast, in guideline development, whether the participants were valuing the importance of bleeding after taking warfarin or after taking aspirin matters.

3.6. Summary

This article describes how the GRADE approach can be applied to assess the certainty of evidence for the relative importance of outcomes when considering risk of bias and indirectness. When assessing certainty of evidence for the relative importance of outcomes evidence starts at "high" for all study designs, with rating down if risk of bias or indirectness are a serious concern. Users rate down by one or two levels depending on the specific considerations for the two domains.

Risk of bias assessment presents challenges. We have proposed a guiding set of questions to consider risk of bias issues; the reliability or validity of our suggested approaches remains unaddressed. Pending this work, using the signalling

questions and examples we have provided will help make judgments regarding risk of bias transparent.

In the next article, we will discuss the application of the other GRADE domains (imprecision, inconsistency, publication bias and upgrading domains) in assessment of certainty of relative importance of outcome and values and preferences evidence.

Table 3.1. Example of GRADE assessment for the certainty of evidence

Evidence profile
Author(s): Ray Yuan Zhang, Pablo Alonso Coello, Holger Schunemann Date: 2016-05-01
Question: What are the views about the relative value/importance of outcomes of interest in decision making for patients with antithrombotic treatment?
Setting: not specified **Bibliography:** MacLean S. Chest 2012; 141:e1S-e23S.

Quality assessment							Value (95%CI or other measure of variability)	Quality
Outcome	**Study Design/ Measurement instrument**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other**		
Stroke								
Non-fatal severe stroke	7 Cross-sectional studies, 580 participants VAS, SG, TTO	not serious[1,2,3,4]	No serious inconsistency	No serious indirectness	No serious imprecision	none	0.1-0.39 (range of the point estimates) 0.149, 95% CI: 0.135-0.163	⊕⊕⊕⊕ High
Moderate stroke	5 cross-sectional studies, 339 participants TTO, SG	not serious	Serious inconsistency[5,6]	No serious indirectness	No serious imprecision	none	0.29-0.77 (range of the point estimates) 0.664, 95% CI: 0.643 - 0.684	⊕⊕⊕◯ Moderate
Bleeding								
Major (unspecified) GI Bleeding	3 cross-sectional studies, 153 participants VAS, TTO and SG	not serious[1,3]	No serious inconsistency	No serious indirectness	No serious imprecision	none	0.65-0.84 (range of the point estimates) 0.789, 95% CI: 0.758 - 0.820	⊕⊕⊕⊕ High
PPS								
Severe PPS	2 cross-sectional studies, 66participants SG	not serious[7]	No serious inconsistency	Serious indirectness[8]	Serious imprecision[9]	none	0.93 - 0.982 (range of the point estimates) 0.973, 95% CI: 0.964 - 0.982	⊕⊕◯◯ Low
DVT								
DVT and VTE, and bleeding	1 cross-sectional study[10], 124 participants	not serious	No serious inconsistency	No serious indirectness	No serious imprecision	none	If there are a 3% chance of a major bleeding event, and a 2% chance of a recurrent episode of venous	⊕⊕⊕⊕ High

	Time trade off						thromboembolism in the next 2 years, the rates of recurrence of DVT without treatment varied from 5%, 10% to 15%, the percentage of participants choosing to stop the VKA treatments are 21%, 23% and 8%, respectively.	
Burden of treatment								
Burden of treatment: warfarin	7 Cross-sectional studies, 466participants VAS, SG, TTO	not serious [1,2,3,4]	No serious inconsistency	No serious indirectness	No serious imprecision	none	0.66-1 (range of estimates across included studies) 0.938, 95% CI: 0.934-0.942	⊕⊕⊕⊕ High
Burden of treatment: anticoagulant/ warfarin	1 qualitative study, 21 participants Semi-structured interview[11]	not serious	No serious inconsistency	No serious indirectness	Serious imprecision[1 2]	none	The majority (specific percentage not reported) of participants had not experienced complications due to warfarin. Many participants reported only minor inconveniences, such as taking a pill every day, regular blood tests, and dietary changes.	⊕⊕⊕◯ Moderate

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

AF: Atrial Fibrillation; CI: Confidence interval; GI bleeding: Gastrointestinal bleeding; PPS: postphlebitic syndrome, SG: Standard Gamble; TTO: Time Trade Off; VAS: Visual Analogue Scale.

1. The representativeness of the studies was impacted by a low response. However, this only impacted a small proportion of the included study population.
2. In Protheroe 2000, 97 of 260 invited patients responded.
3. In Thomson 2000, 57 of the 180 invited patients completed the interview.
4. 17.4% of participants in Gage 1995 did not understand the time trade off technique.
5. Wide variation across point estimates.
6. The included study population were patients with atrial fibrillation (Gage 1996), 30 community volunteer (Lenert 1997), three different patient population (patients with a 1st or 2nd episode of venous thromboembolism, with oral anticoagulants had been started, patients who had experienced an episode of major bleeding during oral anticoagulant treatment, and patients with a postthrombotic syndrome in Locadia 2004), both patients with DVT and without DVT (O'Mera 1994) as well as ischemic stroke survivors and age-matched control subjects (Slot 2009).
7. One of the studies was judged to be of high risk of bias. However, this study had similar estimates with the other one with low risk of bias.
8. The certainty of evidence was downgraded for indirectness. The included studies have different population than the patients facing the choice: 30 community volunteer (Lenert 1997), patients with DVT and without DVT (O'Mera 1994).
9. Small sample size: 66 participants from 2 studies.
10. Locadia 2004 is a cross-sectional study interviewing participants with decision analysis.
11. Dantas 2004 is a qualitative study on the burden of anticoagulant/ warfarin treatment.
12. Only one qualitative study identified to address this phenomenon.

Table 3.2. Risk of bias subdomains and signalling questions

Subdomain	Signalling questions
Risk of bias due to selection of participants into the study	Was an appropriate study sample selected from the sampling frame?
Risk of bias due to missing data	Was the attrition sufficiently low to minimize the risk of bias?
Risk of bias due to the measurement instrument	Is the chosen the instrument for eliciting relative importance of outcomes valid and reliable?
	Was the instrument administered in the intended way?
	Was a valid representation of the outcome (health state) utilized?
	Did the researchers check the understanding of the instrument?
Risk of bias due to confounding	Were the results analyzed appropriately to avoid influence of confounding?

Table 3.3. Signalling questions for indirectness

Sources of indirectness	Signalling questions
Indirectness due to PICO elements	Was the population studied matching the population of interest?
	Were the options studied matching the alternative options of interest?
	Were the outcomes matching the outcomes of interest?
Indirectness due to methodological elements	Were the participants answering questions directly valuing the relative importance of outcomes? • Were direct methodologies for outcome utilities rather than indirect methodologies used? • Was the utility directly estimated from an instrument to elicit utility rather than mapped from instrument whose purpose are not eliciting utility?

Abbreviations

AUR: acute urinary retention

CERQUAL: The Confidence in the Evidence from Reviews of Qualitative research

DCE: discrete choice experiment

EBM: evidence-based medicine

EQ-5D: EuroQual-5-dimension (a quality of life measurement tool)

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HUI: health utility index

MeSH: Medical Subject Headings

PICO: population, intervention, comparison and outcome

RIO: relative importance of outcomes

QWB: quality of wellbeing

SF-6D: Short form-6-dimension (a quality of life measurement tool)

Declarations

Ethics approval and consent to participate

Not required. This study does not involve de novo patient data collection. No patient informed consent and Institutional Review Board approval have been sought.

Consent for publication

Not applicable.

Availability of Data and Materials

The datasets supporting the conclusions of this article are included within the article and its additional file.

Authors' contributions

YZ, PA, GG, HJS designed the methodology for this project. HJS conceived of the project. YZ, JJYN, and YC summarized the certainty assessment for relevant items in systematic reviews; YZ, PA, GG, and HJS proposed the adapted GRADE domains and subdomains for certainty of evidence assessment; all authors participated in the methodological discussions. All authors read and approved the final manuscript.

Competing interests

All the authors are member of the GRADE working group. None of the authors have finial conflict of interests.

Acknowledgements

We are grateful to Dr. Amiram Gafni from McMaster University for the comments on the manuscripts.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. It was funded through internal research funds at McMaster University available to HJS.

3.7. Reference

1. Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *JAMA : the journal of the American Medical Association* 2008;300(21):2483; author reply 83-4. doi: 10.1001/jama.2008.730 [published Online First: 2008/12/04]
2. Schunemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014;186(3):E123-42. doi: 10.1503/cmaj.131237 [published Online First: 2013/12/18]
3. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed)* 1996;312(7023):71-2. [published Online First: 1996/01/13]
4. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e1S-23S. doi: 10.1378/chest.11-2290 [published Online First: 2012/02/15]
5. Stiggelbout AM, Van der Weijden T, De Wit MP, et al. Shared decision making: really putting patients at the centre of healthcare. *BMJ (Clinical research ed)* 2012;344:e256. doi: 10.1136/bmj.e256 [published Online First: 2012/01/31]
6. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. *Evidence Based Medicine* 2002;7(2):36-38. doi: 10.1136/ebm.7.2.36
7. van der Weijden T, Legare F, Boivin A, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. *Implementation science : IS* 2010;5:10. doi: 10.1186/1748-5908-5-10 [published Online First: 2010/03/09]
8. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of clinical epidemiology* 2013;66(7):726-35. doi: 10.1016/j.jclinepi.2013.02.003 [published Online First: 2013/04/11]
9. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA : the journal of the American Medical Association* 2008;300(4):436-8. doi: 10.1001/jama.300.4.436 [published Online First: 2008/07/24]
10. Gafni A. The standard gamble method: what is being measured and how it is interpreted. *Health services research* 1994;29(2):207-24. [published Online First: 1994/06/01]
11. Torrance GW. Measurement of health state utilities for economic appraisal. *Journal of health economics* 1986;5(1):1-30. [published Online First: 1986/02/09]

12. Torrance GW. Utility measurement in healthcare: the things I never got to. *PharmacoEconomics* 2006;24(11):1069-78. [published Online First: 2006/10/28]
13. Churchill DN, Torrance GW, Taylor DW, et al. Measurement of quality of life in end-stage renal disease: the time trade-off approach. *Clinical and investigative medicine Medecine clinique et experimentale* 1987;10(1):14-20. [published Online First: 1987/01/01]
14. Dolan P, Gudex C, Kind P, et al. The time trade-off method: results from a general population study. *Health economics* 1996;5(2):141-54. doi: 10.1002/(sici)1099-1050(199603)5:2<141::aid-hec189>3.0.co;2-n [published Online First: 1996/03/01]
15. Torrance GW, Feeny D, Furlong W. Visual analog scales: do they have a role in the measurement of preferences for health states? *Medical decision making : an international journal of the Society for Medical Decision Making* 2001;21(4):329-34. [published Online First: 2001/07/28]
16. Morimoto T, Fukui T. Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. *Journal of epidemiology / Japan Epidemiological Association* 2002;12(2):160-78. [published Online First: 2002/05/30]
17. Craig BM, Busschbach JJ, Salomon JA. Modeling ranking, time trade-off, and visual analog scale values for EQ-5D health states: a review and comparison of methods. *Medical care* 2009;47(6):634-41. doi: 10.1097/MLR.0b013e31819432ba [published Online First: 2009/05/13]
18. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. *Annals of medicine* 2001;33(5):337-43. [published Online First: 2001/08/09]
19. Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied health economics and health policy* 2003;2(1):55-64. [published Online First: 2003/11/19]
20. Ryan M. Discrete choice experiments in health care. *BMJ (Clinical research ed)* 2004;328(7436):360-61.
21. Stevens TH, Belkner R, Dennis D, et al. Comparison of contingent valuation and conjoint analysis in ecosystem management. *Ecological Economics* 2000;32(1):63-74. doi: http://dx.doi.org/10.1016/S0921-8009(99)00071-3
22. Alonso-Coello P, Montori VM, Diaz MG, et al. Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. *Health expectations : an international journal of public participation in health care and health policy* 2014 doi: 10.1111/hex.12201 [published Online First: 2014/05/13]
23. Devereaux PJ, Anderson DR, Gardner MJ, et al. Differences between perspectives of physicians and patients on anticoagulation in patients with atrial fibrillation: observational study. *BMJ (Clinical research ed)* 2001;323(7323):1218-22. [published Online First: 2001/11/24]
24. Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of

approaches and recommendations for standardization. *Patient Education and Counseling* 2010;78(1):12-23. doi: http://dx.doi.org/10.1016/j.pec.2009.05.011

25. King M, Nazareth I, Lampe F, et al. Conceptual framework and systematic review of the effects of participants' and professionals' preferences in randomised controlled trials. *Health Technology Assessment* 2005;9(35):iii-68.
26. Cronin M, Meaney S Fau - Jepson NJA, Jepson Nj Fau - Allen PF, et al. A qualitative study of trends in patient preferences for the management of the partially dentate state. (1741-2358 (Electronic))
27. DeJean D, Giacomini M, Vanstone M, et al. Patient experiences of depression and anxiety with chronic disease: a systematic review and qualitative meta-synthesis. *Ontario health technology assessment series* 2013;13(16):1-33. [published Online First: 2013/11/15]
28. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical research ed)* 2008;336(7650):924-6. doi: 10.1136/bmj.39489.470347.AD [published Online First: 2008/04/26]
29. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]
30. Lewin S, Glenton C, Munthe-Kaas H, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS medicine* 2015;12(10)
31. Yepes-Nuñez JJ, Zhang Y, Xie F, et al. 42 systematic reviews generated 22 items for assessing the risk of bias in Values and Preferences´ studies, 2016.
32. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology* 2011;64(12):1311-6. doi: 10.1016/j.jclinepi.2011.06.004 [published Online First: 2011/08/02]
33. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *Journal of clinical epidemiology* 2011;64(4):383-94. doi: 10.1016/j.jclinepi.2010.04.026 [published Online First: 2011/01/05]
34. Iorio A, Spencer FA, Falavigna M, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *BMJ (Clinical research ed)* 2015;350:h870. doi: 10.1136/bmj.h870 [published Online First: 2015/03/18]
35. Viswanathan M, Ansari MT, Berkman ND, Chang S, Hartling L, McPheeters LM, Santaguida PL, Shamliyan T, Singh K, Tsertsvadze A, Treadwell JR. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. Agency for Healthcare Research and Quality Methods Guide for Comparative Effectiveness Reviews. March 2012. AHRQ Publication No. 12-EHC047-EF. Available at: http://www.effectivehealthcare.ahrq.gov/

36. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ (Clinical research ed)* 2008;336(7653):1106-10. doi: 10.1136/bmj.39500.677199.AE [published Online First: 2008/05/17]
37. Levin KA. Study design III: Cross-sectional studies. *Evid-based Dent* 2006;7(1):24-25.
38. Fincham JE. Response rates and responsiveness for surveys, standards, and the Journal. *American journal of pharmaceutical education* 2008;72(2):43. [published Online First: 2008/05/17]
39. Lenert LA, Soetikno RM. Automated computer interviews to elicit utilities: potential applications in the treatment of deep venous thrombosis. *Journal of the American Medical Informatics Association : JAMIA* 1997;4(1):49-56. [published Online First: 1997/01/01]
40. Thomson R, Parkin D, Eccles M, et al. Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. *Lancet (London, England)* 2000;355(9208):956-62. doi: 10.1016/s0140-6736(00)90012-6 [published Online First: 2000/04/18]
41. Polonsky WH, Fisher L, Hessler D, et al. Patient perspectives on once-weekly medications for diabetes. *Diabetes, obesity & metabolism* 2011;13(2):144-9. doi: 10.1111/j.1463-1326.2010.01327.x [published Online First: 2011/01/05]
42. Arnesen T, Trommald M. Roughly right or precisely wrong? Systematic review of quality-of-life weights elicited with the time trade-off method. *Journal of health services research & policy* 2004;9(1):43-50. doi: 10.1258/135581904322716111 [published Online First: 2004/03/10]
43. Arnesen T, Trommald M. Are QALYs based on time trade-off comparable?--A systematic review of TTO methodologies. *Health economics* 2005;14(1):39-53. doi: 10.1002/hec.895 [published Online First: 2004/09/24]
44. Lin MR, Hwang HF, Chung KP, et al. Rating scale, standard gamble, and time trade-off for people with traumatic spinal cord injuries. *Physical therapy* 2006;86(3):337-44. [published Online First: 2006/03/02]
45. Gage BF, Cardinalli AB, Albers GW, et al. Cost-effectiveness of warfarin and aspirin for prophylaxis of stroke in patients with nonvalvular atrial fibrillation. *JAMA : the journal of the American Medical Association* 1995;274(23):1839-45. [published Online First: 1995/12/20]
46. Guyatt GH, Oxman AD, Vist G, et al. GRADE guidelines: 4. Rating the quality of evidence--study limitations (risk of bias). *Journal of clinical epidemiology* 2011;64(4):407-15. doi: 10.1016/j.jclinepi.2010.07.017 [published Online First: 2011/01/21]
47. Joy SM, Little E, Maruthur NM, et al. Patient preferences for the treatment of type 2 diabetes: A scoping review. *PharmacoEconomics* 2013;31(10):877-92. doi: http://dx.doi.org/10.1007/s40273-013-0089-7
48. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. *Journal of clinical epidemiology*

2011;64(12):1303-10. doi: 10.1016/j.jclinepi.2011.04.014 [published Online First: 2011/08/02]

49. Emberton M. Medical treatment of benign prostatic hyperplasia: physician and patient preferences and satisfaction. *Int J Clin Pract* 2010;64(10):1425-35. doi: http://dx.doi.org/10.1111/j.1742-1241.2010.02463.x
50. Brazier JE, Yang Y, Tsuchiya A, et al. A review of studies mapping (or cross walking) non-preference based measures of health to generic preference-based measures. *The European journal of health economics : HEPAC : health economics in prevention and care* 2010;11(2):215-25. doi: 10.1007/s10198-009-0168-z [published Online First: 2009/07/09]

Chapter 4. Development of GRADE guidance for assessing the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 2. Inconsistency, Imprecision and other issues

Yuan Zhang[1], Pablo Alonso Coello[1,2], Gordon Guyatt[1], Juan Jose Yepes-Nunez[1], Elie A. Akl[1,3], Glen Hazlewood[4], Hector Pardo-Hernandez[2], Itziar Etxeandia-Ikobaltzeta[1], Amir Qaseem[5], John W. Williams Jr. [6], Peter Tugwell[7], Yaping Chang[1], Yuqing Zhang[1], Reem A. Mustafa [1, 8], Holger Schünemann[1, 9]
1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada
2. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
3. Department of Internal Medicine, Faculty of Medicine, American University of Beirut, Lebanon
4. Department of Medicine and Department of Community Health Sciences, University of Calgary, Canada
5. American College of Physicians, Philadelphia, Pennsylvania, USA
6. Center of Innovation for Health Services Research in Primary Care at the Durham Veterans Affairs Medical Center and Duke University, Durham, NC 27701, USA.
7. Department of Medicine, University of Ottawa, Canada
8. Department of Internal Medicine, Division of Nephrology and Hypertension, University of Kansas Medical Center, Kansas City, Kansas, USA
9. Department of Medicine, McMaster University, Canada

Corresponding author:
Holger J Schünemann, MD, PhD
Chair, Department of Health Research Methods, Evidence, and Impact
McMaster University Health Sciences Centre, Room 2C16
1280 Main Street West
Hamilton, ON, L8N 4K1, Canada
Email: schuneh@mcmaster.ca
Tel: +1 905 525 9140 x 24931

Although this manuscript is final for the doctoral thesis of Yuan Zhang, this is a GRADE working group project, and the final manuscript will require approval by the GRADE Working Group. Members of the GRADE working group making appropriate contributions to this manuscript at that stage may become authors on the byline.

Abstract

This is the second of two articles by the GRADE Working Group providing guidance on how users can assess the certainty of a body of evidence for values and preferences. We conceptualize values and preferences as the *relative importance of outcomes (RIOs)*. Following the discussion in the previous article on risk of bias and indirectness, we describe the rationale for considering the remaining GRADE domains when rating the certainty in a body of evidence for relative importance of outcomes. As meta-analyses are uncommon in this context, inconsistency and imprecision assessments are challenging. We are aware of confusion about inconsistency, imprecision and true variability in relative importance of outcomes. To clarify this issue, we suggest that the true variability in relative importance of outcomes is neither equivalent to inconsistency nor to imprecision. Specifically, inconsistency arises from PICO and methodological elements that should be explored and, if possible, explained. The width of the confidence interval and sample size should inform judgments about imprecision. We also provide suggestions on how to detect publication bias and discuss the domains to rate up the certainty. In conclusion, we describe guidance for rating inconsistency, imprecision and other domains when evaluating a body of evidence describing relative importance of outcomes according to GRADE and provide an example for a complete rating.

4.1. Introduction

Incorporating values is gaining increasing attention in evidence-based decision-making, but it is still unclear how users should assess certainty of this type of evidence.[1-6] The GRADE working group has developed approaches for assessing certainty of evidence for treatment, diagnosis and prognosis questions. The GRADE working group defines values and preferences as the relative importance people place on the outcomes of interest resulting from a decision.[7] We have addressed the GRADE domains of risk of bias and indirectness for relative importance of outcome evidence in the previous article.[8] Here, we will provide guidance on rating inconsistency and imprecision, and describe other related issues including publication bias, rating up the certainty of evidence, and variability.

4.2. Methodology

We described the detailed methods for this work in the previous article.[8] Briefly, we utilized a multi-pronged approach to develop guidance for assessing the certainty of a body of evidence addressing the importance of outcomes or values and preferences, from here on called the relative importance of outcomes (RIO). We applied this GRADE approach to examples from systematic reviews in group discussions of GRADE project group meetings and consulted stakeholders for feedback. We applied the same GRADE domains (risk of bias, inconsistency, indirectness, imprecision, publication bias, and domains to rate up the evidence)

to relative importance of outcomes ratings in these systematic reviews and developed guidance based on these examples.[9]

4.3. Inconsistency

According to the GRADE approach, certainty of the evidence can be lowered if there is unexplained inconsistency or heterogeneity. Assessment of inconsistency of evidence about relative importance of outcomes is challenging for several reasons. Existing systematic reviews often lack a clear definition of values and preferences and include a diverse set of methods and instruments to assess them.[10-13] As a result quantitative synthesis of relative importance of outcomes is uncommon because systematic review authors are hesitant to pool estimates obtained with different instruments such as the standard gamble, time trade off or rating scales. This creates a dilemma for interpretation of systematic reviews as qualitative rather than quantitative syntheses. In other situations where methods such as discrete choice, willingness to pay, rankings, or other scales are used there is often only one single study available. The judgment about inconsistency is straightforward in the latter case because inconsistency does not exist in the context of single study evidence (a body of evidence based on one study will likely be rated down for one or more of the other GRADE domains). While we suggest that raters attempt to statistically pool the relative importance of outcomes if appropriate, the assessment of inconsistency follows the same principles that we suggest below if no pooled estimates are available. We focus here on quantitative

estimates of relative importance of outcomes. For qualitative evidence we refer readers to the work of the GRADE-CERQUAL project group. [14]

We propose raters examine inconsistency for certainty of a body of evidence of relative importance of outcomes in the following way: 1) answering a signalling question, 2) exploring heterogeneity if the results across studies are inconsistent and not rating down if inconsistency can be explained (Figure 4.1), 3) discussing the credibility of subgroup effects if they are detected. We will begin with describing the signalling questions.

4.3.1. Signalling question: are the results across the included studies consistent?

The four items for assessing inconsistency in results are: similarity in point estimates, overlap in confidence intervals, statistical test for heterogeneity, and I^2 statistics. We suggest that the evaluation of point estimates and confidence intervals is by visual inspection. If meta-analyses are available, the statistical test for heterogeneity and I^2 allow for quantitative estimates of inconsistency.[15]

1. Similarity in point estimates: Large differences in point estimates suggest important heterogeneity across studies; whether or not these differences are due to chance is informed by examining confidence intervals to determine whether they are wide and overlap. If they do, random error or chance could be a plausible explanation for the observed difference.
2. Overlap in confidence intervals: If the confidence intervals overlap, random error becomes a likely explanation. Very large studies may be so

precise that even small or even trivial differences in point estimates have no overlap in confidence intervals.

3. Statistical test for heterogeneity: If a meta-analysis is available, a low *P* value suggests that differences in study results have a low probability of being due to chance. In the case of small sample sizes in the included studies, the test of heterogeneity may not have sufficient power and a simple "yes" or "no" answer according to statistical significance could be misleading.
4. I^2: I^2 quantifies the proportion of the variation explained by among-study differences as opposed to the total observed variance. I^2 may be misleading when the study sample sizes are small or very large.[16] Even if the point estimates vary, the I^2 may be low by chance, while when the sample size is large, even a small difference between studies can lead to a large I^2. [16] Simple thresholds for the I^2 are therefore misleading. Values above 50% should lead to very careful examination of heterogeneity in the context of the other inconsistency items.

If examination of the items above suggests no important inconsistency, raters label the inconsistency domain as "not serious" and affirm the signalling question. If examination of point estimates, confidence interval, statistical test and I^2 suggests substantial inconsistency, raters should consider exploring the source.

4.3.2. Detailed exploration of inconsistency

When exploring inconsistency focusing on the PICO elements, raters should consider:

1. ***Are the populations studied consistent across studies? (Answer options: Yes; probably yes; probably no; no)***
2. ***Are the Intervention and comparison consistent across studies? (Answer options: Yes; probably yes; probably no; no)***
3. ***Are the type of outcomes consistent across studies? (Answer options: Yes; probably yes; probably no; no)***

The answers "yes" or "probably yes" suggests that the item is not a cause of inconsistency, while "probably no" or "no" indicate that an explanation for inconsistency may exist.

When results are not consistent, raters explore inconsistency in the following ways. Raters should evaluate differences in the population, e.g. demographic characteristics such as gender, age or cultural background, the options, e.g. the dose, duration or administration of medication in direct choice experiments, and outcomes assessed, e.g., same outcome but different severities. Raters should be aware that these elements are not completely independent, e.g. when patients assess the relative importance of outcomes of their own health outcome, disease severity is an element both related to population and the outcome. As discussed in the indirectness section, for the relative importance of outcomes, the difference in options would be a signal to suggest potential difference in outcomes. The consideration of treatment options is more relevant when a study evaluates direct choice and infers the outcome importance accordingly.

Methodological inconsistency may arise from the risk of bias stemming from the study design, measurement methodology (e.g., standard gamble or time trade off,

and visual analogue scale for utility), or description of outcomes or health states, such as narrative versus point by point format of health states, detailed versus less detailed descriptions. An example is that inconsistency may be explained if different specific outcomes are measured such as ischemic stroke versus hemorrhagic stroke or widely differing descriptions for severe stroke.

Thus, raters should assess whether the methodological approach represents a plausible explanation for inconsistency. Raters should ask:

1. ***Are the study designs consistent across studies?***
2. ***Are instruments consistent across studies?***
3. ***Are the descriptions or definitions of disease severities and outcomes consistent across studies?***

The answer options for the three questions (for each of the three methodological approaches) above are:

1. ***All included studies had similar study designs/instruments/ descriptions or definitions of disease severities and outcomes.***
2. ***Most studies included had a similar study designs/instruments/ descriptions or definitions of disease severities and outcomes.***
3. ***All included studies had different study designs/instruments/ descriptions or definitions of disease severities and outcomes.***

If inconsistency remains unexplained, raters should lower the certainty of evidence for inconsistency by either one or two levels.[15]

4.3.3. Credibility of subgroup estimates

Raters should formulate *a priori* hypotheses to explore inconsistency due to potential subgroup effects. If subgroup analyses show differences, raters should judge the credibility of the pre-specified subgroup effects. While frameworks for evaluating subgroup effects of treatments exist, [17] they do not exist for assessing relative importance of outcomes. Until further guidance is available, we suggest raters ask the following signalling question (Figure 4.1):

If a subgroup analysis was conducted to explore the source of inconsistency, are the subgroup estimates credible? [17] (***Answer options: Yes; probably yes; probably no; No***)

The answers "yes" or "probably yes" suggests that subgroups are a cause of inconsistency, while "probably no" or "no" indicate that the subgroup effects are credible.

4.3.4. Different strategies for systematic review authors and guideline panellists

Systematic review authors should only combine results if the results are similar enough. This can begin with pooling across studies and then test the assumption of similarity across studies. If there is substantial heterogeneity and systematic review authors discover that PICO or methodological elements are a source of heterogeneity, they should summarize and present the results for these groups of patients or people, compared alternatives, or outcomes.

Guideline developers can then formulate recommendations separately for subgroups with different values or they can formulate a conditional or weak

recommendation across populations indicating that how differences in values affect implementation of the recommendation.

4.3.5. Variability versus inconsistency

When referring to inconsistency or heterogeneity across studies for relative importance of outcomes we suggest avoiding the term "variability". Variability may be misinterpreted as broad distribution of values or relative importance of outcomes within included studies. True variability of relative importance of outcomes within studies requires a separate assessment.

4.3.6. Example of inconsistency

A systematic review summarized the relative importance of outcomes of psoriasis patients using willingness to pay and utilities. Two included studies elicited willingness to pay for health states using the same instrument. However, important differences existed for willingness to pay for physical comfort, social comfort, emotional health, self-care, intimacy, ability to sleep, ability to work/volunteer and ability to concentration across these two studies ($2000 vs $10,000, $1000 vs $2000, $2000 vs $5000, $1500 vs $9500, $1000 vs $5000, $625 vs $10000, $1600 vs $10000 and $875 vs $7500, respectively).[18] There were also no PICO elements that explained these differences and one could justify rating down the certainty of evidence for serious inconsistency.

4.4. Imprecision

Rating imprecision for relative importance of outcomes includes an assessment of both the confidence interval and sample size for the body of evidence. This

assessment is often challenging because there rarely are meta-analyses and, thus, no calculated confidence intervals. For the same reason, there is not a simple way to calculate the minimum sample size to produce a sufficiently narrow estimate with sufficient power for relative importance of outcomes.[19] However, we suggest raters take the following approach (Figure 4.2).

4.4.1. Confidence interval of relative importance of outcomes

Systematic review authors are often not in the best position to judge whether the confidence interval (CI) around the estimate is sufficiently narrow for a specific decision. This is because this rating is often dependent on the context. Therefore, they should make their rationale for their judgments explicit, such as accepting a certain range or assuming that decisions would be influenced or not influenced by the width of the confidence interval. Furthermore, due to diversity in study designs, instruments and presentation of results, confidence intervals may not be available. For example, the result could be reported as the proportion of people willing to accept a prophylactic treatment if the risk of adverse events decreased by 5%. Under those circumstances rating imprecision will likely be based on the number of studied people (sample size) alone.

For guideline panellists, we suggest rating imprecision based on whether the confidence intervals of the relative importance of outcomes evidence cross a decision threshold. This requires taking particular absolute effect estimates of interventions on the outcomes into account for which the relative importance of outcomes are obtained. Imprecision is not present if the net benefits clearly outweigh net harms after combing the relative importance of outcomes and the

absolute effect estimates, or vice versa, regardless of whether the upper or lower limit of the confidence interval of the relative importance of outcomes estimate is assumed to be true. If the decision would be overturned by assuming alternative estimates for the relative importance of outcomes stemming from the confidence interval around them, raters would judge the evidence as seriously or very seriously imprecise.

Thus, raters should first inspect the width of the confidence interval, and ask the question:

1. ***Is the confidence interval narrow (for systematic reviews)? (Answer options: Yes; probably yes; probably no; no)***
2. ***Does the confidence interval exclude the clinical decision threshold between recommending and not recommending a treatment (for clinical practice guidelines)? (Answer options: Yes; probably yes; probably no; no)***

The answers "yes" or "probably yes," indicate no important imprecision, while "probably no" or "no" suggest important imprecision.

4.4.2. Sample size

Raters should also consider the sample size across studies when assessing imprecision:

Is the sample size large enough to sufficiently reduce the risk of chance (both for systematic reviews and guideline development)? (Answer options: Yes; probably yes; probably no; no)

The answers "yes" or "probably yes" suggest no imprecision, while "probably no" or "no" suggest imprecision. To assess imprecision, the review information size

could be used as a threshold and its calculation is likely to be different depending on the estimates used to indicate the relative importance of outcomes (e.g., utility, rank, or willingness to pay for an attribute).[20]

For studies on direct choice, a potential approach when the benefits and harms are closely balanced and a judgment about the direction of a recommendation (for or against an option) needs to be made, is to use a threshold of 55% (i.e. that more than 55% of patients would make the same choice) for making a recommendation in one direction or the other.[21 22] For a single group the sample size required to estimate this proportion (55%) with a confidence level of 95% and a desired precision of 5% would be 380 people.[21] This leads to a rule of thumb for judgments about whether to make a recommendation in one direction or the other when the relative importance of the benefits and harms is critical to that judgment. Estimates of patients' values for the relative importance of outcomes should be based on studies that include at least 380 participants. In situations when there is a potential large net benefit and one needs to decide about the strength of the recommendation (strong versus conditional) the threshold suggested by the GRADE Working Group that more than 80% or 90% of patients would make the same choice could be used.[21 22] For a direct choice the sample size required to estimate these proportions (80% or 90%) with a confidence level of 95% and a desired precision of 5% would be 246 and 139 participants, respectively. Thus, a rule of thumb for judgments about whether to make a strong recommendation when the relative importance of the benefits and harms is critical to that judgment might be that estimates of patients' values should be based on studies that include

at least 250 patients or 140 patients depending on if 80% or 90% are required for a strong recommendation.

In most situations direct choice studies are not available and for systematic reviews judgments are made on a per outcome level. Under those circumstances we suggest making a priori assessments of acceptable width of the confidence interval for decision-making or using, again, a rule of thumb for sufficient precision of the relative importance of outcomes estimate. For example, a width of the confidence interval of 0.1 on a utility scale could be utilized to calculate the required review information size based on a significance level (α), statistical power level (1-β) and mean value for the relative importance of outcomes.

4.4.3. Example of imprecision in relative importance of outcomes

In a systematic review summarizing the utilities of head and neck cancer related outcomes, the utility of radiotherapy was 0.66 (measured with rating scale), 0.70 (time trade off) and 0.61 (standard gamble) for laryngeal cancer and 0.78 (rating scale), 0.72 (time trade off) and 0.683 (standard gamble) for floor-of-the-mouth cancer. No confidence intervals were reported. Noticeably, these estimates came from a study of patients with previous history of disease (10 patients with a previous history of laryngeal cancer and 10 patients with a previous history of floor-of-the-mouth cancer), as well as physicians. [23] We rated down the certainty of evidence because of imprecision due to the small sample size.

4.5. Publication bias

Publication bias may be important for evidence addressing relative importance of outcomes. While the reasons for publication bias for this type of evidence may differ importantly from those of intervention studies where for-profit interest is likely to play a role, other reasons for failure to publish (in a timely manner) may be similar. Conceivable reasons for delayed or unsuccessful publication include: the results are not consistent with what was previously shown, results are redundant and language or cultural circumstances lead to delays or failure to publish. Unfortunately, we are not aware of empirical evidence about the extent or how to properly assess publication bias in this field. Only in the situation that users have proof (through knowledge of conducted but unpublished studies) or strong suspicion of publication bias, they should consider rating down the certainty of evidence for publication bias.[24]

4.6. Rating up

The theoretical basis and empirical examples for applying existing reasons for rating up the certainty of evidence (a large effect, dose-response gradient and direction of plausible confounding) in evidence of relative importance of outcomes is limited. Thus far, we do not have clear guidance for when the evidence of relative importance of outcomes should be rated up but we will describe some plausible scenarios here. As we will outline below, the certainty of a body of evidence summarizing relative importance of outcomes starts as high certainty of the evidence. In other GRADE guidance for increasing the certainty

of evidence, we suggest that raters only upgrade studies that are unlikely to be prone to bias.[25] The same considerations apply here. Thus, rating up relative importance of outcomes evidence is likely to rarely apply. Conceivable situations for what would be analogous to dose-effect relations are clear gradients in relative importance of outcomes across different severity levels of marker states. Indeed, in a systematic review on how chronic obstructive pulmonary disease (COPD) patients value their outcomes, the pooled estimates for EQ-5D measurements of mild, moderate, severe, and very severe COPD are 0.821 (95% CI: 0.814-0.828), 0.760 (95% CI: 0.756-0.765), 0.727 (95% CI: 0.722-0.732), respectively. Although we observed heterogeneity for the utility of COPD states across studies, we also identified a gradient of disutility as the disease progresses. This should increases our overall confidence, but we need to further explore whether or not we could rate up for each outcome. Other plausible situations are when the main consideration is if two health states differ importantly in their relative importance of outcomes value. If the difference is precise and the studies not importantly biased, overall certainty in the difference is increased by large observed differences, in particular those that exceed the minimal important difference of relative importance of outcomes such as 0.05 to 0.07 on a 0 to 1 visual analogue scale.[26] As we continue to develop the GRADE approach for evidence addressing the relative importance of outcomes, the current GRADE domains for rating up certainty of evidence may evolve or other reasons to rate up may arise.

4.7. Distribution (variability) of the relative importance of outcomes

4.7.1. Distribution across individuals and decision-making scenarios

In many clinical and methodological discussions, people use the term "variability" of values but there is ambiguity in how the term is used. The term has been indistinctly used to refer to the inconsistency of results across studies (inconsistency), the width of the confidence intervals (imprecision), or the distribution within a population. We suggest that these concepts are kept apart. We described our approach for addressing inconsistency and imprecision and the reasons for rating down for these domains above. When we refer to variability we mean biological variability for which there is no (current) explanation. For example, patients show a large degree of variability with regards to how they value the relative importance of gastrointestinal bleeds in the context of stroke prevention (Figure 4.3). The reasons for why some patients are very averse to bleeding events and others only somewhat averse are not well understood. This variability can lead to both inconsistency across studies (if patients with different although unknown predictors for how they rate the relative importance of outcomes are included) or imprecision (if there is a large number of patients with different values for the relative importance of bleeding within studies).

4.7.2. Deciding on the importance of variability

We would further explore the assessment of the distribution or the variability of the evidence about relative importance of outcomes. In this guidance, we do not

suggest rating the certainty in variability of relative importance of outcomes but making the potential for underlying variability transparent. Systematic review authors are not in the best position to judge whether the variability of the relative importance of outcomes is important or not because it is a context-specific decision. This decision requires balancing of all outcomes and other GRADE Evidence to Decision (EtD) criteria. However, they should make explicit their rationale.

Variability in how patients value the main outcomes will influence the strength of a recommendation. Guideline panellists should consider whether there is potential variability important enough for them to make different recommendations across the range of variable relative importance of outcomes. If this is the case, panel should consider making a conditional or weak recommendation.

4.8. Summary

We describe how the GRADE approach could be applied to assess the certainty of evidence for relative importance of outcomes. The general process of assessing certainty of evidence for relative importance of outcomes evidence is starting at "high" for all outcome assessments. Raters lower the certainty if risk of bias, inconsistency, indirectness, imprecision or publication bias are serious or very serious for a specific domain before determining the final certainty of a body of evidence as high, moderate, low or very low by considering the judgments across all domains (Figure 4.4).

In the example shown in Table 3.1, we assessed the certainty of evidence for utility of non-fatal severe stroke. Seven studies on 580 participants were included,[27-33] the pooled estimate was 0.149 (95% CI: 0.135-0.163) on a scale from 0 to 1. For some of the studies included, we had risk of bias concern due to either low response rate or not understanding of the technique.[28 31 33] However, this only impacted a small proportion of the included study population, and the estimates based on low risk of bias studies were similar to those from studies subject to risk of bias. Therefore, we did not rate down the certainty for risk of bias for the body of evidence assessment. The estimates in the included studies were similar. The pooled estimate was based on a large sample size for estimating the single value, and the confidence interval was narrow. So we did not rate down for inconsistency or imprecision either. The included participants were taking antithrombotic treatment and at the risk of developing stroke or recurrent stroke, and they are the population of interest for answering the research question of "the relative importance of outcomes of interest in decision making for patients with antithrombotic treatment". We were unable to detect anecdotal evidence suggesting publication bias. In sum, the certainty of evidence for this assessment is high, and we are very confident that most people find severe stroke has a large impact on lives and the severe stroke is a critical outcome to consider in decision-making.

The major challenge of rating the evidence of relative importance of outcomes is the diversity of research that exists in this field. Due to this diversity, evidence

synthesis and, particularly, meta-analyses are uncommon. While the former is a challenge to conduct any rating of the evidence, the latter is more problematic for rating inconsistency and imprecision. Better standardization of conduct and reporting of studies in this field should alleviate the challenges over time. This will have to be accompanied by further development of systematic review methodology.

We also noticed that those summarizing and presenting evidence may not clearly separate the variability in how patients value main outcomes and mix them with the assessment of inconsistency and imprecision.

Despite all the challenges, we provide an explicit, structured and transparent approach to assess the certainty of the evidence for relative importance of outcomes. Both the expansion of GRADE to this field of evidence and the assessment of a body of evidence in this area in general are innovative. Health researchers, including systematic review authors, assessors of health technologies and guideline developers, will now be in a position to assess the evidence about effects of interventions on outcomes and how important these outcomes are for the target populations. While GRADE will have to elucidate how the overall certainty of evidence is expressed when ratings from intervention effects and relative importance of outcomes are combined we suggest that decision makers consider both when balancing the desirable and undesirable consequences of alternative options. [34 35]

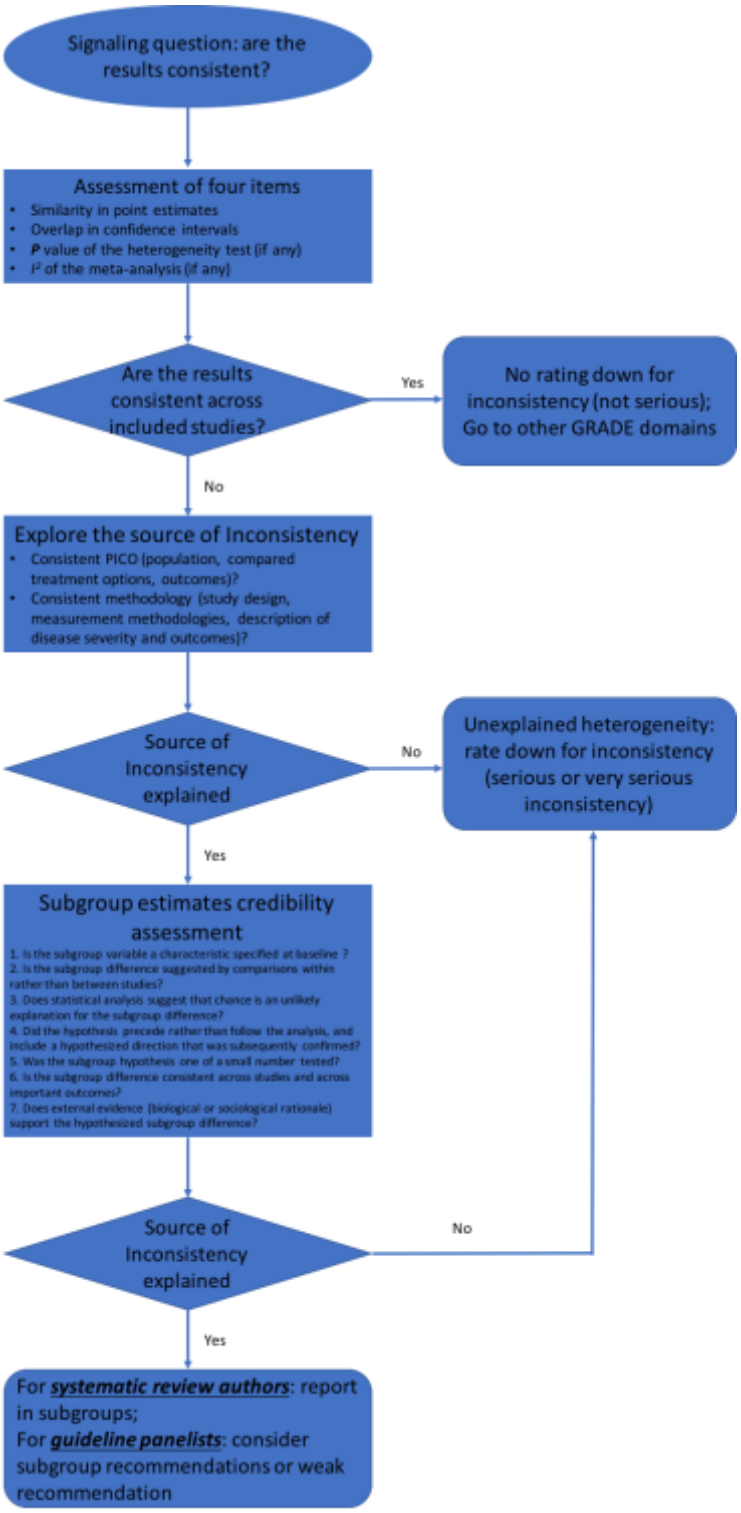


Figure 4.1. Flow chart for assessment of inconsistency

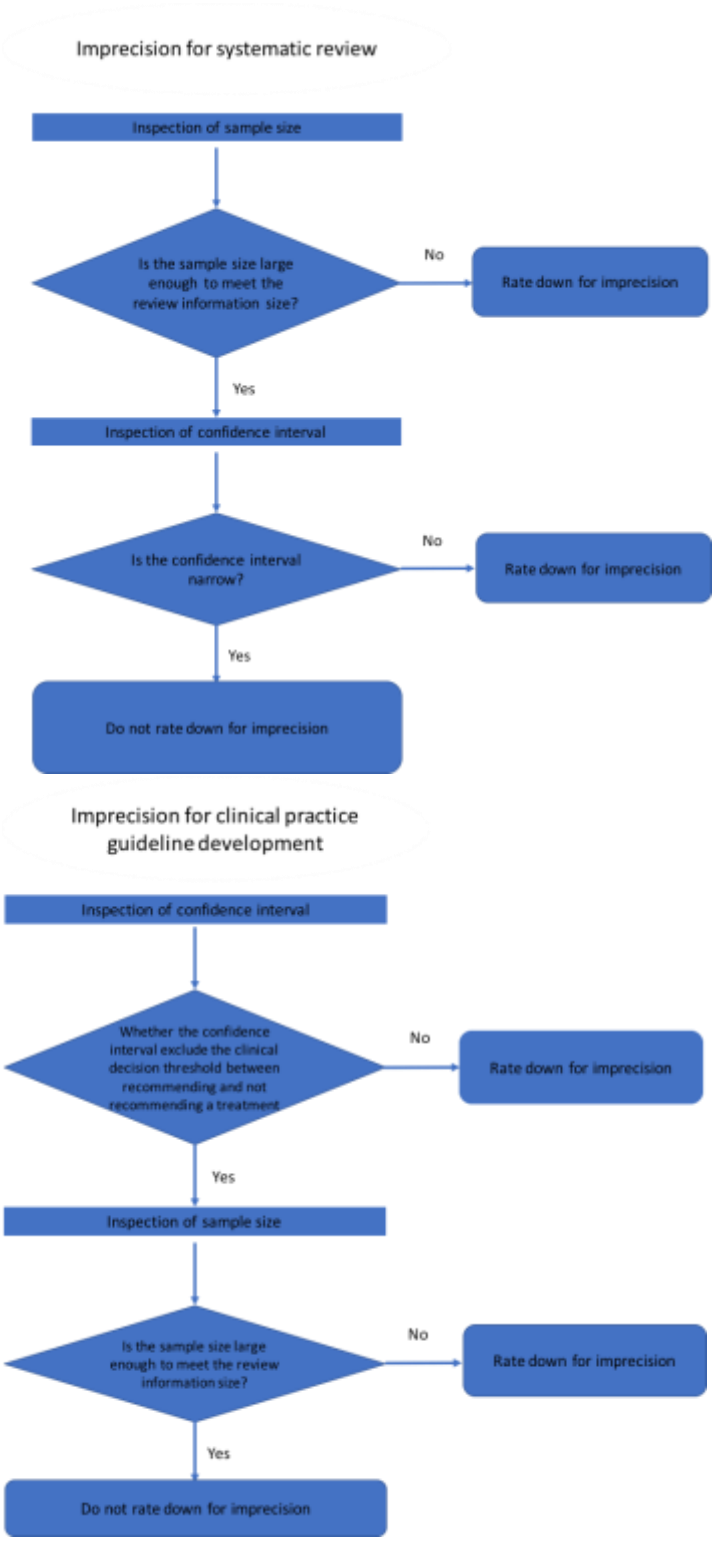


Figure 4.2. Flow chart for assessment of imprecision

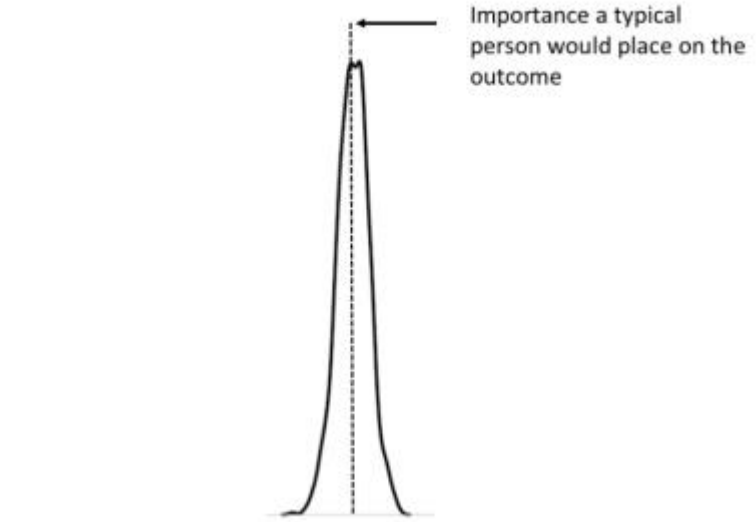


Figure 4.3. Variability: the wide distribution of relative importance of outcome across individuals and/or decision making scenarios

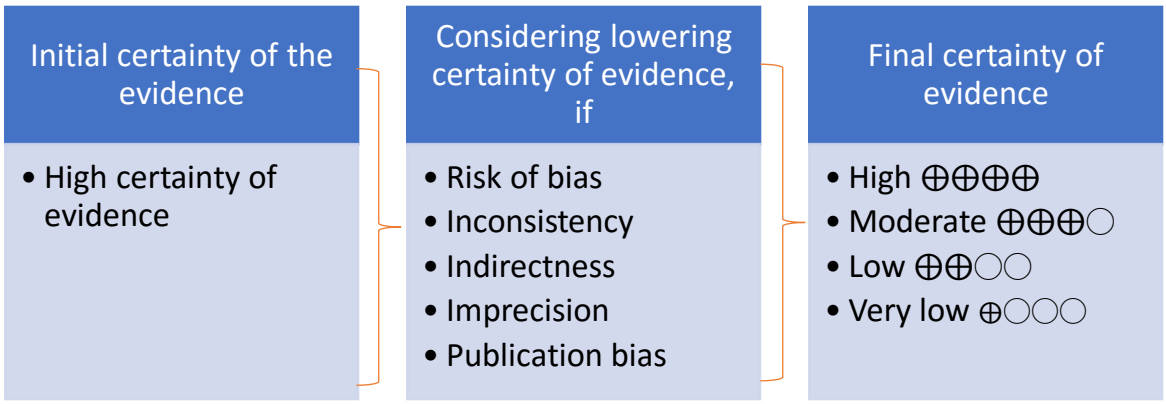


Figure 4.4. The process of GRADE ratings for certainty of evidence for the relative importance of outcomes

Abbreviations

CERQUAL: The Confidence in the Evidence from Reviews of Qualitative research

CI: confidence interval

COPD: Chronic obstructive pulmonary disease

DCE: discrete choice experiment

EBM: evidence-based medicine

EQ-5D: EuroQual-5-dimension (a quality of life measurement tool)

EtD: Evidence to decision

GRADE: Grading of Recommendations Assessment, Development and Evaluation

HUI: health utility index

MeSH: Medical Subject Headings

PICO: population, intervention, comparison and outcome

RIO: relative importance of outcomes

QWB: quality of wellbeing

SF-6D: Short form-6-dimension (a quality of life measurement tool)

Declarations

Ethics approval and consent to participate

Not required. This study does not involve de novo patient data collection. No patient informed consent and Institutional Review Board approval have been sought.

Consent for publication

Not applicable.

Availability of Data and Materials

The datasets supporting the conclusions of this article are included within the article and its additional file.

Authors' contributions

YZ, PA, GG, HJS designed the methodology for this project. YZ, JJYN, and YC summarized the certainty assessment of relevant items in systematic reviews; YZ, PA, GG, and HJS proposed the adapted GRADE domains and subdomains for certainty of evidence assessment; HJS conceived of the project. All authors participated in the methodological discussions. All authors read and approved the final manuscript.

Competing interests

All the authors are member of the GRADE working group. None of the authors have finial conflict of interests.

Acknowledgements

We are grateful to Dr. Amiram Gafni from McMaster University for the comments on the manuscripts.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. It was funded through internal research funds at McMaster University available to HJS.

4.9. Reference

1. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed)* 1996;312(7023):71-2. [published Online First: 1996/01/13]
2. van der Weijden T, Legare F, Boivin A, et al. How to integrate individual patient values and preferences in clinical practice guidelines? A research protocol. *Implementation science : IS* 2010;5:10. doi: 10.1186/1748-5908-5-10 [published Online First: 2010/03/09]
3. Murad MH, Montori VM, Guyatt GH. Incorporating patient preferences in evidence-based medicine. *JAMA : the journal of the American Medical Association* 2008;300(21):2483; author reply 83-4. doi: 10.1001/jama.2008.730 [published Online First: 2008/12/04]
4. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e1S-23S. doi: 10.1378/chest.11-2290 [published Online First: 2012/02/15]
5. Andrews JC, Schunemann HJ, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. *Journal of clinical epidemiology* 2013;66(7):726-35. doi: 10.1016/j.jclinepi.2013.02.003 [published Online First: 2013/04/11]
6. Krahn M, Naglie G. The next step in guideline development: incorporating patient preferences. *JAMA : the journal of the American Medical Association* 2008;300(4):436-8. doi: 10.1001/jama.300.4.436 [published Online First: 2008/07/24]
7. Schunemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014;186(3):E123-42. doi: 10.1503/cmaj.131237 [published Online First: 2013/12/18]
8. Zhang Y, P A-C, Yepes-Nuñez J, J., et al. GRADE guidance for rating the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 1. Risk of bias and indirectness. In: University M, ed., 2017.
9. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]

10. Joy SM, Little E, Maruthur NM, et al. Patient preferences for the treatment of type 2 diabetes: A scoping review. *PharmacoEconomics* 2013;31(10):877-92. doi: http://dx.doi.org/10.1007/s40273-013-0089-7
11. Torrance GW. Preferences for health states: a review of measurement methods. *Mead Johnson Symposium on Perinatal and Developmental Medicine* 1982(20):37-45. [published Online First: 1982/01/01]
12. Ryan M, Scott DA, Reeves C, et al. Eliciting public preferences for healthcare: A systematic review of techniques. *Health Technology Assessment* 2001;5(5):iii-v+1-177.
13. Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization. *Patient Education and Counseling* 2010;78(1):12-23. doi: http://dx.doi.org/10.1016/j.pec.2009.05.011
14. Lewin S, Glenton C, Munthe-Kaas H, et al. Using qualitative evidence in decision making for health and social interventions: an approach to assess confidence in findings from qualitative evidence syntheses (GRADE-CERQual). *PLoS medicine* 2015;12(10)
15. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. *Journal of clinical epidemiology* 2011;64(12):1294-302. doi: 10.1016/j.jclinepi.2011.03.017 [published Online First: 2011/08/02]
16. Rucker G, Schwarzer G, Carpenter JR, et al. Undue reliance on I(2) in assessing heterogeneity may mislead. *BMC medical research methodology* 2008;8:79. doi: 10.1186/1471-2288-8-79 [published Online First: 2008/11/28]
17. Sun X, Briel M, Walter SD, et al. Is a subgroup effect believable? Updating criteria to evaluate the credibility of subgroup analyses. *BMJ (Clinical research ed)* 2010;340:c117. doi: 10.1136/bmj.c117 [published Online First: 2010/04/01]
18. Umar N, Yamamoto S, Loerbroks A, et al. Elicitation and use of patients' preferences in the treatment of psoriasis: A systematic review. *Acta Dermato-Venereologica* 2012;92(4):341-46. doi: http://dx.doi.org/10.2340/00015555-1304
19. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines 6. Rating the quality of evidence--imprecision. *Journal of clinical epidemiology* 2011;64(12):1283-93. doi: 10.1016/j.jclinepi.2011.01.012 [published Online First: 2011/08/16]
20. Schunemann HJ. Interpreting GRADE's levels of certainty or quality of the evidence: GRADE for statisticians, considering review information size or less emphasis on imprecision? *Journal of clinical epidemiology* 2016;75:6-15. doi: 10.1016/j.jclinepi.2016.03.018 [published Online First: 2016/04/12]
21. Jaeschke R, Guyatt GH, Dellinger P, et al. Use of GRADE grid to reach decisions on clinical practice guidelines when consensus is elusive. *BMJ*

(Clinical research ed) 2008;337:a744. doi: 10.1136/bmj.a744 [published Online First: 2008/08/02]
22. Guyatt GH, Norris SL, Schulman S, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):53s-70s. doi: 10.1378/chest.11-2288 [published Online First: 2012/02/15]
23. Komatsuzaki Y, Gramegna P, Stephens JM, et al. Preferences and Utilities of Health Outcomes and Treatments Associated with Head and Neck Cancer. *American Journal of Cancer* 2006;5(1):27-34. doi: 10.2165/00024669-200605010-00004
24. Guyatt GH, Oxman AD, Montori V, et al. GRADE guidelines: 5. Rating the quality of evidence--publication bias. *Journal of clinical epidemiology* 2011;64(12):1277-82. doi: 10.1016/j.jclinepi.2011.01.011 [published Online First: 2011/08/02]
25. Guyatt GH, Oxman AD, Sultan S, et al. GRADE guidelines: 9. Rating up the quality of evidence. *Journal of clinical epidemiology* 2011;64(12):1311-6. doi: 10.1016/j.jclinepi.2011.06.004 [published Online First: 2011/08/02]
26. Schunemann HJ, Griffith L, Jaeschke R, et al. Evaluation of the minimal important difference for the feeling thermometer and the St. George's Respiratory Questionnaire in patients with chronic airflow obstruction. *Journal of clinical epidemiology* 2003;56(12):1170-6. [published Online First: 2003/12/19]
27. Alonso-Coello P, Montori VM, Diaz MG, et al. Values and preferences for oral antithrombotic therapy in patients with atrial fibrillation: physician and patient perspectives. *Health expectations : an international journal of public participation in health care and health policy* 2015;18(6):2318-27. doi: 10.1111/hex.12201 [published Online First: 2014/05/13]
28. Gage BF, Cardinalli AB, Albers GW, et al. Cost-effectiveness of warfarin and aspirin for prophylaxis of stroke in patients with nonvalvular atrial fibrillation. *JAMA : the journal of the American Medical Association* 1995;274(23):1839-45. [published Online First: 1995/12/20]
29. Gage BF, Cardinalli AB, Owens DK. The effect of stroke and stroke prophylaxis with aspirin or warfarin on quality of life. *Archives of internal medicine* 1996;156(16):1829-36. [published Online First: 1996/09/09]
30. Man-Son-Hing M, Laupacis A, O'Connor AM, et al. Patient preference-based treatment thresholds and recommendations: a comparison of decision-analytic modeling with the probability-tradeoff technique. *Medical decision making : an international journal of the Society for Medical Decision Making* 2000;20(4):394-403. doi: 10.1177/0272989x0002000403 [published Online First: 2000/11/04]
31. Protheroe J, Fahey T, Montgomery AA, et al. The impact of patients' preferences on the treatment of atrial fibrillation: observational study of patient based decision analysis. *BMJ (Clinical research ed)* 2000;320(7246):1380-4. [published Online First: 2000/05/20]

32. Slot KB, Berge E. Thrombolytic treatment for stroke: patient preferences for treatment, information, and involvement. *Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association* 2009;18(1):17-22. doi: 10.1016/j.jstrokecerebrovasdis.2008.06.009 [published Online First: 2008/12/27]
33. Thomson R, Parkin D, Eccles M, et al. Decision analysis and guidelines for anticoagulant therapy to prevent stroke in patients with atrial fibrillation. *Lancet (London, England)* 2000;355(9208):956-62. doi: 10.1016/s0140-6736(00)90012-6 [published Online First: 2000/04/18]
34. Alonso-Coello P, Oxman AD, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. *BMJ (Clinical research ed)* 2016;353:i2089. doi: 10.1136/bmj.i2089
35. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical research ed)* 2016;353:i2016. doi: 10.1136/bmj.i2016

Chapter 5. Relative importance of outcomes (values and preferences) for Chronic Obstructive Pulmonary Patients: A systematic review

Yuan Zhang,[1] Rebecca Morgan,[1] Pablo Alonso-Coello,[1, 2] Anna Selva,[3,4] Housne Ara Begum,[1] Gian Paolo Morgano,[1] Wojtek Wiercioch,[1] Małgorzata M. Bała,[5] Rafał R. Jaeschke,[6] Marcin Waligóra,[7] Krzysztof Stycze ń,[6] Arnav Agarwal,[1, 8] Matthew Ventresca,[1] Lídia Blanco-Silvente,[9] Janna-Lina Kerth,[10] Mengxiao Wang,[1] Yuqing Zhang,[1] Saiprasad Narsingam,[11] Yutong Fei,[12] Gordon Guyatt,[1] Holger J. Schünemann[1]

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada
2. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
3. Clinical Epidemiology and Cancer Screening, Corporació Sanitària Parc Taulí, Sabadell, Spain.
4. Research Network on Health Services in Chronic Diseases (REDISSEC), Spain
5. Department of Hygiene and Dietetics, Faculty of Medicine, Jagiellonian University Medical College, Poland
6. Section of Affective Disorders, Department of Psychiatry, Jagiellonian University Medical College, Kraków, Poland
7. REMEDY, Research Ethics in Medicine Study Group, Department of Philosophy and Bioethics, Jagiellonian University Medical College, Krakow, Poland
8. School of Medicine, University of Toronto, Toronto, Canada
9. TransLab Research Group, Department of Medical Sciences, University of Girona, Spain
10. Dept. for Medical Didactics and Curricular Development, Medical Faculty RWTH Aachen University, Germany
11. Instructor, Department of Medicine, Dartmouth Medical School, Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, USA
12. Centre for Evidence-Based Chinese Medicine, Beijing University of Chinese Medicine. Beijing, China

Corresponding author:
Holger J Schünemann, MD, PhD, MSc, FRCP(C)
Chair, Department of Health Research Methods, Evidence, and Impact (formerly "Clinical Epidemiology and Biostatistics")
McMaster University, 1280 Main Street W, Hamilton, ON L8S 4K1
Tel: +1 905 525 9140 x 24931
Email: schuneh@mcmaster.ca

Abstract

OBJECTIVE

To systematically summarize information related to how patients value COPD outcomes and to discuss the methodological challenges in conducting such systematic reviews.

DESIGN

Systematic review.

DATA SOURCES

PubMed, Embase, PsycInfo, and CINAHL.

REVIEW METHODS

Eligible reports assessed relative outcome importance (values and preferences) in COPD patients and used one or more of the following measurement instruments: utilities and health state values, forced choice, probabilistic trade off, discrete choice exercise, willingness to pay, preference trials, other structured quantitative questionnaires or instruments for relative importance of outcome. Two authors independently determined the eligibility of studies through title and abstract screening and, subsequently full text screening, abstracted the eligible studies and assessed the risk of bias. We narratively summarized eligible studies and meta-analyzed the utilities of the same outcomes, or proportions of the same choices. We *a priori* set the disease severity following GOLD criteria as a subgroup factor to consider. We assessed the certainty of evidence using the developed GRADE approach.

RESULTS

Our review identified 170 quantitative studies that reported patient relative importance of COPD related outcomes. Investigators most commonly used quantitative approaches, including direct and indirect utility measurement of outcomes (101 studies), discrete choice exercise, probability trade-off, and forced choice techniques. We identified information for the outcomes of exacerbation, hospitalization, adverse events, intubation, COPD with different severities, and symptom relief. Patients rated both exacerbation of COPD and hospitalization as very important. Patients also rated adverse events as important, but on average less important than symptom relief. The proportion of patients willing to accept mechanical ventilation ranged from 26% to 77%. Although we observed heterogeneity for the utility of COPD states across studies, we also identified a gradient of disutility with severity.

CONCLUSION

Hospitalization and exacerbation are the outcomes that COPD patients rate as most important. We showed the applicability of systematic review methodology as a potential strategy for summarizing evidence how patients value outcomes.

SYSTEMATIC REVIEW REGISTRATION

CRD42015015206

5.1. Background

Considering patient values and preferences regarding the benefits and harms of a health intervention is essential for evidence-based healthcare decision-making.[1] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group has recently operationalized patient values and preferences as "the relative importance patients place on outcomes". [2 3] Systematic reviews about this type of evidence should provide the best available evidence from individual studies to inform shared decision-making. [4-6] Although such reviews could potentially provide very important information, [5 7 8] studies reporting on this aspect vary in their methods and perspectives.[9 10] The methods for conducting systematic reviews in this field, including literature search strategies, eligibility criteria and evidence synthesis remain largely unaddressed.

Chronic Obstructive Pulmonary Disease (COPD) is a global problem and is widely studied. COPD is a progressive lung disease characterized by chronic poor airflow.[11 12] Symptoms of COPD include coughing, wheezing, shortness of breath, chest tightness; these symptoms, and the resulting functional limitations, greatly impact patients' quality of life. [13 14] Using a variety of measurement methodologies, numerous original studies have addressed how patients value COPD outcomes. However, no comprehensive systematic review has summarized the available evidence. Considering the disease burden of COPD,[12] such a review would inform decision-making for a large patient community globally.

We therefore systematically addressed the question "what is the relative importance patients place on chronic obstructive pulmonary disease related

outcomes?"[2 3] Given the lack of guidance on how to summarize research evidence about the relative importance of outcomes, we also discuss the methodological challenges of systematic reviews in this field using COPD as an example.

5.2. Methods

5.2.1. Protocol and registration

We conducted the systematic review of the literature in accordance with the Preferred Reporting in Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[15] The review protocol is registered on PROSPERO (registration number: CRD42015015206).

5.2.2. Eligibility criteria

We included studies that reported patient values and preferences of COPD patients. We set no limits on the type of study design, language, or treatments of COPD. Given that patients' preference for or against an intervention is conceptually equivalent to the importance people place on the outcomes that follow from the decision to undergo an intervention, we included studies of COPD patients' preferences for alternative interventions. Thus, studies with the following characteristics were eligible as methods for reporting relative importance of outcomes:

1. **Patient utility and health state value studies:** Studies that examine how patients value alternative health states and experiences with treatment. The eligible measurement techniques used are: standard gamble, time trade off, visual analogue scale, or mapping results based on either generic (EuroQol-5D, or SF-

36)[9] or specific measurement (i.e. Chronic Respiratory Questionnaire) of health related quality of life.

2. **Direct choice studies:** Studies that examine patients' choice when they were presented with a description of hypothetical states or during decision making for their own actual health states (i.e., forced choice when presented with decision aid, probabilistic trade off techniques, discrete choice, willingness to pay, RCTs for preferences, etc.).
3. **Non-utility measurement of health states studies:** Studies that quantitatively examine the patients' view, attitude or preferences on outcome importance through questionnaires or instruments that are not utility studies.

We only included quantitative studies reporting COPD as comorbidity if they reported COPD relative importance of outcomes information separately. We excluded non-original studies such as clinical practice guidelines, reviews, commentaries, communications, letters, or viewpoints; we also excluded case reports, case series, and health economic evaluation studies without original utility elicitation. Qualitative studies that explore patients' views, attitudes or preferences related to different treatment options using focus groups and individual interviews will be excluded from this review but included and reported in another review.

5.2.3. Information sources

We searched Medline (through PubMed), Embase, PsycInfo, and CINAHL from inception date to January 2015 with an extensive search strategy developed for retrieving this type of evidence (see Appendix 5.1),[16] conference abstracts and reference lists of identified studies.

5.2.4. Study selection

Two authors independently determined the eligibility of studies by review of abstracts and, for studies judged potentially relevant on abstract review, through screening of full text articles with a standardized and piloted abstraction form. We resolved disagreement by discussion or through third party adjudication.

5.2.5. Data collection and items

Two authors independently recorded data: principle author, publication year, participant demographics (sample size, age, sex, etc.), survey techniques or methodologies used, relative importance of outcome results, and existing risk of bias assessments.

5.2.6. Risk of bias in individual and certainty of evidence across studies

Since there is no accepted risk of bias or study quality assessment tool for this type of evidence, we used the GRADE approach to rate certainty of values and preference evidence that we developed, validated and reported in a separate project, which includes a risk of bias proposal.[17] The key items to assess the risk of bias included sample selection, response rate (or attrition rate if follow-ups involved), choice and administration of the instrument, outcome (or health state) presentation, and understanding of the methodology and data analysis (if applicable). We assessed the certainty of evidence using the GRADE approach and classified it as high, moderate, low, and very low based on the risk of bias and the additional domains including inconsistency, imprecision, indirectness and publication bias.

5.2.7. Data analysis

Based on the severity of airflow limitation, the Global Initiative for Chronic Obstructive Lung Disease (GOLD) categorizes COPD into four severity levels.[14] We *a priori* set the disease severity following GOLD criteria as the subgroup factor to consider. We aimed to perform subgroup analysis according to different clinical stages and obtain patient values for each particular stage. Information on relative importance of outcomes exists in a variety of formats, including the utility of outcomes or disease stages, proportion of choice, rankings or scores on a scale. For the sake of simplicity, we report all estimates using the descriptive term "utility" although many studies did not fulfill methodological requirements for true utilities.[18] We narratively summarized eligible studies and meta-analyzed utility estimates or proportions of the same choices for the same outcomes. To pool estimates we used the random-effects inverse variance method utilizing Stata 11.0.[19]

5.3. Results

5.3.1. Study selection

We identified 42,993 records. After excluding duplicates, we screened 33,601 titles and abstracts and retrieved 2,805 articles for full text screening. We included 170 quantitative studies eligible reporting patient values and preferences on COPD related outcomes in the systematic review (See Figure 5.1. Flow Diagram).

5.3.2. Study characteristics

102 of 170 studies reported utilities for COPD outcomes. Of these 102 utility or health state values, 53 utilized feeling thermometer or visual analogue scale (VAS) including the EQ-5D VAS, 8 utilized standard gamble (SG), and 6 time trade-off (TTO). For indirect measurements, 55 studies reported EQ-5D utilities, 10 SF-6D utilities, 7 health utility index (HUI), 7 15D, and 3 quality of well-being (QWB) utilities. Of 51 direct choice studies, 37 used forced choice techniques, eight discrete choice exercise/conjoint analysis or willingness to pay, four probability trade-off, and two ranking methods (see appendix table 5.1).

Regarding the study design, 101 were cross-sectional studies, 16 cohort studies, 11 repeated surveys, 36 randomized controlled trials and 6 quasi-randomised trials. The studied outcomes typically were exacerbation, hospitalization, adverse events, intubation or mechanical ventilation, different severities of COPD and symptom relief (see Table 5.1 and appendix table 5.1). Despite the large number of included studies, few studies reported the relative importance of outcome information on the same outcomes. Meta-analyses were restricted to studies focusing on different COPD severities measured with VAS and EQ-5D utility. Overall, we were unable to clearly detect publication bias.

5.3.3. Risk of bias within included studies

Appendix Table 5.2 summarizes the risk of bias assessment. Major issues arose from the validity and reliability of the measurement tools: we classified 56 studies asking directly the choice over a set of options as "serious risk of bias" in this item. For the sampling frame, we classified nine studies with convenience

sampling strategy or recruiting a volunteer sample as "serious risk of bias". For the response rate, 30 had response rates of less than 50% and were classified as "serious risk of bias". For other domains, we mostly classified studies as "low risk of bias".

5.3.4. Importance of exacerbation

The measurements used to elicit the importance of exacerbations include visual analogue scale (including EQ-5D VAS) in four studies,[20-23] time trade off in one study,[23] and EQ-5D utility in three studies.[20-22] The estimates vary across included studies, from 0.259 to 0.466 with VAS measurement, and 0.43 to 0.683 with EQ-5D utility. We conducted meta-analysis using the inverse variance method to pool the estimates based on VAS and EQ-5D, yielding utility of exacerbation of 0.377 (95% CI: 0.294-0.461, $I^2 = 97.4\%$, $P < 0.001$ for the heterogeneity test) on a 0-1 visual analogue scale, and 0.525 (95% CI: 0.434-0.615, $I^2 = 95.5\%$, $P < 0.001$ for the heterogeneity test) with the EQ-5D utility. For both pooled estimates, the difference in the point estimates, I^2 and the statistical test suggested potential heterogeneity. We could not explain this inconsistency and rated down the certainty of evidence to moderate (see Table 5.1). For studies that used the EQ-5D utility measurement, we further rated down for indirectness given the indirect measurement tool used. So the certainty of evidence is moderate for VAS measurement, and low for the EQ-5D measurement, respectively. Rutten van Molken and colleagues suggested that disutility is related to the number and severity of exacerbations. They reported the disutility related to one non-serious exacerbation, two non-serious exacerbations, one serious exacerbation and one

non-serious and one serious exacerbation as 0.037, 0.068, 0.090 and 0.130 according to visual analogue scale, and 0.010, 0.021, 0.042 and 0.088 according to time trade off, respectively (see Table 5.2).[23] The certainty of evidence is high.

5.3.5. Importance of hospitalization

Three studies separately reported the utilities of hospitalized COPD patients on the visual analogue scale ranging from 0.259 to 0.551,[24-26] with the pooled estimates of 0.363 (95% CI: 0.161-0.565, $I^2 = 98.9\%$, $P < 0.001$ for the heterogeneity test) while one report using the EQ-5D utility suggested hospitalized patients compared to non-hospitalized suffer a utility decrease of 0.077 (see Table 5.3).[26] We rated down the certainty of evidence due to unexplained inconsistency across the included studies. The certainty of evidence is moderate.

5.3.6. Intubation and mechanical ventilation

One report based visual analogue scale suggested the mean value of intubated COPD patients was 0.572, with the standard deviation of 0.182.[27] We rated down the certainty of evidence for risk of bias due to low response rate and the overall certainty is moderate. We also included studies on forced choice experiments evaluating whether or not to accept mechanical ventilation. Across 12 studies, the proportion of people willing to accept mechanical ventilation ranged from 26% to 77% (see Table 5.4).[28-39] One study reported the willingness to accept life-sustaining ventilation would decrease as the disease burden increases. Of these studies, two studies were on decision aids, both suggesting positive effects of decision aids on the decision-making process.[30 39] These studies asked participants

to choose from a set of options, but the reliability and validity of the measurement was deemed unclear. Other than risk of bias we identified very serious heterogeneity across the included studies. We rated down one level for risk of bias and two levels for very serious inconsistency, so the certainty of evidence is very low.

5.3.7. Adverse events

Table 5.5 summarized the results related to the importance of adverse events. One of the two included discrete choice studies compared the possibility of adverse effects with the extent to which treatment seems to relieve symptoms, the extent to which the doctor gives sufficient time to listen to the patient, costs of treatment, the extent to which the patient sees the same doctor each time, and extent to which the doctor treats the patient as an entire person.[40] The extent of symptom relief was deemed to be more important than adverse effects, but the possibility of adverse effects more important than the other outcomes. Another discrete choice study suggested symptom relief to be the most important outcome, while the possibility of adverse events was considered more important than the timing and use of (rescue) medicine use.[41] The latter study was an online voluntary study in 515 participants which we rated down for selection bias and limited validity of the instrument. The overall certainty is moderate.

In one of the two forced choice studies, a cross-sectional study, 38% of 1100 participants chose fewer side effects as ideal characteristics of COPD therapy, which was less important than quick symptom relief and longer intervals between flare-ups, but more important than better ability to cope with daily chores, lower

costs of treatment, and better doses.[42] The second cross-sectional study suggested 12% of participants chose side effects as main negatives of nebulization treatment.[43] We rated down the certainty of evidence due to risk of bias stemming from the limited reliability and validity of the measurement tool. The certainty of evidence is moderate.

5.3.8. Utility of COPD

Most studies addressing the utility of the experience of COPD itself were based on EQ-5D, HUI and 15D. The estimates range from 0.465 to 0.89 for the EQ-5D, 0.520 to 0.939 for the HUI, and 0.730 to 0.810 for the 15D (see Table 5.6). For direct measurements, the utilities ranged from 0.44 to 0.706 for the visual analogue scale, 0.550 to 0.910 for the time trade off, and 0.550 to 0.950 for the standard gamble.

Based on the EQ-5D, utilities reported across different GOLD (The Global Initiative for Chronic Obstructive Lung Disease) stages,[14] we observed a gradient of disutility as the disease progresses. The pooled estimates for EQ-5D measurements of mild, moderate, severe, and very severe COPD are 0.821 (95% CI: 0.814-0.828 , $I^2 = 72.7\%$, $P < 0.001$ for the heterogeneity test),[23 44-53] 0.760 (95% CI: 0.756-0.765, $I^2 = 98.9\%$, $P < 0.001$ for the heterogeneity test),[23 44-48 50-56] 0.727 (95% CI: 0.722-0.732, $I^2 = 98.5\%$, $P < 0.001$ for the heterogeneity test), [23 44-48 50-58] and 0.681 (95% CI: 0.675-0.686, $I^2 = 97.1\%$, $P < 0.001$ for the heterogeneity test) (See Figure 5.2).[23 44 45 47 48 50 52-56] We summarized the median of EQ-5D utility estimates from the studies, and the medians and interquartile ranges (IQR) were 0.83 (IQR: 0.805-0.845), 0.73 (IQR: 0.68-0.81), 0.73 (IQR:

0.653-0.782) and 0.623 (IQR: 0.565-0.72), respectively. Studies reporting utilities according to other criteria, such as BODE index, quartile of FEV1 predicted, and less than or more than 50% FEV1 predicted, also suggested utilities decreases associated with greater impairment of lung function.[26 52 56 59 60] Across different severities, we observe the statistical heterogeneity even when the same measurement tool was used. Consequently, we rated down the certainty of evidence for these utilities due to unexplained inconsistency and for indirectness of the measurement tool (EQ-5D). So the certainty of evidence is low for the EQ-5D utility measurements for very severe, severe, moderate and mild COPD.

5.3.9. Symptom relief

Eligible studies addressed the extent of symptom relief and speed of symptom relief. For extent of symptom relief, we identified two discrete choice studies, and both suggested the extent of symptom relief more important than adverse effects, the doctor giving sufficient time to listen to the patient, costs of treatment, seeing the same doctor each time, being treated as an entire person, onset time of medication, ease of medication use, and use of rescue medication.[40 41] We rated down for the serious risk of bias because a large proportion of the study population were recruited by online survey, the eligibility of the participants and the validity of the answers were in question; thus, the certainty of evidence is moderate. Three other forced choice studies corroborated this result: in a survey addressing expectation of treatment, 82.3% of the respondents chose greater symptomatic relief as the most important outcome.[61] In another survey, extent of symptom relief was considered important, only secondary to "not to be kept alive

on life support when there is little hope for a meaningful recovery".[62] A third survey reported more than half of the participants with end stage COPD would prefer treatment focusing on relieving pain and discomfort rather than extending life.[29] Because of lacking of validity for the instruments in these surveys, we rated down the certainty of evidence for risk of bias, and the certainty is moderate. Studies were variable in reports of the importance of speed of symptom relief. Three studies suggested that although speed of symptom relief was not more important than extent of symptom relief, it was more important compared with impact on mood, and sleep quality.[41 42 63] In contrast, two studies suggested speed of symptom relief was the least important outcome.[61 64] We rated down for the serious risk of bias because of the problematic online survey. Because the importance of symptom relief speed could not be certainly determined due to unexplained differences, we further rated down for inconsistency. The certainty of evidence is low.

5.3.10. Forced choice and Preference trials

We identified 16 trials evaluating the preference for different interventions. [39 65-79] Of these 16 trials, eight examined the preferences of inhalers,[65 66 72 75-79] four different pharmacotherapies,[67-69 71] one mechanical ventilation,[39] two the place of treatment, [73 74] and one for the model of COPD care.[70] Of the 16 trials, 14 asked participants to choose between accepting the options or not; two asked participants to rank the preferences of four different inhalers.[77 79]

The identified studies also evaluated the importance of other outcomes, including utility associated with FEV1 predicted,[26 52 56 59 60] ease of use of inhalers,[75 76 79-81] and sleep quality[63 64 82] (see appendix 5.3).

5.4. Discussion

5.4.1. Main findings

We have conducted the most comprehensive systematic review of how COPD patients value outcomes with 170 individual studies. The studies were highly variable in their designs, measurement instruments used, and outcomes addressed. Patients rated intubation, exacerbation of COPD and COPD needing hospitalization as very important. Willingness to accept mechanical ventilation varied greatly. Studies, primarily using the EQ-5D, consistently reported that the utility associated with living with COPD decreases as disease progresses. Patients considered symptom relief important, and more important than adverse events from treatment.

5.4.2. Strengths and limitations of the study

Our study has several strengths. Our literature search using a sensitive search filter for patient values and preferences studies likely identified the vast majority of published relevant studies. Our pre-specified eligibility criteria, in accordance with the definition of "relative importance of outcome", proofed useful for identifying examples of different types of outcome descriptions and methods for their assessment. We rated the certainty in the evidence by applying, for the first

time, GRADE specific guidance to assess the certainty of evidence for the relative importance of outcomes.[83]

Our study has some limitations. First, because of variability in measurement instruments, and the paucity of evidence based on standard gamble and time trade-off, we were only able to conduct meta-analysis across severity levels of EQ-5D utility and EQ-5D VAS measurements.

Second, we identified a relatively small number of discrete choice and probability trade-off studies. These studies could provide information on the threshold for a change in a decision.[33] However, both the probability trade-off and the discrete choice exercise have the merit of allowing "to customize" the methodology according to the study objectives. Consequently, the studies included reported a variety of attributes and different levels of the same attributes studied.[84-86] This created obstacles for evidence synthesis and interpretation.

In addition, we were unable to pre-specify all the outcomes relevant to COPD decision-making and set the review scope on these outcomes. Lastly, given the lack of empirical knowledge in what manner and to what extent publication bias may affect our systematic review results, our assessment of publication bias is limited.

5.4.3. Context to other studies

Several aspects distinguish our work from previous published literature reviews.[87-91] Our work is more comprehensive than the work Moayeri and colleagues who evaluated EQ-5D utilities of COPD stages, though the pooled EQ-5D utilities proved similar.[87] Two reviews included only multi-attribute utility results. [88 91]

Brooker and colleagues identified ten studies on patient preferences for mechanical ventilation in COPD, most of them cross-sectional surveys with forced choice questions. [91] Our work yielded more studies because of the broad definition focusing on the importance of outcomes and including all types of relevant studies and measurement tools.

A second aspect in which our work differs is the critical assessment, both on the individual study level and on the body of evidence level with the GRADE approach. Without this critical assessment, not performed in detail in any of the other reviews, assessment of the confidence one should place on the results is very limited.[92]

5.4.4. Implications of the study

The GRADE working group has recently operationalized "values and preferences" as how patients value outcomes (relative importance of outcome). This has been reflected for example in the Evidence to Decision (EtD) framework as an individual criterion ("Is there important uncertainty and/or variability in how patients [or those affected by the decision] value the main outcomes?").[2 3] The underlying rationale is that individuals make decisions by explicitly or implicitly considering the consequences this decision would incur. This conceptualization does not exclude the preferences for different management strategies; rather this definition is a way to understand the preferred and non-preferred strategies from the consequences the strategies will incur. Considering the latter, we also included forced choice studies; most addressing whether the patients would accept an intervention, or which treatment the patients would prefer.

Given the breadth of findings, this systematic review provides empirical evidence to support using "relative importance of outcome" to inform values and preferences. We included a number of different approaches to measure utilities. For example, in the included studies participants expressed their preferences or importance they place on outcomes through utilities or health state values. Researchers can use discrete choice and probability trade-off techniques to manipulate the level of attributes, observing the choice participants would make as these attributes vary, and quantifying the importance of these attributes. They could also simply ask participants to rank a set of options.

5.4.5. Unanswered questions and future studies

Although we used a risk of bias tool to assess the quality of included studies, we are uncertain the weight each factor should bear within and across study level. For considering risk of bias, one concern is the merits of measurement tools involving a valuation of hypothetical scenarios in relation to measurements of an actual outcome that participants experience. If the participants are valuing a prescribed outcome, it is likely they are valuing that same outcome. But if participants were valuing the outcome they are experiencing or they have experienced, the underlying outcome is actually different across participants, varying because of disease severity or functions affected. In this case, variability around the relative importance of outcome, expressed as utility or disutility, may be a result of true variability for the valuation on the same outcome or because the participants are inherently valuing different outcomes. Additionally, we have not completed the validation of our risk of bias instrument for utility studies. Further studies are

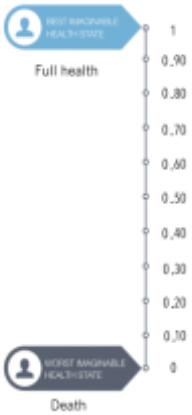
necessary to validate the search strategy for this type of studies. This sensitive but not specific search strategy leads to a large number of hits, which could be a burden for systematic review authors and guideline panelists. Researchers could also explore the possibility of using machine learning models to aid the systematic review process and improve efficiency.

5.5. Conclusions

Our systematic review showed that patients value the outcomes of hospitalization and COPD exacerbation as very important. The willingness to accept a treatment decreases as the disease burden increases. We observed large variability in the utility associated with COPD severity, but on a population level. We identified a gradient of disutility as the disease progresses. Quantitative approaches, including direct and indirect utility measurement of outcomes, discrete choice exercise, probability trade-off, and forced choice are the predominant measurement instruments to address the importance patients place on relevant outcomes. Although further studies necessary to explore the unsolved methodological questions to synthesize the evidence, through this systematic review process, we also showed the usefulness of systematic reviews as a potential strategy for summarizing evidence in this field, and inform decision makers, both in the context of health technology assessments and guidelines.

Table 5.1. Summary of finding table

Question: What are the views about the relative value/importance of outcomes of interest in decision making for patients with chronic obstructive pulmonary disease?



*Utilities represent the value individuals place on different outcomes. They are measured on an interval scale, with zero reflecting states of health equivalent to death/worst imaginable health and one (or 100 in some cases) reflecting perfect health/ best imaginable health.

Health state/Outcome (Categories of values and preferences)	**Estimates of outcome importance** (range across studies /pooled mean, 95% CI)	**No. of participants /studies**	**Certainty of evidence**	**Interpretation of findings**
Exacerbation (Utility* measured with visual analogue scale [1])	range across studies: 0.259-0.466/ pooled mean: 0.377 (95% CI: 0.294, 0.461) [2]	1076 participants/ 4 studies [2]	⊕⊕⊕◯ Moderate certainty due to inconsistency [2]	Most people find exacerbation of COPD probably has a large impact on lives. There is likely no important variability for this assessment.
Exacerbation (EQ-5D Utility [3])	range across studies 0.43-0.683/ pooled mean: 0.525 (95% CI: 0.434, 0.615) [4]	927 participants/ 3 studies [4]	⊕⊕◯◯ Low certainty due to inconsistency and indirectness [4,5]	Most people find exacerbation of COPD probably has a large impact on lives. There is likely no important variability for this assessment.
Exacerbation (disutility) [6]	Visual analogue scale: One non-serious exacerbation: -0.037 (0.005); Two non-serious exacerbations: -0.068 (0.005); One serious exacerbation: -0.090 (0.007); One non-serious and one serious exacerbation: -0.130 (0.007) Time trade off: One non-serious exacerbation: -0.010 (0.007); Two non-serious exacerbations: -0.021 (0.007); One serious exacerbation: -0.042 (0.009); One non-serious and one serious exacerbation: -0.088 (0.009)	239 participants/ 1 study	⊕⊕⊕⊕ High certainty	Most people find exacerbation of COPD has an impact on lives, which grows larger as the severity of exacerbation progresses. There is likely no important variability for this assessment.

Hospitalization (Utility measured with visual analogue scale) [7]	range across studies: 0.259-0.551/ pooled mean: 0.363 (95% CI: 0.161, 0.565) [8]	356 participants/ 3 studies	⊕⊕⊕◯ Moderate certainty due to inconsistency 8	Most COPD patients find hospitalization probably has a large impact on lives. There is likely no important variability for this assessment.
Intubated (utility measured with visual analogue scale) [9]	Mean (SD): 0.572 (0.182), Median (IQR): 0.55 (0.45, 0.70)	171 participants/ 1 study	⊕⊕⊕◯ Moderate certainty due to risk of bias [9]	Most people find intubation probably has a moderate impact on lives. There is likely no important variability for this assessment.
mechanical ventilation (forced choice) [10]	The proportion of willing to accept the mechanical ventilation ranges from 26% to 77%. Two studies on decision aid suggested decision aid reduced decision conflict or uncertainty.	3470 participants/ 12 studies	⊕◯◯◯ Very low ow certainty due to risk of bias [10] and very serious inconsistency [11]	People seem to prefer to accept mechanical ventilation. There is likely important variability for this assessment.
Adverse event (discrete choice) [12]	Two studies suggested patients consider adverse events as important outcomes. One study suggested adverse events more important than onset time of medicine, ease of use, rescue medicine use; another suggested adverse events more important than costs of treatment, extent to which the patient sees the same doctor each time, and extent to which the doctor treats the patient as an entire person. Both studies concluded symptom relief more important than adverse events.	564 participants/ 2 studies	⊕⊕⊕◯ Moderate certainty due to risk of bias [12]	People probably consider adverse events as important outcome. There is likely no important variability for this assessment
Adverse event (forced choice) [13]	In one cross-sectional study, 38% of the participants chose fewer side effects as ideal characteristics of a COPD therapy. Another cross-sectional study suggested 12% of participants chose side effects as main negatives of nebulization treatment.	1500 participants/ 2 studies	⊕⊕⊕◯ Moderate certainty due to risk of bias [13]	People probably consider adverse events as important outcome. There is likely no important variability for this assessment

GRADE Working Group grades of evidence: here we assess the certainty of evidence on mean outcome importance. We use "certainty of

evidence", "certainty in estimates", "quality of evidence" and "strength of evidence" interchangeably.
High certainty: We are very confident that the true value of outcome importance lies close to that of the estimate.
Moderate certainty: We are moderately confident in the estimate: The true value of outcome importance is likely to be close to the estimate, but there is a possibility that it is substantially different
Low certainty: Our confidence in the estimate is limited: The true value of outcome importance may be substantially different from the estimate
Very low certainty: We have very little confidence in the estimate: The true value of outcome importance is likely to be substantially different from the estimate

CI: Confidence interval; IQR: interquartile range; SD: standard deviation; SG: Standard Gamble; TTO: Time Trade Off;
VAS: Visual Analogue Scale.

1. Four studies including Cross 2010, Goossens 2011, Miravitlles 2011a, and O'Reilly 2007 used EQ-5D visual analogue scale to elicit health state values on exacerbation of COPD.
2. Across four included studies, the point estimates range from 0.259 to 0.466. Using inverse-variance method to pool the estimates, the I^2 (97.4%) and statistical test (<0.001) suggest potential heterogeneity across studies. The difference in study population cannot explain the source of heterogeneity (the participants in the three studies were exacerbation patients, exacerbation patient not needing hospitalization, ambulatory patients, and hospitalized patients due to exacerbation).
3. Three studies including Cross 2010, Goossens 2011, Miravitlles 2011a used EQ-5D utility to elicit the importance of outcome.
4. Across three included studies, the point estimates range from 0.43 to 0.683. Using inverse-variance method to pool the estimates, the I^2 (95.5%) and statistical test (<0.001) suggest potential heterogeneity across studies. The

difference in study population cannot explain the source of heterogeneity (the participants in the three studies were exacerbation patients, exacerbation patient not needing hospitalization, and ambulatory patients).

5. We rated down the quality of evidence for indirectness because indirect measurement tool (EQ-5D) was used to elicit the utility on outcomes.
6. Rutten van Molken 2009 reported the disutility due to the exacerbations. The measurement tools include visual analogue scale and time trade off. The researchers estimated the disutility due to exacerbation using random effects regression analysis.
7. Alcazar 2012, Antoniu 2014, and O'Reilly 2007 measured the importance of hospitalization to participants using EQ-5D visual analogue scale.
8. The point estimates vary from 0.259 to 0.551. Using inverse-variance method to pool the estimates, the I^2 (95.5%) and statistical test (<0.001) suggest potential heterogeneity across studies. The difference in study population cannot explain the source of heterogeneity.
9. Wildman 2009 was a cohort study recruiting consecutive participants, and measures their health state value with visual analogue scale. The study was in risk of bias due to a low response rate from the patients contacted (39.4%).
10. The studies reported participants' preference of ventilation include Chakrabarti 2009, Claessens 2000, Dales 1999, Downey 2013, Gaber 2004, Janssen 2011b, Janssen 2011c, Reinke 2011, Rinnenburger 2012, Stapleton 2005,

Travaline 1995, and Wilson 2005. All the studies directly asked the participants to choose from a set of options, with the reliability and validity of the measurement tools unclear.

11. The preferences of ventilation range from 26% to 77%, and we cannot explain the variation through known factors.
12. Bulcun 2014 compared extent of symptom relief with extent to which the doctor gives sufficient time to listen to the patient, possibility of experiencing adverse effects from treatment, costs of treatment, extent to which the patient sees the same doctor each time, and extent to which the doctor treats the patient as an entire person. Kawata 2014 was an online voluntary survey on the comparison of importance of symptom relief, speed of symptom relief, rescue medicine use, and side effects; with 515 participants recruited, the eligibility of the participants, and their answers are in risk of bias.
13. Miravitlles 2007 and Sharafkhaneh 2013 reported the importance of adverse events in COPD treatment decision making. In Miravitlles 2007, the researchers asked participants to choose the ideal characteristics of a COPD therapy. While in Sharafkhaneh 2013, the participants chose the primary disadvantages of nebulization therapy from a set of options. The reliability and validity of the measurement tools were unclear.

Table 5.2. Utility of exacerbation

Study ID	Instrument	Report format	Results
Cross 2010	EQ-5D VAS	Mean (SD)	Exacerbation of COPD: MCP arm 44.95 (21.03), no MCP arm 46.64 (21.42)
	EQ-5D utility	Mean (SD)	Exacerbation of COPD: MCP arm 0.45 (0.32), no MCP arm 0.43 (0.36)
Goossens 2011	EQ-5D VAS	Mean (SD)	Exacerbation (at enrollment): 36.68 (25.244)
	EQ-5D utility	Mean (SD)	Exacerbation (at enrollment): 0.683 (0.209)
Menn 2010	EQ-5D utility	Mean (SD)	EQ-5D Admission Stage III: 0.62 (0.26)
			EQ-5D Admission Stage IV: 0.60 (0.26)
			EQ-5D Discharge Stage III: 0.84 (0.20)
			EQ-5D Discharge Stage IV: 0.75 (0.22)
	SF-6D utility	Mean (SD)	SF-12-SF-6D Admission Stage III: 0.61 (0.13)
			SF-12-SF-6D Admission Stage IV: 0.54 (0.08)
			SF-12-SF-6D Discharge Stage III: 0.65 (0.12)
			SF-12-SF-6D Discharge Stage IV: 0.58 (0.08)
Miravitlles 2011a	EQ-5D utility	Mean (SD), Range	EQ-5D index baseline (exacerbation): 0.54 (0.23)
			EQ-5D index 1 month follow-up: 0.61 (0.21)
			EQ-5D index Mean change: 0.07 (0.17)
	EQ-5D VAS	Mean (SD), Range	EQ VAS baseline (exacerbation): 34.4 (27.4)
			EQ VAS 1-month follow-up: 41.8 (31.2)
			EQ VAS Mean change: 7.68 (13.8)
Rutten van Molken 2009	VAS	regression coefficients (SEM)	One non-serious exacerbation- -0.037 (0.005); Two non-serious exacerbations - -0.068 (0.005); One serious exacerbation - -0.090 (0.007); One non-serious and one serious exacerbation - -0.130 (0.007)
	TTO		One non-serious exacerbation- -0.010 (0.007); Two non-serious exacerbations - -0.021 (0.007); One serious exacerbation - -0.042 (0.009); One non-serious and one serious exacerbation - -0.088 (0.009)
Solem 2013	EQ-5D utility	Mean (SD)	patients recently experiencing a severe exacerbation： 0.627 (0.210) patients recently experiencing a moderate exacerbation： 0.698 (0.197) patients who had experienced three or more exacerbations in the previous year： 0.638 (0.212) patients who had experienced two exacerbations in the previous year： 0.684 (0.204) patients who had experienced one exacerbation in the previous year： 0.727 (0.175)

			current health (last exacerbation)： 0.552 (0.283) thought back, patients experiencing a severe exacerbation (last exacerbation)： 0.471 (0.313) thought back, patients experiencing a moderate exacerbation (last exacerbation)： 0.595 (0.257) very severe COPD (last exacerbation)： 0.494 (0.312) severe COPD (last exacerbation)： 0.590 (0.256) patients who had experienced three or more exacerbations in the previous year (last exacerbation)： 0.520 (0.282) patients who had experienced two exacerbations in the previous year (last exacerbation)： 0.552 (0.306) patients who had experienced one exacerbation in the previous year (last exacerbation)： 0.610 (0.254)
Torrance 1999	HUI	Mean (SD)	first AECB, for Ciprofloxacn group: 0.72 (0.20), usual care group: 0.68 (0.19)
			remaining AECB (Excluding first AECB), Ciprofloxacn group: 0.74 (0.18), usual care group: 0.69 (0.22)
Punekar 2007	EQ-5D utility	Mean (95% CI)	No exacerbations in Primary care physician setting: 0.78 (0.75-0.8)
			1-2 exacerbations in Primary care physician setting: 0.74 (0.72- 0.77)
			>3 exacerbations in Primary care physician setting: 0.61 (0.59-0.64)
			No exacerbations in respiratory specialist setting: 0.75 (0.72-0.77)
			1-2 exacerbations in respiratory specialist setting: 0.73 (0.71-0.76)
			>3 exacerbations in respiratory specialist setting: 0.57 (0.54-0.60)

Table 5.3. Utility of Hospitalization

First Author	Instrument	Reported format	Hospitalization
Alcazar 2012	EQ-5D VAS	Mean (SD)	Hospitalized patients: 0.5509 (0.1968); Non-hospitalized patients: 0.6484 (0.1860)
Antoniu 2014	VAS (EQ-5D)	Mean (SD)	Hospitalized patients: 27.91 (25.18)
O'Reilly 2007	EQ-5D utility	Mean (SD)	-0.077 (0.397)
	EQ-5D VAS		25.9 (17.0)

Table 5.4. Willingness to accept mechanical ventilation

Study ID	Instrument	Reported format	Results
Chakrabarti 2009	forced choice: treatment	Choice or proportion of choice	Willingness to accept a IMV during an exacerbation after stage 4: 60% (30/50) willing, 30% (15/50) unwilling, 10% (5=50) unsure; after stage 5: 70% (35/50) willing, 24% (12/50) unwilling, 6% (3/50) unsure.
Claessens 2000	Forced choice: treatment	Choice or proportion of choice	"Very unwilling" or "Would rather die" than be attached to a ventilator "all the time" : 78%
Dales 1999	Probability trade off	Choice or proportion of choice	Baseline Choice ventilation: 35% wanting MV. After using decision aid: 40% wanting MV
Downey 2013	Preference Rating	Mean (SD); Choice or proportion of choice	Mechanical Ventilation with Current Health Preference Rating (from 1 definitely no to 4 definitely yes) mean (SD): 2.7 (1.2) Probably or Definitely Wants Treatment, n (%): 111 (61)
Gaber 2004	Forced choice: treatment	Choice or proportion of choice	Number of patients: Patient's views towards "yes" CPR, IV and NIV: 48 Patient's views towards "yes" IV and NIV: 19 Patient's views towards "yes" IV: 10 Patient's views towards "no" CPR, IV and NIV: 12
Janssen 2011b	Probability trade off	Choice or proportion of choice	COPD patients preferring CPR: 70.50% COPD patients preferring MV: 70.50% Patients want life-sustaining treatments under different scenarios: as disease burden increases, the proportion of preferring MV decreases.
Janssen 2011c	Forced choice: treatment	Choice or proportion of choice	Patients' preferences in their current health state for MV: 70.5% of Dutch population and 58.2% of US patients reported they would accept
Reinke 2011	Forced choice: treatment	Choice or proportion of choice	preferences on MV Total: 220 (64.2%); history of depression: 78 (60.5%); no history of depression: 143 (66.4%)
Rinnenburger 2012	Preferences of decision making mode	Choice or proportion of choice	If respiratory failure rapidly develops. intubation may be necessary to provide more effective ventilation. Would you agree to this procedure? yes: 63 (75%) Sometimes. when non-invasive ventilation is no longer effective or cannot be

			performed due to other reasons. tracheostomy may be necessary with the subsequent insertion of an endotracheal tube connected to a ventilator. Would you agree to this procedure? yes: 47 (55.9%)
Stapleton 2005	Forced choice: treatment	Choice or proportion of choice	want mechanical ventilation: 62.20%
Travaline 1995	Forced choice: treatment	Choice or proportion of choice	decision to use MV yes 15 (40%); no 8 (22%); unsure: 14 (38%)
Wilson 2005	Forced choice: treatment, importance of mechanical ventilation	Choice or proportion of choice	MV choices after the decision aid After reviewing the decision aid, 31 participants (94%) reported that they had reached a decision about whether they personally would accept or forego MV in the event of a serious exacerbation; only two individuals remained completely uncertain. Of those participants who did arrive at a decision, 23 (74%) indicated that they would forego the MV option in favor of SC without MV.

IMV: invasive mechanical ventilation; MV: mechanical ventilation

Table 5.5. Importance of adverse events

Study ID	Instrument	Reported format	Results
Bulcun 2014	Conjoint analysis/Discrete choice analysis	Influence or contribution or weight of certain aspects/attributes	Possibility of experiencing adverse effects from treatment 20%: -0.9 10%: -0.06 4%: 1.0 Difference between highest and lowest utility levels: 8.2 Extent to which the doctor gives sufficient time to listen to the patient, possibility of experiencing adverse effects from treatment, costs of treatment, extent to which the patient sees the same doctor each time, and extent to which the doctor treats the patient as an entire person.
Kawata 2014	Willingness to pay, Conjoint analysis/Discrete choice analysis	Mean (95% CI)	Utility score Mild side effects (no side effects as reference) : -0.29 (-0.33, -0.24) Moderate to severe side effects (no side effects as reference) : -1.13 (-1.18, -1.09) Willingness to pay Mild side effects (no side effects as reference) : $14.81 (12.40–17.22) Moderate to severe side effects (no side effects as reference) : $58.69 (56.28–61.11)
Miravitlles 2007	Ideal characteristics of a COPD therapy	Choice or proportion of choice	Ideal characteristics of a COPD therapy as listed by survey respondents Fewer side effects 36% Quick symptom relief > longer intervals between flare-ups > fewer side effects > better ability to cope with daily chores again > lower costs of treatment > better doses
Sharafkhaneh 2013	Primary disadvantages of nebulization therapy	Choice or proportion of choice	Question: what do you see as the main negatives or disadvantages of nebulization? No negatives: 86 (21%) Side effects: 46 (12%) Device immobile/bulky/cumbersome > time-consuming = side effects > inconvenient/ don't like doing it > having to use it several times a day > care and cleanup after use > too expensive

Table 5.6. Utility of different COPD severities

Study ID	Instrument	Reported format	GOLD classifications			
			Mild COPD	Moderate COPD	Severe COPD	Very Severe COPD
Boland 2014	EQ-5D utility	Mean (SD)	0.81 (0.22)	0.62 (0.3)	0.76 (0.24)	0.64 (0.28)
	EQ-5D VAS		73.6 (14.0)	58.1 (17.0)	68.3 (14.4)	54.9 (17.7)
Boros 2012	VAS	Mean (95% CI, SD)	73.04 (95% CI 72.16 73.92; SD 16.357)	62.56 (95% CI 62.08 63.03; SD 16.447)	44.56 (95% CI 43.89 45.22; SD 16.072)	32.05 (95% CI 30.19 33.91; SD 17.062)
Chen 2014	EQ-5D utility	Mean			0.686	0.565
	SF-6D utility (HK value set)				0.646	0.597
Fletcher 2011	EQ-5D utility	Mean (SEM)	0.836 (0.007)	0.579 (0.009)	0.409 (0.015)	
	EQ-5D VAS	Mean (SEM)	73.3 [0. 5]	56.1 [0. 6]	45.9 [0.9]	
Kim 2014	EQ-5D utility	Mean (SD), Adjusted mean (SE)	0.83 (0.17) adjusted 0.83, SE: 0.04	0.88 (0.12) adjusted 0.88 (0.02)	0.82 (0.16) adjusted 0.81 (0.03)	0.61 (0.26) adjusted 0.60, SE (0.04)
	EQ-VAS		71.9 (18.9) adjusted 73.9, SE: 5.4	71.9 (17.8) adjusted 75.1, SE: 2.9	65.0 (20.6) adjusted 68.9, SE: 3.3	60.9 (13.9) adjusted 65.1, SE: 5.6
Lin 2014	EQ-5D VAS	Mean (SD)	0.81 (0.14)	0.81 0.14)	0.76 (0.17)	0.74 (0.15)
	EQ-5D utility		76.6 (17.5)	72.6 (19.1)	65.7 (20.1)	61.1 (19.7)
Menn 2010	EQ-5D utility	Mean (SD)			0.62 (0.26)	0.60 (0.26)
	SF-6D utility	Mean (SD)			0.61 (0.13)	0.54 (0.08)
Pickard 2011	EQ-5D utility (US value set)	Mean (SD)	0.8 (0.13)	0.7 (0.21)	0.72 (0.19)	0.72 (0.16)
	EQ-5D utility (UK value set)		0.73 (0.19)	0,59 (0,32)	0.63 (0.25)	0.63 (0.24)
	EQ-5D VAS		74.3 (16.3)	66.2 (20.0)	60.1 (18.4)	58.7 (15.8)
Punekar 2007	EQ-5D utility	Mean (95% CI)	0.77 (0.73-0.81) in primary care setting	0.68 (0.62 - 0.74)	0.62 (0.56-0.68)	
			0.68 (0.64-0.72) in respiratory specialist care setting	0.72 (0.69-0.75)	0.64 (0.61-0.72)	

Rodriguez Gonzalez-Moro 2009	EQ-5D VAS	Mean (95% CI)		58.9 (58.1-59.9)	45.9 (44.9 -46.7)	
Rutten van Molken 2006	EQ-5D VAS	Mean (SD) or Mean (95% CI)		67.74 (66.51-68.97)	62.45 (60.97-63.92)	57.84 (54.52-61.16)
	EQ-5D utility UK value set			0.787 (0.771-0.802)	0.750 (0.731–0.768)	0.647 (0.598–0.695)
	EQ-5D utility US value set			0.832 (0.821–0.843)	0.803 (0.790–0.816)	0.731 (0.699–0.762)
Rutten van Molken 2009	VAS	Mean [SE] (utility of mild COPD, and disutility in relation to mild COPD)	Mild COPD: 0.811 [0.011]	Moderate: -0.133 [0.006]	Severe: -0.354 [0.006]	Very severe: -0.508 [0.006]
	TTO		Mild COPD: 0.974 [0.017]	Moderate: -0.045 [0.008]	Severe: -0.257 [0.008]	Very severe: -0.452 [0.008]
Scharf 2011	HUI utility	Mean (SD); Median, IQR	0.40 (0.33)	0.58 (0.36)	0.53 (0.35)	0.39 (0.51)
Schunemann 2007	Standard Gamble	Mean (SD)	0.79 (0.20)	0.62 (0.27)	0.42 (0.27)	
	VAS		0.80 (0.12)	0.61 (0.12)	0.36 (0.14)	
Solem 2013	EQ-5D utility	Mean (SD)			0.707 (0.174)	0.623 (0.234)
Stahl 2005	EQ-5D VAS	Mean (SD)	0.73 (0.21)	0.65 (0.24)	0.62 (0.21)	0.37 (0.28)
	EQ-5D utility		0.84 (0.15)	0.73 (0.23)	0.74 (0.25)	0.52 (0.26)
Starkie 2011	EQ-5D utility	Mean (Range)		Observed utilty for moderate COPD 0.752 (0.22) Predicted OLS for moderate COPD 0.752 (0.14) Predicted GLM for moderate COPD 0.754 (0.15) Predicted 2 PART for moderate COPD 0.755 (0.15)	Observed utilty for severe COPD 0.708 (0.23) Predicted OLS for severe COPD 0.704 (0.15) Predicted GLM for severe COPD 0.705 (0.15) Predicted 2 PART for severe COPD 0.706 (0.15)	Observed utilty for very severe COPD 0.672 (0.22) Predicted OLS for very sever COPD 0.667 (0.15) Predicted GLM for very severe COPD 0.667 (0.14)

						Predicted 2 PART for very severe COPD 0.666 (0.14)
Szende 2009	EQ-5D utility	Mean (SD); Median, Range	0.85 (0.16)	0.73 (0.21)	0.74 (0.24)	0.53 (0.28)
	SF-6D utility	Mean (SD); Median, Range	0.80 (0.13)	0.73 (0.13)	0.73 (0.14)	0.62 (0.15)
Vestbo 2014	EQ-5D utility	Mean	0.94	0.78	0.97	0.62

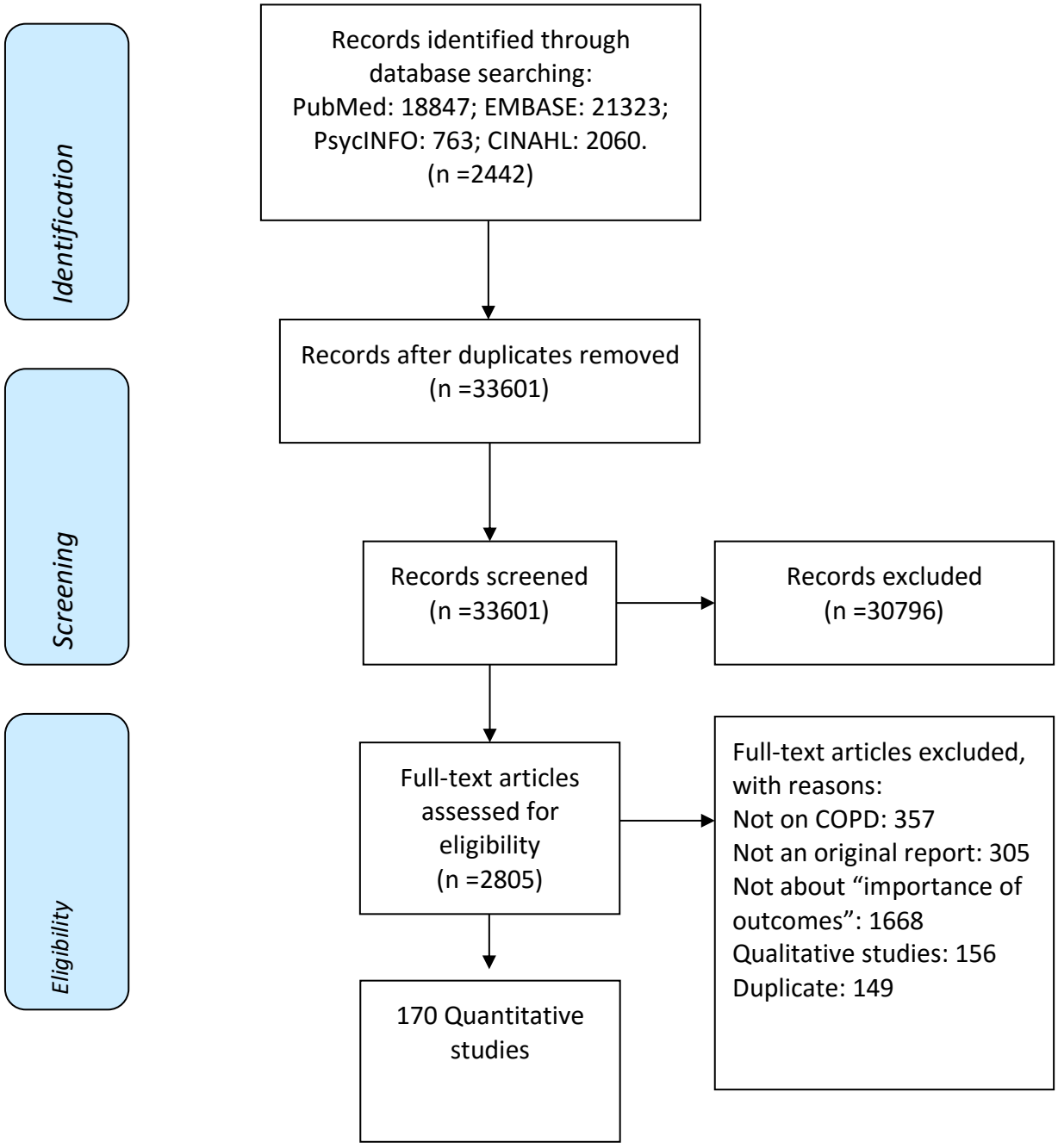


Figure 5.1. Flow Diagram for systematic review on COPD patients' values and preferences

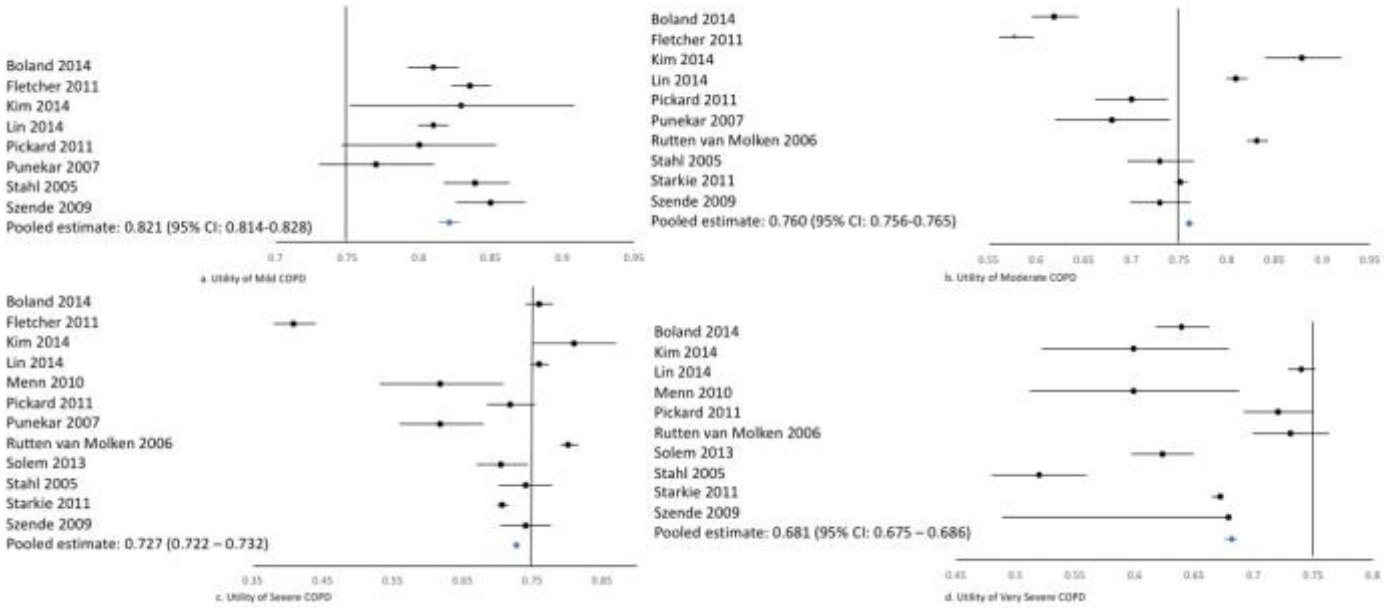


Figure 5.2. Forest plots for utility of different COPD severities

Abbreviations

CINAHL: Cumulative Index to Nursing and Allied Health Literature

CI: confidence interval

COPD: Chronic Obstructive Pulmonary Disease

DCE: discrete choice experiment

EQ-5D: EuroQual-5-dimension (a quality of life measurement tool)

FEV: Forced expiratory volume

GRADE: Grading of Recommendations Assessment, Development and Evaluation

GOLD: Global Initiative for Chronic Obstructive Lung Disease

HUI: health utility index

IQR: interquartile range

PRISMA: Preferred Reporting in Systematic Reviews and Meta-Analyses

QWB: quality of wellbeing

RCT: randomized controlled trial

SF-6D: Short form-6-dimension (a quality of life measurement tool)

SG: standard gamble

TTO: time trade off

VAS: visual analogue scales

Contributors: YZ, PA, AA, GG, and HJS designed the systematic review methodology for patient values and preferences; YZ, RM, PA, AS, HA, GPM, WW, MV, MMB, RJ, MW, KS, AA, JK, LBS, MW, YZ, SN, and YF screened

the literature and abstracted the data; YZ, RM, PA, GG, and HJS drafted the manuscript; All authors read and approved the final manuscript.

Acknowledgement: We are grateful to Dr. Amiram Gafni from McMaster University for the comments on the manuscripts, and Dr. Sean Doran from University of Missouri-Kansas City for the title and abstract screening.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. It was funded through internal research funds at McMaster University available to HJS.

Competing interests: All authors have completed the ICMJE uniform disclosure from at http://www.icmje.org/coi_disclosure.pdf and declare: None of the authors have finial conflict of interests.

Ethical approval: Not required.

Data sharing: The datasets supporting the conclusions of this article are included within the article and its additional file.

Transparency: The lead authors affirm that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

5.6. References

1. Sackett DL, Rosenberg WM, Gray JA, et al. Evidence based medicine: what it is and what it isn't. *BMJ (Clinical research ed)* 1996;312(7023):71-2. [published Online First: 1996/01/13]
2. Schunemann HJ, Wiercioch W, Etxeandia I, et al. Guidelines 2.0: systematic development of a comprehensive checklist for a successful guideline enterprise. *CMAJ : Canadian Medical Association journal = journal de l'Association medicale canadienne* 2014;186(3):E123-42. doi: 10.1503/cmaj.131237 [published Online First: 2013/12/18]
3. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical research ed)* 2016;353:i2016. doi: 10.1136/bmj.i2016
4. Bremner KE, Chong CA, Tomlinson G, et al. A Review and meta-analysis of prostate cancer utilities. *Medical Decision Making* 2007;27(3):288-98. doi: http://dx.doi.org/10.1177/0272989X07300604
5. MacLean S, Mulla S, Akl EA, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest* 2012;141(2 Suppl):e1S-23S. doi: 10.1378/chest.11-2290 [published Online First: 2012/02/15]
6. Pickard AS, Wilke CT, Lin H-W, et al. Health utilities using the EQ-5D in studies of cancer. *PharmacoEconomics* 2007;25(5):365-84.
7. Sepucha K, Ozanne EM. How to define and measure concordance between patients' preferences and medical treatments: A systematic review of approaches and recommendations for standardization. *Patient Education and Counseling* 2010;78(1):12-23. doi: http://dx.doi.org/10.1016/j.pec.2009.05.011
8. Joy SM, Little E, Maruthur NM, et al. Patient preferences for the treatment of type 2 diabetes: A scoping review. *PharmacoEconomics* 2013;31(10):877-92. doi: http://dx.doi.org/10.1007/s40273-013-0089-7
9. Gafni A, Birch S. Preferences for outcomes in economic evaluation: an economic approach to addressing economic problems. *Social science & medicine (1982)* 1995;40(6):767-76. [published Online First: 1995/03/01]
10. Torrance GW. Utility measurement in healthcare: the things I never got to. *PharmacoEconomics* 2006;24(11):1069-78. [published Online First: 2006/10/28]
11. Lopez-Campos JL, Tan W, Soriano JB. Global burden of COPD. *Respirology (Carlton, Vic)* 2016;21(1):14-23. doi: 10.1111/resp.12660 [published Online First: 2015/10/24]
12. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. *Lancet (London, England)* 2007;370(9589):765-73. doi: 10.1016/s0140-6736(07)61380-4 [published Online First: 2007/09/04]

13. Chronic obstructive pulmonary disease: Definition, clinical manifestations, diagnosis, and staging http://www.uptodate.com/contents/chronic-obstructive-pulmonary-disease-definition-clinical-manifestations-diagnosis-and-staging?source=search_result&search=COPD&selectedTitle=1~150 .Accessed on Jan 31, 2015.
14. 2017 Global Initiative for Chronic Obstructive Lung Disease. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Pulmonary Disease [Available from: http://www.goldcopd.org/.
15. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic reviews* 2015;4:1. doi: 10.1186/2046-4053-4-1 [published Online First: 2015/01/03]
16. Selva A SI, Y Z, AJ S, et al. Development of a search strategy for studies about patients' values and preferences (submitted for publication).
17. Zhang Y, P A-C, Yepes-Nuñez J, J., et al. GRADE guidance for rating the certainty of a body of evidence describing the relative importance of outcomes or values and preferences: 1. Risk of bias and indirectness. In: University M, ed., 2017.
18. Birch S, Ismail AI. Patient preferences and the measurement of utilities in the evaluation of dental technologies. *Journal of dental research* 2002;81(7):446-50. [published Online First: 2002/08/06]
19. Peasgood T, Brazier J. Is Meta-Analysis for Utility Values Appropriate Given the Potential Impact Different Elicitation Methods Have on Values? *PharmacoEconomics* 2015;33(11):1101-5. doi: 10.1007/s40273-015-0310-y [published Online First: 2015/07/03]
20. Cross J, Elender F, Barton G, et al. A randomised controlled equivalence trial to determine the effectiveness and cost-utility of manual chest physiotherapy techniques in the management of exacerbations of chronic obstructive pulmonary disease (MATREX). *Health technology assessment (Winchester, England)* 2010;14(23):1-147, iii-iv. doi: 10.3310/hta14230 [published Online First: 2010/05/22]
21. Goossens LM, Nivens MC, Sachs P, et al. Is the EQ-5D responsive to recovery from a moderate COPD exacerbation? *Respir Med* 2011;105(8):1195-202. doi: 10.1016/j.rmed.2011.02.018 [published Online First: 2011/03/29]
22. Miravitlles M, Naberan K, Cantoni J, et al. Socioeconomic status and health-related quality of life of patients with chronic obstructive pulmonary disease. *Respiration; international review of thoracic diseases* 2011;82(5):402-8. doi: 10.1159/000328766 [published Online First: 2011/07/23]
23. Rutten-van Molken MP, Hoogendoorn M, Lamers LM. Holistic preferences for 1-year health profiles describing fluctuations in health: the case of chronic obstructive pulmonary disease. *PharmacoEconomics* 2009;27(6):465-77. doi: 10.2165/00019053-200927060-00003 [published Online First: 2009/07/31]

24. Alcazar B, Garcia-Polo C, Herrejon A, et al. Factors associated with hospital admission for exacerbation of chronic obstructive pulmonary disease. [Spanish]. *Archivos de Bronconeumologia* 2012;48(3):70-76. doi: http://dx.doi.org/10.1016/j.arbres.2011.10.009

25. Antoniu SA, Puiu A, Zaharia B, et al. Health status during hospitalisations for chronic obstructive pulmonary disease exacerbations: The validity of the Clinical COPD Questionnaire. *Expert Review of Pharmacoeconomics and Outcomes Research* 2014;14(2):283-87. doi: http://dx.doi.org/10.1586/14737167.2014.887446
26. O'Reilly JF, Williams AE, Rice L. Health status impairment and costs associated with COPD exacerbation managed in hospital. *Int J Clin Pract* 2007;61(7):1112-20. doi: 10.1111/j.1742-1241.2007.01424.x [published Online First: 2007/06/20]
27. Wildman MJ, Sanderson CF, Groves J, et al. Survival and quality of life for patients with COPD or asthma admitted to intensive care in a UK multicentre cohort: the COPD and Asthma Outcome Study (CAOS). *Thorax* 2009;64(2):128-32. doi: 10.1136/thx.2007.091249 [published Online First: 2008/10/15]
28. Chakrabarti B, Sulaiman MI, Davies L, et al. A study of patient attitudes in the United Kingdom toward ventilatory support in chronic obstructive pulmonary disease. *Journal of Palliative Medicine* 2009;12(11):1029-35. doi: http://dx.doi.org/10.1089/jpm.2009.0160
29. Claessens MT, Lynn J, Zhong Z, et al. Dying with lung cancer or chronic obstructive pulmonary disease: insights from SUPPORT. Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments. *Journal of the American Geriatrics Society* 2000;48(5 Suppl):S146-53. [published Online First: 2000/05/16]
30. Dales RE, O'Connor A, Hebert P, et al. Intubation and mechanical ventilation for COPD: development of an instrument to elicit patient preferences. *Chest* 1999;116(3):792-800. [published Online First: 1999/09/24]
31. Downey L, Au DH, Curtis JR, et al. Life-sustaining treatment preferences: matches and mismatches between patients' preferences and clinicians' perceptions. *Journal of pain and symptom management* 2013;46(1):9-19. doi: 10.1016/j.jpainsymman.2012.07.002 [published Online First: 2012/09/29]
32. Gaber KA, Barnett M, Planchant Y, et al. Attitudes of 100 patients with chronic obstructive pulmonary disease to artificial ventilation and cardiopulmonary resuscitation. *Palliative medicine* 2004;18(7):626-9. [published Online First: 2004/11/16]
33. Janssen DJ, Spruit MA, Schols JM, et al. A call for high-quality advance care planning in outpatients with severe COPD or chronic heart failure. *Chest* 2011;139(5):1081-8. doi: 10.1378/chest.10-1753 [published Online First: 2010/09/11]
34. Janssen DJA, Curtis JR, Au DH, et al. Patient-clinician communication about end-of-life care for Dutch and US patients with COPD. *European*

Respiratory Journal 2011;38(2):268-76. doi: http://dx.doi.org/10.1183/09031936.00157710
35. Reinke LF, Slatore CG, Udris EM, et al. The association of depression and preferences for life-sustaining treatments in veterans with chronic obstructive pulmonary disease. *Journal of pain and symptom management* 2011;41(2):402-11. doi: 10.1016/j.jpainsymman.2010.05.012 [published Online First: 2010/12/15]
36. Rinnenburger D, Alma MG, Bigioni D, et al. End-of-life decision making in respiratory failure. The therapeutic choices in chronic respiratory failure in a 7-item questionnaire. *Annali dell'Istituto superiore di sanita* 2012;48(3):328-33. doi: Doi: 10.4415/ann_12_03_14 [published Online First: 2012/09/26]
37. Stapleton RD, Nielsen EL, Engelberg RA, et al. Association of depression and life-sustaining treatment preferences in patients with COPD. *Chest* 2005;127(1):328-34. doi: 10.1378/chest.127.1.328 [published Online First: 2005/01/18]
38. Travaline JM, Silverman HJ. Discussions with outpatients with chronic obstructive pulmonary disease regarding mechanical ventilation as life-sustaining therapy. *Southern medical journal* 1995;88(10):1034-8. [published Online First: 1995/10/01]
39. Wilson KG, Aaron SD, Vandemheen KL, et al. Evaluation of a decision aid for making choices about intubation and mechanical ventilation in chronic obstructive pulmonary disease. [References]. *Patient Education and Counseling* 2005;57(1):88-95. doi: http://dx.doi.org/10.1016/j.pec.2004.04.004
40. Bulcun E, Ekici M, Ekici A. Assessment of patients' preferences regarding the characteristics associated with the treatment of chronic obstructive pulmonary disease. *International journal of chronic obstructive pulmonary disease* 2014;9:363-8. doi: 10.2147/copd.s56229 [published Online First: 2014/05/03]
41. Kawata AK, Kleinman L, Harding G, et al. Evaluation of Patient Preference and Willingness to Pay for Attributes of Maintenance Medication for Chronic Obstructive Pulmonary Disease (COPD). *Patient* 2014 doi: 10.1007/s40271-014-0064-1 [published Online First: 2014/06/04]
42. Miravitlles M, Anzueto A, Legnani D, et al. Patient's perception of exacerbations of COPD-the PERCEIVE study. *Respir Med* 2007;101(3):453-60. doi: http://dx.doi.org/10.1016/j.rmed.2006.07.010
43. Sharafkhaneh A, Wolf RA, Goodnight S, et al. Perceptions and attitudes toward the use of nebulized therapy for COPD: patient and caregiver perspectives. *Copd* 2013;10(4):482-92. doi: 10.3109/15412555.2013.773302 [published Online First: 2013/07/24]
44. Boland MRS, Tsiachristas A, Kruis AL, et al. Are GOLD ABCD groups better associated with health status and costs than GOLD 1234 grades? A cross-sectional study. *Primary Care Respiratory Journal* 2014;23(1):30-37. doi: http://dx.doi.org/10.4104/pcrj.2014.00002

45. Boros PW, Lubinski W. Health state and the quality of life in patients with chronic obstructive pulmonary disease in Poland: a study using the EuroQoL-5D questionnaire. *Polskie Archiwum Medycyny Wewnetrznej* 2012;122(3):73-81. [published Online First: 2012/02/23]
46. Fletcher MJ, Upton J, Taylor-Fishwick J, et al. COPD uncovered: an international survey on the impact of chronic obstructive pulmonary disease [COPD] on a working age population. *BMC public health* 2011;11:612. doi: 10.1186/1471-2458-11-612 [published Online First: 2011/08/03]
47. Kim SH, Oh YM, Jo MW. Health-related quality of life in chronic obstructive pulmonary disease patients in Korea. *Health and quality of life outcomes* 2014;12:57. doi: 10.1186/1477-7525-12-57 [published Online First: 2014/04/25]
48. Lin FJ, Pickard AS, Krishnan JA, et al. Measuring health-related quality of life in chronic obstructive pulmonary disease: properties of the EQ-5D-5L and PROMIS-43 short form. *BMC medical research methodology* 2014;14:78. doi: 10.1186/1471-2288-14-78 [published Online First: 2014/06/18]
49. Punekar YS, Rodriguez-Roisin R, Sculpher M, et al. Implications of chronic obstructive pulmonary disease (COPD) on patients' health status: a western view. *Respir Med* 2007;101(3):661-9. doi: 10.1016/j.rmed.2006.06.001 [published Online First: 2007/01/02]
50. Scharf SM, Maimon N, Simon-Tuval T, et al. Sleep quality predicts quality of life in chronic obstructive pulmonary disease. *International journal of chronic obstructive pulmonary disease* 2011;6:1-12. doi: 10.2147/copd.s15666 [published Online First: 2011/02/12]
51. Schunemann HJ, Norman G, Puhan MA, et al. Application of generalizability theory confirmed lower reliability of the standard gamble than the feeling thermometer. *Journal of clinical epidemiology* 2007;60(12):1256-62. doi: 10.1016/j.jclinepi.2007.03.010 [published Online First: 2007/11/14]
52. Stahl E, Lindberg A, Jansson SA, et al. Health-related quality of life is related to COPD disease severity. *Health and quality of life outcomes* 2005;3:56. doi: 10.1186/1477-7525-3-56 [published Online First: 2005/09/13]
53. Szende A, Leidy NK, Stahl E, et al. Estimating health utilities in patients with asthma and COPD: evidence on the performance of EQ-5D and SF-6D. *Qual Life Res* 2009;18(2):267-72. doi: 10.1007/s11136-008-9429-z [published Online First: 2008/12/24]
54. Pickard AS, Yang Y, Lee TA. Comparison of health-related quality of life measures in chronic obstructive pulmonary disease. *Health and quality of life outcomes* 2011;9:26. doi: 10.1186/1477-7525-9-26 [published Online First: 2011/04/20]
55. Rutten-van Molken MP, Oostenbrink JB, Tashkin DP, et al. Does quality of life of COPD patients as measured by the generic EuroQol five-dimension questionnaire differentiate between COPD severity stages? *Chest* 2006;130(4):1117-28. doi: 10.1378/chest.130.4.1117 [published Online First: 2006/10/13]

56. Starkie HJ, Briggs AH, Chambers MG, et al. Predicting EQ-5D values using the SGRQ. *Value in health : the journal of the International Society for Pharmacoeconomics and Outcomes Research* 2011;14(2):354-60. doi: 10.1016/j.jval.2010.09.011 [published Online First: 2011/03/16]
57. Menn P, Weber N, Holle R. Health-related quality of life in patients with severe COPD hospitalized for exacerbations - comparing EQ-5D, SF-12 and SGRQ. *Health and quality of life outcomes* 2010;8:39. doi: 10.1186/1477-7525-8-39 [published Online First: 2010/04/20]
58. Solem CT, Sun SX, Sudharshan L, et al. Exacerbation-related impairment of quality of life and work productivity in severe and very severe chronic obstructive pulmonary disease. *International journal of chronic obstructive pulmonary disease* 2013;8:641-52. doi: 10.2147/copd.s51245 [published Online First: 2014/01/01]
59. Eskander A, Waddell TK, Faughnan ME, et al. BODE index and quality of life in advanced chronic obstructive pulmonary disease before and after lung transplantation. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation* 2011;30(12):1334-41. doi: 10.1016/j.healun.2011.06.006 [published Online First: 2011/07/26]
60. Stavem K, Kristiansen IS, Olsen JA. Association of time preference for health with age and disease severity. *The European journal of health economics : HEPAC : health economics in prevention and care* 2002;3(2):120-4. doi: 10.1007/s10198-002-0102-0 [published Online First: 2002/06/01]
61. Kuyucu T, Guclu SZ, Saylan B, et al. A cross-sectional observational study to investigate daily symptom variability, effects of symptom on morning activities and therapeutic expectations of patients and physicians in COPD-SUNRISE study. *Tuberkuloz ve Toraks* 2011;59(4):328-39. doi: http://dx.doi.org/10.5578/tt.3268
62. Rocker GM, Simpson AC, Horton R, et al. Opioid therapy for refractory dyspnea in patients with advanced chronic obstructive pulmonary disease: patients' experiences and outcomes. *CMAJ open* 2013;1(1):E27-36. doi: 10.9778/cmajo.20120031 [published Online First: 2013/01/01]
63. Haughney J, Partridge MR, Vogelmeier C, et al. Exacerbations of COPD: Quantifying the patient's perspective using discrete choice modellling. *European Respiratory Journal* 2005;26(4):623-29. doi: http://dx.doi.org/10.1183/09031936.05.00142704
64. Pisa G, Freytag S, Schandry R. Chronic obstructive pulmonary disease (COPD) patients' disease-related preferences : a study using conjoint analysis. *Patient* 2013;6(2):93-101. doi: 10.1007/s40271-013-0007-2 [published Online First: 2013/03/26]
65. Brophy C, Kastelik JA, Gardiner E, et al. Quality of life measurements and bronchodilator responsiveness in prescribing nebulizer therapy in COPD. *Chronic Respiratory Disease* 2008;5(1):13-18. doi: http://dx.doi.org/10.1177/1479972307087652
66. Chapman KR, Fogarty CM, Peckitt C, et al. Delivery characteristics and patients' handling of two single-dose dry-powder inhalers used in COPD.

International journal of chronic obstructive pulmonary disease 2011;6:353-63. doi: 10.2147/copd.s18529 [published Online First: 2011/07/16]

67. Hansen NCG, May O. Domiciliary nebulized terbutaline in severe chronic airways obstruction. *European Respiratory Journal* 1990;3(4):463-64.
68. Mahler DA, Waterman LA, Ward J, et al. Comparison of dry powder versus nebulized beta-agonist in patients with COPD who have suboptimal peak inspiratory flow rate. *Journal of aerosol medicine and pulmonary drug delivery* 2014;27(2):103-9. doi: 10.1089/jamp.2013.1038 [published Online First: 2013/06/12]
69. Ohno T, Wada S, Hanada S, et al. Efficacy of indacaterol on quality of life and pulmonary function in patients with COPD and inhaler device preferences. *International journal of chronic obstructive pulmonary disease* 2014;9:107-14. doi: 10.2147/copd.s56777 [published Online First: 2014/02/04]
70. Ojoo JC, Moon T, McGlone S, et al. Patients' and carers' preferences in two models of care for acute exacerbations of COPD: results of a randomised controlled trial. *Thorax* 2002;57(2):167-9. [published Online First: 2002/02/06]
71. Siler TM, LaForce CF, Kianifard F, et al. Once-daily indacaterol 75 micro g in moderate- to-severe COPD: Results of a Phase IV study assessing time until patients' perceived onset of effect. *International Journal of COPD* 2014;9:919-25. doi: http://dx.doi.org/10.2147/COPD.S67356
72. Sutherland ER, Brazinsky S, Feldman G, et al. Nebulized formoterol effect on bronchodilation and satisfaction in COPD patients compared to QID ipratropium/albuterol MDI. *Current medical research and opinion* 2009;25(3):653-61. doi: 10.1185/03007990802708152 [published Online First: 2009/02/24]
73. Utens CM, Goossens LM, van Schayck OC, et al. Patient preference and satisfaction in hospital-at-home and usual hospital care for COPD exacerbations: results of a randomised controlled trial. *International journal of nursing studies* 2013;50(11):1537-49. doi: 10.1016/j.ijnurstu.2013.03.006 [published Online First: 2013/04/16]
74. Utens CM, van Schayck OC, Goossens LM, et al. Informal caregiver strain, preference and satisfaction in hospital-at-home and usual hospital care for COPD exacerbations: results of a randomised controlled trial. *International journal of nursing studies* 2014;51(8):1093-102. doi: 10.1016/j.ijnurstu.2014.01.002 [published Online First: 2014/02/04]
75. van der Palen J, Ginko T, Kroker A, et al. Preference, satisfaction and errors with two dry powder inhalers in patients with COPD. *Expert opinion on drug delivery* 2013;10(8):1023-31. doi: 10.1517/17425247.2013.808186 [published Online First: 2013/06/12]
76. van der Palen J, van der Valk P, Goosens M, et al. A randomised cross-over trial investigating the ease of use and preference of two dry powder inhalers in patients with asthma or chronic obstructive pulmonary disease.

Expert opinion on drug delivery 2013;10(9):1171-8. doi: 10.1517/17425247.2013.817387 [published Online First: 2013/07/03]
77. Wilson DS, Gillion MS, Rees PJ. Use of dry powder inhalers in COPD. *Int J Clin Pract* 2007;61(12):2005-8. doi: 10.1111/j.1742-1241.2007.01593.x [published Online First: 2007/11/14]
78. Mutterlein R, Schmidt G, Fleischer W, et al. A new inhalation system for bronchodilatation. [German]. *Fortschritte der Medizin* 1990;108(11):61-66.
79. Oliver S, Rees PJ. Inhaler use in chronic obstructive pulmonary disease. *Int J Clin Pract* 1997;51(7):443-5. [published Online First: 1998/04/16]
80. Molimard M, Colthorpe P. Inhaler Devices for Chronic Obstructive Pulmonary Disease: Insights from Patients and Healthcare Practitioners. *Journal of aerosol medicine and pulmonary drug delivery* 2014 doi: 10.1089/jamp.2014.1142 [published Online First: 2014/09/30]
81. Moore AC, Stone S. Meeting the needs of patients with COPD: patients' preference for the Diskus inhaler compared with the Handihaler. *Int J Clin Pract* 2004;58(5):444-50. [published Online First: 2004/06/23]
82. Polatli M, Bilgin C, Saylan B, et al. A cross sectional observational study on the influence of chronic obstructive pulmonary disease on activities of daily living: The COPD-Life study. *Tuberkuloz ve Toraks* 2012;60(1):1-12. doi: http://dx.doi.org/10.5578/tt.3414
83. Schunemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *BMJ (Clinical research ed)* 2008;336(7653):1106-10. doi: 10.1136/bmj.39500.677199.AE [published Online First: 2008/05/17]
84. Ryan M. Discrete choice experiments in health care. *BMJ (Clinical research ed)* 2004;328(7436):360-61.
85. Ryan M, Farrar S. Using conjoint analysis to elicit preferences for health care. *BMJ (Clinical research ed)* 2000;320(7248):1530-33. doi: 10.1136/bmj.320.7248.1530
86. Sculpher M, Bryan S, Fry P, et al. Patients' preferences for the management of non-metastatic prostate cancer: discrete choice experiment. *BMJ (Clinical research ed)* 2004;328(7436):382.
87. Moayeri F, Hsueh YS, Clarke P, et al. Health State Utility Value in Chronic Obstructive Pulmonary Disease (COPD); The Challenge of Heterogeneity: A Systematic Review and Meta-Analysis. *Copd* 2016;13(3):380-98. doi: 10.3109/15412555.2015.1092953 [published Online First: 2015/12/19]
88. Petrillo J, van Nooten F, Jones P, et al. Utility estimation in chronic obstructive pulmonary disease: a preference for change? *PharmacoEconomics* 2011;29(11):917-32. doi: 10.2165/11589280-000000000-00000 [published Online First: 2011/10/13]
89. Pickard AS, Wilke C, Jung E, et al. Use of a preference-based measure of health (EQ-5D) in COPD and asthma. *Respir Med* 2008;102(4):519-36. doi: http://dx.doi.org/10.1016/j.rmed.2007.11.016

90. Bereza BG, Troelsgaard Nielsen A, Valgardsson S, et al. Patient preferences in severe COPD and asthma: a comprehensive literature review. *International journal of chronic obstructive pulmonary disease* 2015;10:739-44. doi: 10.2147/copd.s82179 [published Online First: 2015/04/29]
91. Brooker AS, Carcone S, Witteman W, et al. Quantitative patient preference evidence for health technology assessment: A case study. *International journal of technology assessment in health care* 2013;29(3):290-300. doi: http://dx.doi.org/10.1017/S0266462313000329
92. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. *Journal of clinical epidemiology* 2011;64(4):401-6. doi: 10.1016/j.jclinepi.2010.07.015 [published Online First: 2011/01/07]

Chapter 6. Conclusion

In this work, we conceptualize patient values and preferences as the importance patients placing on the outcomes of interest, and incorporate this type of evidence in real world examples of systematic review and guideline development projects. We developed the GRADE approach to assess the certainty of evidence in the relative importance of outcomes.

This work is an ongoing effort to further develop the GRADE approach to assess the certainty of bodies of evidence. It will inform the methodological underpinnings of GRADE Evidence to Decision frameworks. Ultimately, the results of this work will further improve the rigorousness of the structured and transparent process to develop GRADE recommendations.

6.1. Summary of findings

6.1.1. Findings on the case example of incorporating evidence about relative importance of outcomes in the guideline development process

We describe an approach for the incorporation of the relative importance of health outcomes in healthcare recommendations. We applied a systematic review strategy complemented by other information sources. We used illustrative examples to show the usefulness of identifying context specific evidence and using their findings in drafting recommendations for a local setting.

6.1.2. Findings of the GRADE guidance for assessing the evidence about the relative importance of outcomes

Following a multi-pronged approach, we developed guidance and described the rationale for considering GRADE domains for the evidence about the importance of outcomes. We illustrate the application of the GRADE approach with systematic review examples. Users start the assessment of the body of evidence at "high certainty", and rate down for serious problems in risk of bias, indirectness, inconsistency, imprecision and publication bias.

With these examples, we also identified the challenges of applying GRADE to assessing the relative importance of outcome evidence. First, there is no reliable and valid assessment tool for the risk of bias in primary studies eliciting the relative importance of outcomes, our proposed questions for considering the risk of bias could be the basis for the future development work. Second, it is uncommon for systematic review authors to quantitatively synthesize the results across studies. This means we often are unable to make judgments about inconsistency and imprecision assisted by heterogeneity test, I^2, and width of confidence intervals and need to make qualitative judgments. Another challenge is that there is no registration for such studies to help decide whether reporting bias may or may not exist.

6.1.3. Findings of the systematic review on COPD related relative importance of outcome

Our systematic review suggests that for COPD patients, exacerbation and hospitalization have great impacts on their lives. Patients consider adverse events important outcomes during the decision-making process, although less important compared with symptom relief. We assessed the certainty of evidence for the outocmes in this systematic review, and summarized the results in a summary of finding tables. With this case example, we conceptualized values and preferences as "how patients (or other affected, such as caregivers) value the main outcomes". Following this definition, we identified sources describing the relative importance of outcome information. The values are based on utility measurements of outcomes, discrete choice exercises, probability trade-off, and forced choice experiments.

6.2. Implications for the clinicians, guideline developers, health policy makers, and researchers

Integrating patient relative importance of outcomes in decision-making means respecting patients as their best agents, respecting their autonomy, and, by doing that, potentially improving patient adherence.[1 2] This integration could happen at different health care decision level. In an individual patient-clinician encounter, it requires that patients and clinicians to take the "shared decision making" approach for their decisions.[2-4]

When guideline panels decide the direction and strength of recommendations, it is crucial to balance benefits and harms.[5 6] Directly involving patient representative is helpful to decide the balance from the patients' perspective, and potentially helpful for the transparency and acceptability of guidelines. However, it also introduces potential personal bias by the representatives when they are to represent those affected.[7-9] Systematic reviews of the importance of outcomes provide an opportunity to systematically acquire, assess, and apply the evidence without relying on the view of few representatives on guideline panels.

Health technology assessment involves valuing outcomes from a perspective different from patients'[10].[11] Our GRADE guidance describes how to deal with indirectness under different decision-making scenarios. More importantly, the GRADE guidance provides a systematic approach to decide how confident health policy makers can be about the relative importance of outcome evidence ("utility" of health states, in health economics). This could improve the quality of input in the decision analytic models.

6.3. Strengths and challenges of this work

This work is strengthened by the foundation of the GRADE working group. The structured and transparent approach to assess the certainty of evidence and determine the strength of recommendation, is the theoretical basis of this work. [5 6 12]

Nevertheless, throughout our work, we were challenged by the variety of methods used to elicit relative importance of outcomes. This variety creates difficulty for

users to synthesize the results from different studies, to interpret the evidence, and further challenges the assessment of inconsistency and imprecision for a body of evidence. Additionally, for both our case examples and our methodological exploration, we were considering relative importance of outcome evidence in an inclusive manner.

6.4. Further research directions

Many questions in this area of research still need to be explored. Systematic review methodology for this type of evidence is in development.

We focused on measures of central tendency of the relative importance of outcomes, which is for a typical or average person. Explaining the variability of people's relative importance for outcomes is an important area of research. Specific to guideline development, the distribution or variability across individuals influences the strength of recommendations. We need to further explore the degree and causes of variability.

As discussed above, the relative importance of outcome should be used in decision-making across different levels. However, we should explore whether the integration of relative importance of outcomes improves the uptake of recommendations on the health system level; whether the integration of relative importance of outcomes improves the adherence to treatment among patients; whether the integration of relative importance in formulating recommendations improves patient important outcomes themselves; and whether application of

different types of relative importance of outcome leads to different decision-making.

6.5. Final remarks

The work in this thesis provided consistency for the definition, measurement, and application of values and preferences. It created new knowledge on how to identify and summarize values and preference studies, how to assess the certainty of evidence, and how to present the evidence for transparent use.

6.6. References

1. van der Weijden T, Pieterse AH, Koelewijn-van Loon MS, et al. How can clinical practice guidelines be adapted to facilitate shared decision making? A qualitative key-informant study. *BMJ quality & safety* 2013;22(10):855-63. doi: 10.1136/bmjqs-2012-001502 [published Online First: 2013/06/12]
2. Gafni A, Charles C, Whelan T. The physician-patient encounter: the physician as a perfect agent for the patient versus the informed treatment decision-making model. *Social science & medicine (1982)* 1998;47(3):347-54. [published Online First: 1998/07/29]
3. Charles C, Gafni A, Whelan T. Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango). *Social science & medicine (1982)* 1997;44(5):681-92. [published Online First: 1997/03/01]
4. Deber RB. Shared decision making in the real world. *J Gen Intern Med* 1996;11(6):377-8. [published Online First: 1996/06/01]
5. Alonso-Coello P, Schünemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ (Clinical research ed)* 2016;353:i2016. doi: 10.1136/bmj.i2016
6. Alonso-Coello P, Oxman AD, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 2: Clinical practice guidelines. *BMJ (Clinical research ed)* 2016;353:i2089. doi: 10.1136/bmj.i2089
7. World Health Organization. WHO Handbook for Guideline Development. 2nd ed. Switzerland: World Health Organization, 2014.
8. National Institute for Health and Clinical Excellence. The guidelines manual (Nov 2012). London: National Institute for Health and Clinical Excellence. Available from: http://www.nice.org.uk/.
9. Lossie AC, Green J. Building Trust: The History and Ongoing Relationships Amongst DSD Clinicians, Researchers, and Patient Advocacy Groups. *Hormone and metabolic research = Hormon- und Stoffwechselforschung = Hormones et metabolisme* 2015;47(5):344-50. doi: 10.1055/s-0035-1548793 [published Online First: 2015/04/14]
10. Hofmann B. Toward a procedure for integrating moral issues in health technology assessment. *International journal of technology assessment in health care* 2005;21(3):312-8. [published Online First: 2005/08/23]
11. Russell LB, Fryback DG, Sonnenberg FA. Is the societal perspective in cost-effectiveness analysis useful for decision makers? *The Joint Commission journal on quality improvement* 1999;25(9):447-54. [published Online First: 1999/09/11]
12. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ (Clinical research ed)* 2008;336(7650):924-6. doi: 10.1136/bmj.39489.470347.AD [published Online First: 2008/04/26]

Appendix 2.1. Search strategy

Values and preferences terms

1. patient$ participation.mp. or exp patient participation/
2. patient$ satisfaction.mp. or exp patient satisfaction/
3. attitude to health.mp. or exp Attitude to health/
4. (patient$ preference$ or patient$ perception$ or patient$ decision$ or patient$ perspective$ or user$ view$ or patient$ view$ or patient$ value$).mp.
5. (patient$ utilit$ or health utilit$).mp.
6. health related quality of life.mp. or exp "quality of life"/
7. (health stat$ utilit$ or health stat$ indicator$ or (health stat$ adj 2 valu$)).mp. or exp Health Status Indicators/
8. 1 or 2 or 3 or 4 or 5 or 6 or 7

Geographic terms:

1. Saudi Arab$.mp,in. or Saudi Arabia/
2. Riyadh.mp,in.
3. Jeddah.mp,in.
4. Kh*bar.mp,in.
5. Dammam.mp,in.
6. 1 or 2 or 3 or 4 or 5
7. Kuwait$.mp,in. or Kuwait/
8. United Arab Emirates.mp,in. or United Arab Emirates/
9. Qatar$.mp,in. or Qatar/
10. Oman$.mp,in. or Oman/
11. Yemen$.mp,in. or Yemen/
12. Bahr*in$.mp,in. or Bahrain/
13. 7 or 8 or 9 or 10 or 11 or 12
14. Middle East$.mp,in. or Middle East/
15. Jordan$.mp,in. or Jordan/
16. Libya$.mp,in. or Libya/
17. Egypt$.mp,in. or Egypt/
18. Syria$.mp,in. or Syria/
19. Iraq$/ or Iraq.mp,in.
20. Morocc$.mp,in. or Morocco/
21. Tunisia$.mp,in. or Tunisia/
22. Leban$.mp,in. or Lebanon/
23. West Bank.mp,in.
24. Iran$.mp,in. or Iran/
25. Turkey/ or (Turkey or Turkish).mp,in.
26. Algeria$.mp,in. or Algeria/
27. Arab$.mp,in. or Arabs/
28. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21 or 22 or 23 or 24 or 25 or 26
29. 27 or 28
30. 6 or 13 or 29

Appendix 3.1. Summary of Proposed GRADE domains for assessing the Certainty of evidence for relative importance of outcomes (values and preferences) evidence

This document serves as the guidance to assess the certainty of evidence (quality of evidence) for relative importance of outcomes /values and preferences evidence.

Risk of bias

We rate down the certainty of evidence if there is serious risk of bias.
The assessment of risk of bias involves three steps:

1. Inspect each individual primary studies for the following sub-domains;
2. Summarize the risk of bias within a study across the following sub-domains;
3. Summarize the risk of bias across studies.

It should be borne in mind that the assessment of risk is a per-outcome assessment.

Step 1: inspect each individual primary studies for the following subdomains (low or moderate, serious, or critical risk of bias)

Subdomains risk of bias	Questions	Guidance to answer the question	Decision rule
Risk of bias due to selection of participants into the study	*Was an appropriate study sample selected from the sampling frame?* • *Yes* • *Probably yes* • *Probably no* • *No*	*Please also consider:* *What is the sampling strategy? i.e., random sample or consecutive sample, etc.* *What subset of the population or population with what specific characteristics more or less likely to be reached with this sampling strategy?* The sampling strategy solely does not determine the risk of bias; if there is a subset of the population more or less likely to be reached, the answer for "was the study sample selected in a manner to ensure the representativeness" is yes or probably yes.	One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in sample selection.
Risk of bias due to missing data	*Was the attrition sufficiently low to minimize the risk of bias?* • *Yes* • *Probably yes* • *Probably no* • *No*	*Please also consider:* *What was the response rate?* *If follow-ups were planned and used, what was the attrition rate during the follow up?* *Did the authors report the characteristics of the participants responded and those not; if yes, were they different?*	One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in attrition.

		Response rate for 80% or higher would be considered high for a cross-sectional study. Users could set their own criteria for the response rate and attrition rate.	
Risk of bias due to measurement methodology: validity and reliability of the methodology	*Was the chosen methodology for eliciting relative importance of outcomes valid and reliable?* • *Yes* • *Probably yes* • *Probably no* • *No*	*Please also consider:* *What was the measurement instrument selected? What about validity and reliability of this methodology? Or whether this instrument is a widely accepted instrument on this area with reliability and validity?* Please consider yes or probably yes for the following methodologies: standard gamble, time trade off, visual analogue scale (or feeling thermometers), discrete choice, treatment trade-off, willingness to pay	One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in choice of methodology.
Risk of bias due to measurement methodology: Administration of the methodology	*Was the instrument administered in the intended way?* • *Yes* • *Probably yes* • *Probably no* • *No*		One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in administration of methodology.
Risk of bias due to measurement methodology: Outcome presentation	*Was a valid representation of the outcome (health state) utilized?* • *Yes* • *Probably yes* • *Probably no* • *No*	If the researchers demonstrated they were using available evidence to support the health state presentation, the answer should be yes or probably yes.	One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in outcome presentation.
Risk of bias due to measurement methodology: Understanding of the methodology	*Did the researchers check the understanding to the measurement techniques?* • *The investigator tested the understanding, and understanding was adequate;* • *The investigators did not formally test the understanding, but there was evidence suggesting adequate understanding;* • *The investigator tested the*	*Please also consider:* *Did the researchers pilot the methodology?* *Was it simple enough to assume understanding?* If the methodology is simple, choosing "the investigators did not formally test the understanding, but the results suggested it was adequate" could be appropriate. If the researchers piloted the methodology, choosing "the investigators did not formally test the understanding, but the results suggested it was adequate" may also be appropriate.	One study will be classified as low or moderate (the understanding was adequate or there was evidence suggesting adequate understanding), serious or critical (the understanding was inadequate or there was evidence suggesting inadequate understanding)) risk of bias in understanding of methodology.

	understanding, and the understanding was inadequate; • *The investigators did not formally test the understanding, but there was evidence suggesting inadequate understanding.*		
Risk of bias due to confounding	*Were the results analyzed appropriately to avoid influence of bias and confounding?* • *Yes* • *Probably yes* • *Probably no* • *No*	*Please also consider:* *Whether the adjustment, stratification, strategy to deal with missing data and model selection, if any, was appropriate.*	One study will be classified as low or moderate (yes, or probably yes) or serious or critical (no, or probably no) risk of bias in data analysis.

Step 2: Summarize the risk of bias within a study across the following sub-domains (low, moderate, serious, and critical risk of bias for each study (per-outcome assessment))

- We suggest users classify one particular study as "low" "moderate" "high" or "critical" risk of bias, and avoid using "unclear".

Response option	Criteria
Low risk of bias	The study is classified as with low risk of bias across subdomains.
Moderate risk of bias	The study is classified as low or moderate risk of bias across subdomains.
Serious risk of bias	The study is classified as serious risk of bias for at least one subdomain, but not classified as critical risk of bias for any subdomain.
Critical risk of bias	The study is classified as critical risk of bias for at least one subdomain.

Step 3: Summarize the risk of bias across studies (no serious, serious or very serious risk of bias for each study (per-outcome assessment))

- If most information is from studies at low risk of bias, the overall judgment for risk of bias should be "low risk of bias" and in GRADE certainty of evidence, do not downgrade because the risk of bias is "not serious".

- The overall judgement could be “serious” or even “very serious”, as the contribution of studies with risk of bias concern increases, and accordingly, users downgrade the certainty of evidence due to “serious” (by one level) or “very serious” (by two levels) risk of bias.

Inconsistency

The assessment of risk of bias involves three steps:

1. Inspect if the results across studies are inconsistent (heterogeneous):
    - **if there is no inconsistency noticed, do not rate down for inconsistency and go to other GRADE domains**
    - **If there is inconsistency noticed, go to next step: explore the source of heterogeneity**
2. Explore the source of inconsistency *if there is inconsistency*;
    - **if heterogeneity is unexplained, rate down for inconsistency**
    - **if heterogeneity explained, do not rate down for inconsistency, go to next step**
3. If inconsistency explained, consider whether subgroup effects or methodological explanations of inconsistency are credible
    - **Present the results separately (for systematic review)**
    - **Consider different recommendation for subgroups (for guideline development)**

Step 1. Inspect if the results across studies are inconsistent (heterogeneous):

According to the following questions, are the results across included studies consistent?

- *Yes*
- *Probably yes*
- *Probably no*
- *No*

To answer this signaling question, consider:

- If a meta-analysis was conducted, is the I^2 sufficiently small to exclude important inconsistency?
- Is the probability of inconsistency due to chance large? (Consider: what is the P value of the test heterogeneity)
- Is the variation of point estimates indicating consistent results across studies?
- Is there overlap of confidence intervals indicating consistent results across studies?

- If the Q statistic p-value does not suggest significant heterogeneity (i.e., >0.10), the I^2 value does not suggest large heterogeneity , estimates across studies tend to be similar, and confidence intervals across studies overlap, then yes or probably yes. **Go to other GRADE domains.**
- If one or more of the four (4) above criteria are violated, consider no or probably no. **Go to step 2.**

Step 2. Explore the source of inconsistency *if there is inconsistency*:

Inconsistency due to PICO elements:

- Are the populations studied consistent across studies?
- Are the Intervention and comparison consistent across studies?

- Are the outcomes consistent across studies?

Inconsistency due to methodological elements:

- Are the study designs consistent across studies?
- Are the instruments consistent across studies?
- Are the descriptions or definitions of outcomes consistent across studies?

- If inspection of PICO elements or methodological elements provides **plausible explanations** for heterogeneity, **go to step 3**.
- If inspection of PICO elements or methodological elements is not able to provide plausible explanations for heterogeneity, ra**te down for inconsistency, and go to other GRADE domains**.

Step 3. If heterogeneity explained, consider whether subgroup effects are credible

- Systematic review authors could choose to present the results separately for the subgroup while not rating down the certainty of evidence; guideline panelists should consider whether the different relative importance of outcomes across subgroup justify separate recommendations.

Indirectness

- If most information for the body of evidence comes from studies deviating from the PICO addressed in the systematic review or that is of interest for a guideline, then **rate down**.
- If none of the studies had different study populations, different outcome definitions, or interventions from the PICO that is addressed, then inspect whether most information of the body of evidence comes from studies using indirect measurement of utility, or mapping other measurement results into utility, if so, **consider rating down**.
- If none of the studies had different PICO, or indirect measurement of utility or mapping results, then **do not rate down**.

Questions	Answer options	Instructions
Was the population studied matching the population of interest?	• *Yes* • *Probably yes* • *Probably no* • *No*	Consider whether the study population are newly diagnosed patients facing a choice, or people who are at high risk of disease possibly facing the choice, or proxies (e.g., spouse, other family members or caregivers); Consider whether the patient population are from other healthcare settings (e.g., the primary or secondary care, or outpatients or inpatients, or the countries); Consider whether the study population have made their choice or decision.
Were the outcomes matching the outcomes of interest?	• *Yes* • *Probably yes* • *Probably no* • *No*	
Were the options studied matching the alternative options of interest?	• *Yes* • *Probably yes* • *Probably no* • *No*	An outcome is not indirect because this particular outcome is empirically considered as "surrogate" outcome or non-patient important outcome, since the purpose of relative importance of outcome evidence is to value the outcome.
Were the participants answering questions directly valuing the relative importance of outcomes?	• *Yes* • *Probably yes* • *Probably no* • *No*	*Please also consider:* *Were direct methodologies for outcome utilities rather than indirect methodologies used?* *Were utility directly estimated from a methodology to elicit utility rather than mapping from methodology whose purpose are not eliciting utility?*

Imprecision

For systematic review:

Inspect if the sample size is large enough to sufficiently reduce the risk of chance

- if no, **rate down** for imprecision
- if yes, **inspect the confidence interval**.

Then inspect the confidence interval

- If the confidence interval is narrow, **do not rate down** for imprecision.
- If the confidence interval is not narrow, **rate down** for imprecision.

For guideline development:

First inspect the confidence interval

- If the confidence interval excludes the clinical decision threshold, **inspect the sample size**.
- If the confidence interval includes the clinical decision threshold, **rate down** for imprecision.

Then inspect if the sample size is large enough to sufficiently reduce the risk of chance

- if no, **rate down** for imprecision
- if yes, then **do not rate down** for imprecision.

Questions	**Answer options**
Is the confidence interval narrow (for systematic reviews)?	• *Yes* • *Probably yes* • *Probably no* • *No*
Does the confidence interval exclude the clinical decision threshold between recommending for or against or formulating a recommendation as strong or conditional (for clinical practice guideline)?	• *Yes* • *Probably yes* • *Probably no* • *No*
Is the sample size large enough to sufficiently reduce the risk of chance?	• *Yes* • *Probably yes* • *Probably no* • *No*

Publication bias

- If empirically users have proof or strong doubt of publication bias, **rate down** for publication bias;
- Consider if a) results are not consistent with what was previously shown; b) results are overly redundant (without expected random variation); c) most of the included studies being small studies funded by private funding or d) language other than English have systematically led to delayed or unsuccessful publication, **rate down** for publication bias;
- If no strong reason to support existing publication bias, **do not rate down**.

Appendix 3.2. Other assessed examples

Example 1.

What is the importance patients with head and neck cancer placed on the treatment related health states (outcomes)?

Patient or population: patients with head and neck cancer

Komatsuzaki, et al. Preferences and Utilities of Health Outcomes and Treatments Associated with Head and Neck Cancer: A Systematic Review. Am J Cancer 2006; 5 (1): 27-34

Quality assessment							Estimate (Mean (SD))	Quality
Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other		
Radiotherapy (Utility)	1 cross-sectional study 20 participants [1, 2, 3] RS, TTO, SG	Serious [1]	not serious [4]	Serious [3]	Serious [5]	None	0.66 (with RS), 0.70 (with TTO) and 0.61 (with SG) for laryngeal cancer and 0.78 (with RS), 0.72 (with TTO) and 0.683 (with SG) for floor-of-the-mouth cancer [2]	⊕◯◯◯Very low
Radiotherapy (End of therapy) (Utility)	1 cross-sectional study 61 participants [6] RS, SG	not serious [6]	not serious [4]	not serious	Serious [5]	None	0.698 (0.182) (with RS), 0.842 (0.163) (with SG) for patients with laryngeal cancer	⊕⊕⊕◯ Moderate
General health before	1 cross-sectional study	Serious [7]	not serious	not serious	Serious [5]	None	0.89 (0.099)	⊕⊕◯◯Low

radiotherapy	66 participants [7]		[4]					
General health after radiotherapy (Utility)	1 cross-sectional study 66 participants [7] LASA	Serious [7]	not serious [4]	not serious	Serious [5]	None	0.81 (0.12)	⊕⊕◯◯Low
Xerostomia (dry mouth) after radiotherapy (month 6) (Utility)	1 cross-sectional study 130 participants [8] LASA	Serious [8]	not serious [4]	not serious [9]	not serious	None	0.33 (0.254)	⊕⊕⊕◯ Moderate
Perfect health (Utility)	1 cross-sectional study 114 participants [10] TTO	Serious [10]	not serious [4]	not serious	not serious	None	0.878 (0.174)	⊕⊕⊕◯ Moderate

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

LASA, linear analog self-assessment; RS, rating scale; SG, standard gamble; TTO, time trade off

1. This is a study (van der Donk 1995) using TTO and SG to measure utility. Several respondents were not able to express a preference on >= one occasions. Although no report on the proportion of patient participants, a high proportion of missing value was reported, 11 of 39 respondents for SG, and 6 of 39 for TTO.
2. The authors did not report the variation of the measurement results, only mean.
3. This is a study on patients with previous history of disease (10 patients with a previous history of laryngeal cancer and 10 patients with a previous history of floor-of-the-mouth cancer), as well as physicians. The study population is not the optimal study population, who should be facing the decision.
4. Single study.
5. Small sample size.
6. Llewellyn-Thomas 1992 was a cross-sectional study consecutively recruiting patients undergoing, on an outpatient basis, standard four- or five- week radio therapeutic protocols for either head and neck or cervical/endometrial cancer; 97 patients were approached and 67 consented to enter the study. The researchers applied standard gamble and the rating scale to assess both the individualized and standardized health state descriptions.
7. Llewellyn-Thomas 1993 was a cross-sectional study consecutively recruiting new patients. But the systematic review author only included the results from LASA scale measurement, of which the validity was not tested and the transformation of LASA scale to utility score was not well constructed.

8. This was a cross-sectional study (Ringash 2005) suffering from low participation rate and the participants were potentially different from non-participants since participants were somewhat younger and had slightly higher Karnofsky performance status compared to the non-participants
9. Due to risk of bias in the study population selection, the estimates may be biased. However, we decide not to downgrade for indirectness, to avoid double penalty.
10. Ringash 2000 was a secondary analysis based on a previous RCT. Patients with squamous cell carcinoma of the head and neck who were scheduled to receive RT and in whom > 50% of both parotid glands would receive doses > 50 Gy within 4–5 weeks were recruited but no further information related to recruitment was reported, even in the previous reports. Xerostomia was measured on a patient-reported linear analog scale (LASA) at baseline and 1, 3, and 6 months after RT. Formal validation of this scale had not been previously completed and it is unclear this quality of life measurement could indicate the utility of health states.

Example 3.
Evidence profile
Author(s): **Date:** 2016-05-11
Question: What is the importance patients placed on the noninsulin diabetes medications related health states?
Setting: not specified **Bibliography:** Purnell et al. Diabetes Care. July 2014;37(7):2055-2062.

Quality assessment							Value (95%CI or other measure of variability)	Quality
Outcome	Study Design/ Measurement instrument	Risk of bias	Inconsistency	Indirectness	Imprecision	Other		
Diabetes complication: blindness	1 Cross-sectional study, 473 participants [1] TTO	Not serious	No serious inconsistency	serious indirectness[1,2]	No serious imprecision [3]	none	Mean (SD): 0.39 (0.32)	⊕⊕⊕◯ Moderate
Diabetes complication: +3% weight on the basis	1 Cross-sectional study, 129 participants[4] VAS, SG	Not serious	No serious inconsistency	No serious indirectness	No serious imprecision	none	Mean (SD) of disutility: -0.04 (0.08)	⊕⊕⊕⊕ High
Weight loss/ control	4 cross-sectional studies, 2086 participants [5] DCE or Likert-like scale	Not serious	No serious inconsistency	No serious indirectness	No serious imprecision	none	Patients ranked weight loss/control as more important than treatment administration or frequency (4 of 4 comparisons), cost (1 of 1 comparison), glucose testing (2 of 2 comparisons), gastrointestinal effects (3 of 3 comparisons), hypoglycaemia (4 of 4 comparisons), and potential weight gain (2 of 3 comparisons).	⊕⊕⊕⊕ High
Glycemic control	5 cross-sectional studies, 2493 participants [6] DCE, conjoint analysis or Likert-like scale	Not serious	Serious inconsistency [7]	No serious indirectness [8]	No serious imprecision[9]	none	Glycemic control was ranked more important than treatment administration (4 of 4 comparisons), cost (1 of 1 comparison), glucose testing (1 of 1 comparison), gastrointestinal effects (2 of 4 comparisons), risk of hypoglycaemia (5 of 5 comparisons), and potential weight gain (2 of 4 comparisons).	⊕⊕⊕◯ Moderate
Blood pressure control	1 cross-sectional study, 461 participants [10] DCE	Not serious	No serious inconsistency	serious indirectness [2, 10, 11]	No serious imprecision	none	The WTP values for the more intangible aspects investigated were: SEK 387 (€33.71) to have one less antihypertensive treatment; SEK 223 (€19.51) to reduce blood pressure by 5 mm Hg; SEK 294 (€26.17) to reduce blood pressure by 10 mm Hg; and SEK 430 (€37.70) to improve heart	⊕⊕⊕◯ Moderate

							function.	

CI: Confidence interval; DCE: Discrete choice exercise; SG: Standard Gamble; TTO: Time Trade Off; VAS: Visual Analogue Scale.

1. Chin 2008 was a cross-sectional study on type 2 diabetes patients 65 years or older to measure the utility on diabetic complications and treatment intensity with time trade off technique. In this study, only participants 65 years or older were eligible.
2. Only one single eligible study for this outcome importance assessment; the generalizability of the result is limited since this result was based on a narrowly selected study population.
3. There was important variability across the participants on how they value diabetic complication of blindness, but the estimate was precise with the large sample size.
4. Matza 2007 was a cross-sectional study on type 2 diabetes patients in England and Scotland to measure the disutility associated with treatment benefits (including weight loss/control) and side effects (including weight gain, nausea, hypoglycaemia) through standard gamble.
5. Bogelund 2011 was a discrete choice experiment to weigh the importance of treatment benefits (including glycemic control (HbA 1c), weight loss/control, blood pressure control, improved heart function, possess driver' s license) and treatment burdens (including mode of administration, blood glucose testing, payment per month) and side effects (including hypoglycaemia events, transient nausea).; Jendle 2012 and Jendle 2010 were reports of a discrete choice exercise, weighing treatment benefits (including glycemic control (HbA 1c), weight loss/control, blood pressure control, improved heart function), treatment burdens (including mode of administration, blood glucose testing, payment per month), and side effects (including hypoglycaemia events, Transient nausea).; Polonsky 2011 used Likert scale to indicate the importance of treatment benefits (including glycemic control, weight loss/control), treatment burdens (including treatment frequency, costs) and hypoglycaemia events.
6. Bogelund 2011 was a discrete choice experiment to weigh the importance of treatment benefits (including glycemic control (HbA 1c), weight loss/control, blood pressure control, improved heart function, possess driver' s license) and treatment burdens (including mode of administration, blood glucose testing, payment per month) and side effects (including hypoglycaemia events, transient nausea); Hauber 2009 was a discrete choice exercise weighing glycemic control (HbA 1c) with side effects including hypoglycaemia events, water retention, weight gain, mild stomach upset, heart attack risk; Jendle 2010 was a discrete choice exercise, weighing treatment benefits (including glycemic control (HbA 1c), weight loss/control, blood pressure control, improved heart function), treatment burdens (including mode of administration, blood glucose testing, payment per month), and side effects (including hypoglycaemia events, Transient nausea); Polonsky 2011 used Likert scale to indicate the importance of treatment benefits (including glycemic control, weight loss/control), treatment burdens (including treatment frequency, costs) and hypoglycaemia events; Polster 2010 was a conjoint analysis to trade-off glycemic control (HbA 1c) with dosing schedule and nausea, hypoglycaemia.
7. Different methodologies were used and glycemic control was weighted against different outcomes, the result suggested some difference on the importance of glycemic control.
8. This is not a patient important outcome.
9. The total sample size was large (n=2493).
10. Jendle 2012 was a discrete choice exercise on patients with type 2 diabetes patients, weighing treatment benefits (including glycemic control (HbA 1c), weight loss/control, blood pressure control, improved heart function), treatment burdens (including mode of administration, blood glucose testing, payment per month), and side effects (including hypoglycaemia events, transient nausea).
11. The study participants were selected from efficacy trials of liraglutide versus other commonly used glucose lowering agents for type 2 diabetes. The study population was highly selected and the preference was measured specific for liraglutide vs other glucose lowering drugs, rather than the health states related to noninsulin diabetes medication in general.

Example 4.

Evidence profile

Author(s): **Date:** 2016-05-11

Question: What is the importance patients placed on the psoriasis related health states?

Setting: not specified **Bibliography:** Umar N, et al. Acta Derm Venereol. 2012; 92: 341–346.

Quality assessment							Value (95%CI or other measure of variability)	Quality
Outcome	Study Design/ Measurement instrument	Risk of bias	Inconsistency	Indirectness	Imprecision	Other		
Psoriasis								
Psoriasis (utility)	2 Cross-sectional studies, 283 participants [1,2] VAS, TTO, SG	Not serious	No serious inconsistency	No serious indirectness	No serious imprecision	none	0.69 - 0.7 for rating scale 0.88 (SEM: 0.010) for TTO 0.97 (SEM: 0.007) for SG	⊕⊕⊕⊕ High
Psoriasis (willingness to pay)	2 Cross-sectional studies, 282 participants[1,3] Willingness to pay	Not serious	No serious inconsistency	No serious indirectness	No serious imprecision	none	Lundberg 1999 reported patients were willing to pay between 1253 and 1956 Swedish crowns (SEK) per month for a psoriasis cure, and Schmitt 2008 suggested the WTP were €100 monthly for controlled and €200 monthly for uncontrolled psoriasis. [4]	⊕⊕⊕⊕ High
Mild psoriasis (utility)	1 Cross-sectional study, 87 participants[5] VAS, TTO, SG	Not serious	No serious inconsistency	serious indirectness [5, 6]	Serious imprecision[7]	none	0.71 (IQR: 0.52-0.89) with rating scale 0.89 (IQR: 0.88-0.99) with TTO 0.82 (IQR: 0.79-0.99) with SG	⊕⊕◯◯ Low
Psoriasis treatment related attributes								
Preferences for treatment attributes	1 cross-sectional study, 126 participants [8] DCE	Not serious	No serious inconsistency	serious indirectness [6, 8, 9]	No serious imprecision	none	Patients with psoriasis prioritized low risk of skin cancer and liver damage and preferred treatment that resulted in a shorter time to achieve a moderate improvement over a longer time to relapse. Patients were most willing to wait longer for a treatment to work if the likelihood of skin cancer or liver damage was reduced. 29.1% (30 of 103) ranked time to achieve moderate (50%) improvement as "most important"; while 16.5% (17 of	⊕⊕⊕◯ Moderate

							103) chose time to relapse, 16.6% (19 of 103) and 27.2% (28 of 103) for 20-y risk of experiencing liver damage, 20-y risk of experiencing skin cancer, respectively.	
Intimacy, Physical comfort, Self-care, Ability to work or volunteer, Ability to concentrate, Emotional health, Social comfort, Ability to sleep (Choice)	2 cross-sectional studies, 99 participants [2,10] ranking	Not serious	Serious inconsistency [11]	No serious indirectness	Serious imprecision[12]	none	Delfino et al 2008 reported numbers of patient ranked highly of physical comfort, social comfort, emotional health, self-care, intimacy, ability to sleep, ability to work/volunteer and ability to concentration were 33, 30, 29, 21, 16, 13, 9 and 9 (sample size =40). While Hu 2010 reported percentage of patient ranked physical comfort, emotional health, ability to sleep, ability to work or volunteer, social comfort, self-care, intimacy, ability to concentrate, as important domains were 87%, 68%, 60%, 45%, 53%, 37%, 28%, and 38%, respectively.	⊕⊕◯◯ Low
Intimacy, Physical comfort, Self-care, Ability to work or volunteer, Ability to concentrate, Emotional health, Social comfort, Ability to sleep (willingness to pay)	2 cross-sectional studies, 99 participants [2,10] Willingness to pay	Not serious	Very serious inconsistency [13]	No serious indirectness	Serious imprecision[12]	none	Willingness to pay for physical comfort, social comfort, emotional health, self-care, intimacy, ability to sleep, ability to work/volunteer and ability to concentration were $2000 to $10,000, $1000 to $2000, $2000 to $5000, $1500 to $9500, $1000 to $5000, $625 to $10000, $1600 to $10000 and $875 to $7500, respectively.	⊕◯◯◯ Very low

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

CI: Confidence interval; DCE: Discrete choice exercise; IQR: Interquartile range; SEM: standard error of means; SG: Standard Gamble; TTO: Time Trade Off; VAS: Visual Analogue Scale.

1. Lundberg 1999 was a study on psoriasis patients using rating scales, time trade off and standard gamble to elicit the utility of outcomes. This study also measured the willingness to pay through a bidding game.
2. Hu 2010 was a study using rating scale to elicit the utility of outcomes. Questions on willingness to pay to eliminate the impairment in particular domain were also asked. The attributes included intimacy, physical comfort, self-care, ability to work or volunteer, ability to concentrate, emotional health, social comfort, ability to sleep.
3. Schmitt 2008 was a study using willingness to pay to measure the preference for controlled and uncontrolled psoriasis.
4. Close estimate from both studies.
5. Zug 1995 was a study measuring utility for mild, moderate and severe psoriasis using visual analogue scale, time trade off, and standard gamble. The study population sample included patients who were seen in the dermatology section at Dartmouth-Hitchcock Medical Center, Lebanon (a tertiary medical center).
6. Only one single study eligible: the generalizability of the result is limited due to highly selected study participants.

7. Small sample size (n=87).
8. Seston 2007 was a discrete choice exercise to ask participants trade-off between outcomes including time to moderate improvement, relapse, and risk of experiencing skin irritation, high blood pressure, liver damage, and skin cancer.
9. In Seston 2007, patients were recruited from the dermatology departments of 3 Acute National Health Service Hospital Trusts located in northwest England. Although no information about the response rate could be reached, the study asked the participants to provide response voluntarily and in general, the study participants were with long durations of psoriasis (a mean of 22.8 years, range from 1 to 63 years).
10. Delfino 2008 was a study measuring the preference through ranking the importance or state the willingness to pay for attributes including intimacy, physical comfort, self-care, ability to work or volunteer, ability to concentrate, emotional health, social comfort, ability to sleep.
11. Ability to sleep, ability to work/volunteer and intimacy were valued differently in the two studies.
12. Both studies had small sample size. In total, the sample size was 99.
13. Very different results from the two studies on the WTP estimates.

Example 5.

What is the importance patients placed on the benign prostatic hyperplasia related health states

Patient or population: benign prostatic hyperplasia

Emberton M: Medical treatment of benign prostatic hyperplasia: physician and patient preferences and satisfaction. International Journal of Clinical Practice 2010, 64:1425-1435.

Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Estimate	Quality
Symptom improvement, decreased prostate size and risks of acute urinary retention and surgery (Direct choice: Trade-off)	1 cross-sectional study 208 participants [1] Discrete choice trade-off	not serious	not serious [2]	Serious indirectness [3,4]	not serious	none	Men would wait longer for symptom improvement in exchange for decreased prostate size (13 months) than they would in exchange for an absolute 1% decrease in the risks of AUR (2 months) and surgery (8 months)	⊕⊕⊕◯ Moderate
Watchful waiting, treatment with an alpha blocker or TURP (Direct choice: Trade-off)	1 cross-sectional study; 87 participants [5] Probability trade off	not serious	not serious [2]	not serious	Serious imprecision[6]	none	More patients rated watchful waiting as a first choice vs. a-blocker therapy (47% vs. 34% respectively)	⊕⊕⊕◯ Moderate
Surgical or non-surgical treatment (Direct choice: proportion)	1 cross-sectional study 635 participants [7] Questionnaire	Serious risk of bias [7]	not serious [2]	not serious	not serious	none	59.4% of patients had a definite or probable preference for non-surgical therapy, while only 9.1% of patients expressing a preference for surgery; however, patients with severe symptoms were more than twice as likely to prefer surgery then those with mild or moderate symptoms.	⊕⊕⊕◯ Moderate
long-term risks of benign prostatic hyperplasia and immediate symptom relief (Direct choice: paired comparison)	1 cross-sectional study 419 participants Questionnaire	Serious risk of bias [8]	not serious [2]	not serious	not serious	none	70% of the men were more worried about long-term risks of BPH than with immediate symptom relief.	⊕⊕⊕◯ Moderate
Reducing progression to surgery, symptom relief (Direct choice: paired comparison)	1 cross-sectional study 502 participants Questionnaire	Serious risk of bias [7]	not serious [2]	not serious	not serious	none	In general, reducing progression to surgery was favoured over symptom relief regardless of whether patients were receiving an a-blocker or 5ARI	⊕⊕⊕◯ Moderate

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

5ARI, 5α-reductase inhibitors; AUR, acute urinary retention.

1. Watson 2010 reported a discrete choice trade-off experiment based on characteristics of a hypothetical a-blocker and 5ARI (pretreatment assessment) on 208 participants. The attributes investigated in the discrete choice experiment were time to symptom improvement, sexual and nonsexual side effects, the risks of acute urinary retention (AUR) and surgery, cost and prostate size decrease.
2. Single study.
3. This study included men aged >= 40 years from the general population, which are not the optimal study population who are facing the decision making.
4. Only one single study eligible: the generalizability of the result is limited.
5. Lkewellyn-Thomas 1996 was a study using probability trade off to determine the preferences for hypothetical treatment options: watchful waiting, treatment with an alpha blocker or TURP.
6. A small sample size
7. Piercy 1999 is a study on the impact of a shared decision-making program on patient preferences. The patients were men with symptomatic benign prostatic hyperplasia, and the preference was measured using a scale, by directly asking what the patients prefer. The validity of this question remained to be validated.
8. The preference was measured using stated preference questions to ask whether the patients consider the attribute essential or very important. The validity of the question remained to be validated.

Example 6.

Evidence profile

Author(s): **Date:** 2016-05-11

Question: What is the patient preference for psychological vs. pharmacological treatment for psychiatric disorders?

Setting: not specified **Bibliography:** McHugh et al. J Clin Psychiatry. 2013;74(6):595-602.

Quality assessment							Value (95% CI)	Quality
Outcome	**Study Design/ Measurement instrument**	**Risk of bias**	**Inconsistency**	**Indirectness**	**Imprecision**	**Other**		
Preference for psychological and pharmacological treatment	34 Cross-sectional study, 68,612 participants [1] Direct choice/forced choice	Very serious[2]	Serious inconsistency [3]	No serious indirectness[4]	No serious imprecision [5]	none	0.75 (0.69 to 0.80) [6] Subgroup estimates Treatment Seeking Samples Only 0.69 (0.61 to 0.77) Samples given > 2 treatment choices 0.75 (0.68 to 0.80) Samples expressing treatment preference for depression only 0.70 (0.62 to 0.77)	◯◯◯◯ Very low

CI: Confidence interval; DCE: Discrete choice exercise; SG: Standard Gamble; TTO: Time Trade Off; VAS: Visual Analogue Scale.

1. 34 cross-sectional studies used a forced-choice assessment of participant preference for type of treatment for a psychiatric disorder; These studies included treatment options with at least one psychological treatment and one medication and the study sample including individuals with a specific psychiatric disorder diagnosis or unselected samples for which participants were asked to identify their treatment preference if they were to be diagnosed with a particular disorder..
2. No formal quality rating was conducted. This is a systematic review on preferences for treatment options, the measurement technique lacked of validity.
3. I^2 and p-value were not shown. However, it seems that studies were pooled (fig.2). There is not description about the type of instrument used to assess the preference. On the other hand, population, intervention, and outcomes seem to be consistent. Subgroup analysis was explored but it was not for explaining inconsistency in the results.
4. Population and alternative options for treatment were chosen properly.
5. It seems like the 95%CI was narrow [0.75 (95% CI: 0.69 to 0.80)].
6. Effect size: 0 indicating “prefer pharmacological treatment”, 1 indicating “prefer psychological treatment”, while 0.5 meaning equal preferences for two options.

Example 7

What is the importance patient placed on colorectal cancer treatment related health states (outcomes)?

Patient or population: patients facing decision making for colorectal cancer treatment

Currie A, Askari A, Nachiappan S, Sevdalis N, Faiz O, Kennedy R: A systematic review of patient preference elicitation methods in the treatment of colorectal cancer. Colorectal Dis 2015, 17:17-25.

Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Estimate	Quality
a permanent stoma	1 cross-sectional study 122 participants (62 APR and 60 LAR patients) utility time trade off [1]	Serious [1,2]	not serious [3]	not serious	not serious [4]	none	median disutility of a permanent stoma: 0.08 for APR patients while 0.37 for LAR patients	⊕⊕◯◯ Low
monthly incontinence	1 cross-sectional study 122 participants (62 APR and 60 LAR patients) utility time trade off [1]	Serious [1,2]	not serious [3]	not serious	not serious [4]	none	median disutility of monthly incontinence: 0.27 for APR patients and 0.19 for LAR patients	⊕⊕◯◯ Low
daily incontinence	1 cross-sectional study 122 participants (62 APR and 60 LAR patients) direct choice (acceptable risk before switching to another treatment) treatment trade off [1]	Serious [1,2]	not serious [3]	not serious	not serious [4]	none	The acceptable risk of daily incontinence (Maximum risk (%) of daily incontinence after LAR patients accept before switching to APR) was higher for LAR patients compared to APR patients (median: 80% vs 10%).	⊕⊕◯◯ Low
preference for avoiding the treatment with a stoma	1 cross-sectional study 99 participants PMPt (mean proportion of remaining life expectancy traded)	Not serious	not serious [3]	Serious indirectness [6]	not serious	none	0.34	⊕⊕⊕◯ Moderate

	[5]							
preference for avoiding the chemotherapy	1 cross-sectional study 97 participants acceptable mortality risk to gamble standard gamble [7]	Serious [8]	not serious [3]	Serious indirectness [6]	not serious	none	the mean of mortality risk gambled was high, 21.4 percent	⊕⊕◯◯ Low
Stoma	3 cross-sectional studies and 1 follow up study 567 participants	Not serious	not serious [10]	not serious	not serious	none	Patients would like to avoid a stoma. Two studies (Bossema et al. and Zolciak et al.]) suggested patients most likely to select LAR, accepting a higher risk of complications to avoid a stoma; while two other studies (Harrison et al. and Solomon et al.) suggested patients would like to reduce or gamble survival to avoid a stoma. Although two studies (Bossema et al. and Zolciak et al.) suggested previous APR meant stoma was viewed less negatively, another study (Harrison et al.) suggested knowing someone with stoma meant APR viewed even more negatively.	⊕⊕⊕⊕ High
	narrative summary [9]							

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

APR: abdominoperineal excision of the rectum; LAR: low anterior resection; PMPt: the Prospective Measure of Preference method

1. Bossema 2008 was a study using time trade off technique to measure the utilities with a permanent stoma, with monthly incontinence or daily incontinence of rectal cancer patients.
2. In Bossema 2008, researchers randomly selected rectal cancer patients, while the study was classified as "high risk of bias" due to low response rate. In total, 129 patients were selected, of those, 60 patients who had undergone APR (146 eligible), 30 patients who had undergone LAR and previously had a temporary stoma (179 initially eligible, but only 91 previously reported faecal incontinence) and 30 patients who had undergone LAR and never had a stoma (112 initially eligible).
3. Only one single study eligible.
4. The utility measurement results were based on a small sample size, there were 62 APR and 60 LAR patients respectively in the study.
5. Harrison 2008 was a study using the Prospective Measure of Preference method (mean proportion of remaining life expectancy traded) to measure the preference of 99 colorectal cancer patients.
6. Patients were recruited after the operation while the study asked their preferences related to the treatment options. Thus, the participants were not the optimal population who are interested in the decision.

7. Solomon 2008 was a study using standard gamble technique to elicit the acceptable mortality risk of patients.
8. The study was classified as high risk of bias due to validity and reliability; researchers used forced choice questions with unproved validity and reliability: 'what kind of operation would you choose for yourself, if this problem concerned you?' Patients after surgery were asked: 'what type of operation would you have chosen for yourself based upon your experiences after the treatment if this problem had concerned you?' The patients were given three options: (i) APR, (ii) AR and (iii) 'I would have left the decision to the surgeon'.
9. This is a narrative summary of results from 3 cross-sectional studies and 1 follow up studies for patients preferences on stoma.
10. Across included studies, patients placed high importance on avoiding a stoma.

Example 8.

What is the importance patients placed on the schizophrenia related health states (outcomes)?

Patient or population: patients facing decision making for schizophrenia

Eiring O, Landmark BF, Aas E, Salkeld G, Nylenna M, Nytroen K: What matters to patients? A systematic review of preferences for medication-associated outcomes in mental disorders. BMJ Open 2015, 5:e007848.

Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Estimate	Quality
Positive, acute or psychotic symptoms	2 cross-sectional studies 147 participants non-utility importance of outcomes rating scales[1]	Serious [1,2]	not serious [3]	not serious	not serious	none	rating results (ranging from 0-5) of 4.031 (1.29) and ranking results (ranging from 0-20) of 11.856 (5.72) for decreased psychotic symptoms in Bridges 2013 15.0 (9.5) for positive symptoms in Shurnway 2003	⊕⊕⊕◯ Moderate
Negative symptoms	2 cross-sectional studies 147 participants non-utility importance of outcomes rating scales[1]	Serious [1,2]	not serious [3]	not serious	not serious	none	rating results (ranging from 0-5) of 4.103 (1.15) and ranking results (ranging from 0-20) of 13.619 (5.59) for decreased depressive thoughts and feelings in Bridges 2013; 11.5 (9.0) for nageative symptoms in Shurnway 2003	⊕⊕⊕◯ Moderate
Negative symptoms	1 cross-sectional study 49 participants Utility paired comparison [4]	Not serious	not serious [5]	not serious	Serious imprecision[6]	none	outpatient, negative symptoms: 0.30	⊕⊕⊕◯ Moderate
Inpatient	1 cross-sectional study 49 participants Utility paired comparison [4]	Not serious	not serious [5]	not serious	Serious imprecision[6]	none	inpatient, acute positive symptoms: 0.19	⊕⊕⊕◯ Moderate

EPS	1 cross-sectional study and 1 RCT (randomized into two different surveys) 151 participants Non-utility importance of outcomes parameter estimates with conjoint analysis, rating scale [7]	Not serious	not serious [8]	not serious	not serious	none	the parameter estimates and their SE were 0.553 (0.153) with D-efficient design, 0.756 (0.162) orthogonal design and for EPS (the larger the estimate, the more important the attribute is) in Kinter 2012; rating result of 13.5 (9.0) for extrapyramidal symptoms in Shurnway 2003	⊕⊕⊕⊕ High
EPS	1 cross-sectional study 50 participants Utility EQ-5D [9]	Not serious	not serious [5]	not serious	Serious imprecision[6]	none	0.72	⊕⊕⊕◯ Moderate

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

EPS: Extrapyramidal side effects

1. Bridges 2013 was a study recruiting hospital and outpatient psychiatrists and outpatients diagnosed with schizophrenia, and using rating scales to assess the relative importance of positive, acute or psychotic symptoms and negative symptoms. In Shurnway 2003, researchers used the standard gamble and rating scales to elicit the relative importance of positive, acute or psychotic symptoms, negative symptoms and extrapyramidal side effects.
2. The validity and reliability of the measurement methodology of Bridges 2013 (the study with larger sample size of the two studies) were not well constructed. The study used a measurement methodology including a rating method (bounded by very important = 5 and not at all important = 1); a ranking method (bounding by most important = 20 and least important = 1); a self-explicated method, estimated by the product of the rating and ranking method.
3. Bridges 2013 and Shurnway 2003 both suggested positive symptoms were important for patients.
4. Revicki 1998 was a study required subjects to provide numeric ratings for 16 descriptions of health states associated with schizophrenia with a rating scale task.
5. Only one single study eligible.
6. Small sample size
7. Kinter 2012 used discrete choice exercise to elicit the preferences from 101 outpatients diagnosed with schizophrenia. During the survey, the participant was presented with a vignette describing two hypothetical individuals diagnosed with schizophrenia who had begun a new treatment 6 months prior and who had very different experiences. In Shurnway 2003, researchers used the standard gamble and rating scales to elicit the relative importance of positive, acute or psychotic symptoms, negative symptoms and extrapyramidal side effects.

8. Both Kinter 2012 and Shurnway 2003 both suggested positive symptoms were important for patients.
9. In Briggs 2008, researchers used EQ-5D to elicit the utility of 49 adult outpatients with a diagnosis of schizophrenia or schizoaffective disorders.

Example 9.

What is the importance patients placed on the substance abuse related health states (outcomes)?

Patient or population: patients facing decision making for substance abuse related therapy

Friedrichs A, Spies M, Harter M, Buchholz A: Patient Preferences and Shared Decision Making in the Treatment of Substance Use Disorders: A Systematic Review of the Literature. PLoS One 2016, 11:e0145817.

Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Estimate	Quality
Alcohol use	1 cross-sectional study 46 participants Direct choice Conjoint analysis [1]	Not serious	not serious [2]	Serious indirectness [3, 4]	Serious imprecision [5]	none	Mean (SE): 2.23 (0.166) at baseline, 2.35 (0.20) at follow up	⊕⊕◯◯ Low
Alcohol use	1 cross-sectional study 156 participants Non-utility importance of outcomes six goal statements [6]	Serious [7]	not serious [2]	Serious indirectness [3, 6]	not serious	none	Regarding drinking goals, 50.7% wanted to reduce their drinking to a nonproblem social level, 34.0% sought no change in their drinking behavior, and 15.4% wanted abstinence or abstinence if controlled drinking would not be a realistic option for them.	⊕⊕◯◯ Low
Cigarette use	1 cross-sectional study 46 participants Direct choice Conjoint analysis [1]	Not serious	not serious [2]	Serious indirectness [3, 4]	Serious imprecision [5]	none	Mean (SE): 0.51 (0.199) at baseline, 0.73 (0.22) at follow up	⊕⊕◯◯ Low
Alcohol use and cigarette use	1 cross-sectional study 46 participants Direct choice Conjoint analysis [1]	Not serious	not serious [2]	Serious indirectness [3, 4]	Serious imprecision [5]	none	The most preferred vignette at baseline was no drinking and smoking half the pretreatment amount, whereas at follow-up the most preferred vignette was no smoking and no drinking. The least preferred vignette at both baseline and follow-up was maintaining the current levels of smoking and drinking... The group-level regression analysis indicated	⊕⊕◯◯ Low

							that stopping alcohol was a stronger priority than stopping tobacco at both the initial and follow-up surveys.	

GRADE Working Group grades of evidence
High quality: We are very confident that the true effect lies close to that of the estimate of the effect
Moderate quality: We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different
Low quality: Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect
Very low quality: We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect

SE: Standard error

1. In Flach 2004, researchers recruited a sample of consecutive patients in the Substance Abuse Treatment Center at a Veteran's Administration Medical Center and used a conjoint analysis to elicit the preferences on vignettes. The vignettes were developed with a full factorial design, with a vignette for each of the nine possible combinations of three levels of cigarette and alcohol use.
2. Only one single study eligible.
3. Only one single study; the applicability of study results was limited due to highly selective study participants.
4. The study participants were enrolled in the Substance Abuse Treatment Center at a Veteran's Administration Medical Center. Subjects had a diagnosis of alcohol dependence based on DSM-IV criteria (Diagnostic and Statistical Manual of Mental Disorders, 1994) and smoked at least 20 cigarettes per day.
5. A small sample size
6. Dillworth 2009 was a study on a community sample to assess the treatment preference. The measurement methodology included six goal statements from the Motivational Information Section of the Comprehensive Drinker Profile. The study attempted to oversample diverse populations, so the researchers advertised in newspapers and areas that catered to ethnic and sexual minorities.
7. The study was classified as high risk of bias due to the unproved reliability and validity. The treatments were rated on a 5-point scale of likelihood to attend (1=extremely unlikely and 5=extremely likely).

Example 10

What is the importance patients placed on the cancer screening related outcomes?

Patient or population: patients facing decision making for cancer screening

Carol Mansfield, Florence K. L. Tangka, Donatus U. Ekwueme, Judith Lee Smith, Jr GPG, Chunyu Li A, Hauber B: Stated Preference for Cancer Screening: A Systematic Review of the Literature, 1990–2013. Prev Chronic Dis 2016, 13.

Outcome	Study Design/ Measurement	Risk of bias	Inconsistency	Indirectness	Imprecision	Other	Estimate	Quality
sensitivity	1 cross-sectional study 656 participants Direct choice Discrete choice exercise [1]	Serious [2]	not serious [3]	Serious indirectness [4]	not serious	none	sensitivity (level of attribute: 50%, 75%, 100%): 1.129 (0.285) for Model with Main Effects Only; in the Model with Interaction Terms, sensitivity interacted with other factors: Sensitivity × Upper Class： 0.624 (0.695), Sensitivity × Highly Educated： 0.935 (0.730)	⊕⊕◯◯ Low
accuracy	1 cross-sectional study 87 participants Direct choice Discrete choice exercise [5]	Serious [6]	not serious [3]	not serious	Serious imprecision [7]	none	accuracy of screening (70%: -0.997, 80%: -0.347, 90%: 0.164, 100%: 1.179)	⊕⊕◯◯ Low
false positive	3 cross-sectional studies 1159 participants Direct choice Discrete choice exercise [1, 8, 9, 10]	serious[1] [1]	not serious [3]	serious indirectness [12]	not serious	none	false positive (level of attribute: 1 in 1000, 1 in 250, 1 in 150 and 1 in 100 for standard Pap test, and 1 in 2000, 1 in 500, 1 in 150 and 1 in 100 for Liquid based Pap test): -0.4504 (Fiebig 2009) Unnecessary colonoscopy (level of attribute: 2%, 4%, 6% of people who take part in screening): − 6.03 (4.424) for Model with Main Effects Only; − 8.703 (5.455) for Model with Interaction Terms (Nayaradou 2010) Chance of a false positive (level of attribute: 8/1000 tested, 15/1000, 20/1000): - 0.070 (0.003) (Salkeld 2000)	⊕⊕◯◯ Low
harms	1 cross-sectional study 301 participants	Not serious	not serious [3]	Serious indirectn	not serious	none	Number of unnecessary colonoscopies (level of attribute: 100, 300, 600 or 800 per CRC death prevented): –0.00013 (0.000006) for	⊕⊕⊕◯ Moderate

	Direct choice Discrete choice exercise [13]			ess [4]			model 1, and –0.00014 (0.000007) for model 2	
cost	3 cross-sectional studies 1159 participants Direct choice Discrete choice exercise [1, 8, 9, 10]	serious[1] [2]	not serious [3]	serious indirectness [12]	not serious	none	cost: -0.0268 (Fiebig 2009) cost: − 0.016 (0.002) for Model with Main Effects Only; − 0.017 (0.003) for Model with Interaction Terms (Nayaradou 2010) Cost of the test kit (COST) ($): - 0.111 (0.006) (Salkeld 2000)	⊕⊕◯◯ Low
GRADE Working Group grades of evidence **High quality:** We are very confident that the true effect lies close to that of the estimate of the effect **Moderate quality:** We are moderately confident in the effect estimate: The true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different **Low quality:** Our confidence in the effect estimate is limited: The true effect may be substantially different from the estimate of the effect **Very low quality:** We have very little confidence in the effect estimate: The true effect is likely to be substantially different from the estimate of effect								

CRC:

1. Nayaradou 2010 was a study using a discrete choice experiment to elicit population preferences for colorectal cancer screening test. Questionnaires were compiled with a set of pairs of hypothetical colorectal cancer screening scenarios.
2. In Nayaradou 2010, the overall response rate after 2 reminders was 32.8% (656/2000, including 483 after the first request, 184 after 1 reminder, and 178 after 2 reminders; however, 189 questionnaires were returned incomplete, reducing the number of exploitable responses from 845 to 656). This rate of response is low in relation to other comparable studies.
3. Only one single study eligible.
4. The study aimed to capture general population preferences. It did not focus on patients actually facing decision making for cancer screening.
5. Gerard 2003 was a study to use a convenience sample of women in the process of attending for breast screening to elicit preferences for future screening. In this study, hypothetical but realistic options for breast screening services were presented to respondents. Respondents were asked a series of binary choices, i.e. whether or not they would present for re-screening in the future if the service was as described.
6. In Gerard 2003, eighty-seven useable surveys were returned, resulting in an overall response rate of 48%.
7. A small sample size.
8. Three studies reported patients' importance on false positive results.
9. Fiebig 2009 was a study testing two samples of women in the target population (previously screened and never-screened women). The study used a discrete choice exercise methodology. In each case there is a choice between a constant reference alternative of no test/no recommendation, a standard Pap test and a liquid based Pap test.
10. Salkeld 2000 recruited a stratified random sample of 600 individuals who had used the bowel scan test kit on at least two occasions in the previous 3 years. A discrete choice experiment was used to look at consumer preferences for a bowel cancer testing kit.
11. High attrition rate in two of the studies, which may have biased the findings.
12. Two studies had patients potentially facing the decision to undergo cancer screening. While for the other with the largest sample size of the three study one (Nayaradou 2010), participants were selected from general population in this study.
13. In Salkeld 2003, 301 participants (138 men and 163 women) completed the discrete choice exercise. The researchers assigned various plausible levels to each of the three attributes. Levels for CRC deaths prevented and colonoscopies due to a false positive test result were based on a systematic review of trial data for CRC screening.

Appendix 5.1. Search strategy

1. PubMed

Search	Query
#12	Search #6 and #7
#11	Search #5 and #7
#10	Search #4 and #7
#9	Search #3 and #7
#8	Search #2 and #7
#7	Search ("Lung Diseases, Obstructive"[Mesh]) OR ("Pulmonary Disease, Chronic Obstructive"[Mesh]) OR (chronic pulmonary obstructive disease[tiab]) OR (COPD[TIAB]) OR (Obstructive Lung Disease[TIAB]) OR (Obstructive Lung Diseases[TIAB]) OR (Obstructive Pulmonary Disease[TIAB]) OR (Obstructive Pulmonary Diseases[TIAB]) OR (chronic pulmonary obstructive diseases[tiAB]) OR (Acute exacerbation of COPD) OR (acute exacerbation of chronic obstructive pulmonary disease) OR (AECB[TIAB]) OR (AECB) OR (COAD) OR (Restrictive Lung Disease[TIAB])
#6	Search (SF36[tiab]) OR (SF 36[tiab]) OR (SF 12[tiab]) OR (SF12[tiab]) OR (HRQoL[tiab]) OR (QoL[tiab]) OR (Quality of life[tiab]) OR ("Quality of Life"[MeSH])
#5	Search (preference based[tiab]) OR (preference score*[tiab]) OR (multiattribute[tiab]) OR (multi attribute[tiab]) OR (EuroQol 5D[tiab]) OR (EuroQol5D[tiab]) OR (EQ5D[tiab]) OR (EQ 5D[tiab]) OR (SF6D[tiab]) OR (SF 6D[tiab]) OR (HUI[tiab]) OR (15D[tiab])
#4	Search (health[ti] AND utilit*[ti]) OR ("Decision Support Techniques"[MeSH]) OR (gamble*[tiab]) OR (prospect theory[tiab]) OR (preference score[tiab]) OR (preference elicitation[tiab]) OR (health utilit*[tiab]) OR (utility value*[tiab]) OR (Utility score*[tiab]) OR (Utility estimate*[tiab]) OR (health state utilit*[tiab]) OR (health state[tiab]) OR (feeling thermometer*[tiab]) OR (best-worst scaling[tiab]) OR (standard gamble[tiab]) OR (time trade-off[tiab]) OR (TTO[tiab]) OR (probability trade-off[tiab]) OR (utility score[tiab])
#3	Search (((decision*[ti] AND mak*[ti]) OR (decision mak*[tiab]) OR (decisions mak*[tiab])) AND (patient*[tiab] OR user*[tiab] OR men[tiab] OR women[tiab])) OR (discrete choice*[tiab]) OR (decision board*[tiab]) OR (decision analy*[tiab]) OR (decision-support[tiab]) OR (decision tool*[tiab]) OR (decision aid*[tiab]) OR (discrete-choice*[tiab]) OR (decision*[tiab] AND (patient*[ti] OR user*[ti] OR men[ti] OR women[ti]) OR (Decision Making[MAJR] AND (patient*[ti] OR user*[ti] OR men[ti] OR women[ti])))
#2	Search ("Attitude to Health"[MAJR]) OR ("Patient Participation"[MAJR]) OR (preference*[tiab]) OR ("Patient Preference"[MAJR]) OR (choice[ti]) OR (choices[ti]) OR (value*[ti]) OR (health state values[tiab]) OR (valuation*[ti]) OR (expectation*[tiab]) OR (attitude*[tiab]) OR (acceptab*[tiab]) OR (knowledge[tiab]) OR (point of view[tiab]) OR (user participation[tiab]) OR (users participation[tiab]) OR (users' participation[tiab]) OR (user's participation[tiab]) OR (patient participation[tiab]) OR (patients' participation[tiab]) OR (patients' participation[tiab]) OR (patient's participation[tiab]) OR (patient perspective*[tiab]) OR (patients perspective*[tiab]) OR (patients' perspective*[tiab]) OR (patient's perspective*[tiab]) OR (patient perce*[tiab]) OR (patients perce*[tiab]) OR (patients' perce*[tiab]) OR (patient's perce*[tiab]) OR (health perception*[tiab]) OR (user view*[tiab]) OR (users view*[tiab]) OR (users' view*[tiab]) OR (user's view*[tiab]) OR (patient view*[tiab]) OR (patients view*[tiab]) OR (patients' view*[tiab]) OR (patient's view*[tiab])

2. Embase

1	preference.mp. or exp patient preference/
2	choice*.ti.

3 value*.ti.
4 health state value*.mp.
5 valuation*.ti.
6 expectation*.mp.
7 attitude*.mp. or exp patient attitude/ or exp attitude to health/
8 acceptab*.mp.
9 knowledge.mp.
10 point of view.mp.
11 user* participation.mp.
12 patient* participation.mp. or exp patient participation/ or exp patient satisfaction/
13 patient* perspective.mp.
14 patient* perce*.mp.
15 health perception*.mp.
16 user* view*.mp.
17 patient* view*.mp.
18 (decision* and mak*).ti.
19 decision* mak*.mp.
20 (patient* or user* or men or women or man or woman).mp. and (18 or 19)
21 (discrete-choice* or discrete choice*).mp.
22 decision board*.mp.
23 decision analy*.mp.
24 (decision-support* or decision support*).mp.
25 exp decision support system/
26 decision tool*.mp. or exp medical decision making/ or exp patient decision making/
27 decision aid*.mp.
28 prospect theory.mp.
29 ("preference score " or "preference elicitation").mp.
30 health utilit*.mp.
31 ("utility value*" or "Utility score*" or "Utility estimate*").mp.
32 health state utilit*.mp. or exp health status indicator/
33 (health and utilit*).ti.
34 health state*.mp.
35 feeling thermometer*.mp. or exp visual analog scale/
36 best-worst scaling.mp.
37 standard gamble.mp.
38 time trade-off.mp.
39 TTO.mp.
40 probability trade-off.mp.
41 utility score*.mp.
42 preference based.mp.
43 preference score*.mp.
44 multiattribute.mp.
45 multi attribute.mp.
46 EuroQol.mp.
47 EQ5D.mp.
48 EQ 5D.mp.
49 (SF-36 or SF 36).mp.
50 SF 6D.mp.
51 SF6D.mp.
52 SF 12.mp.
53 SF12.mp.
54 15 D.mp.
55 HUI.mp.
56 Health Utilit* Index.mp.
57 HRQoL.mp.
58 health related quality of life.mp.

59 quality of life.mp. or exp "quality of life"/
60 or/1-17
61 or/20-27
62 or/28-41
63 (or/42-56) or 29
64 (or/49-54) or (or/57-59)
65 or/60-64
66 exp chronic obstructive lung disease/
67 emphysema$.mp.
68 (chronic$ adj3 bronchiti$).mp.
69 (obstruct$ adj3 (pulmonary or lung$ or airway$ or airflow$ or bronch$ or respirat$)).mp.
70 COPD.mp.
71 COAD.mp.
72 COBD.mp.
73 AECB.mp.
74 or/66-73
75 60 and 74
76 61 and 74
77 62 and 74
78 63 and 74
79 64 and 74
80 65 and 74

3. PsychInfo

1 preference.mp. or exp Preferences/
2 choice*.ti.
3 value*.ti.
4 health state value*.mp.
5 valuation*.ti.
6 expectation*.mp.
7 attitude*.mp. or attitudes/ or exp consumer attitudes/ or exp health attitudes/ or exp "physical illness (attitudes toward)"/ or exp attitude measurement/ or exp attitude measures/ or exp Client Attitudes/
8 acceptab*.mp.
9 knowledge.mp.
10 point of view.mp.
11 user* participation.mp.
12 patient* participation.mp. or exp Client Participation/ or exp Client Satisfaction/
13 patient* perspective.mp.
14 patient* perce*.mp.
15 health perception*.mp.
16 user* view*.mp.
17 patient* view*.mp.
18 (decision* and mak*).ti.
19 decision* mak*.mp.
20 (patient* or user* or men or women or man or woman).mp. and (18 or 19)
21 (discrete-choice* or discrete choice*).mp.
22 decision board*.mp.
23 decision analy*.mp.
24 decision-support.mp.
25 decision support*.mp. or exp Decision Support Systems/
26 decision tool*.mp. or exp Decision Making/
27 decision aid*.mp.
28 prospect theory.mp.

29 ("preference score " or "preference elicitation").mp.
30 health utilit*.mp.
31 ("utility value*" or "Utility score*" or "Utility estimate*").mp.
32 health state utilit*.mp. or exp psychometrics/ or exp Utility Theory/
33 (health and utilit*).ti.
34 health state*.mp.
35 feeling thermometer*.mp. or exp Rating Scales/
36 best-worst scaling.mp.
37 standard gamble.mp.
38 time trade-off.mp.
39 TTO.mp.
40 probability trade-off.mp.
41 utility score*.mp.
42 preference based.mp.
43 preference score*.mp.
44 multiattribute.mp.
45 multi attribute.mp.
46 EuroQol.mp.
47 EQ5D.mp.
48 EQ 5D.mp.
49 (SF-36 or SF 36).mp.
50 SF 6D.mp.
51 SF6D.mp.
52 SF 12.mp.
53 SF12.mp.
54 15 D.mp.
55 HUI.mp.
56 Health Utilit* Index.mp.
57 HRQoL.mp.
58 health related quality of life.mp.
59 quality of life.mp. or exp "quality of life"/
60 or/1-17
61 or/20-27
62 or/28-41
63 (or/42-56) or 29
64 (or/49-54) or (or/57-59)
65 or/60-64
66 exp chronic obstructive lung disease/
67 emphysema$.mp.
68 (chronic$ adj3 bronchiti$).mp.
69 (obstruct$ adj3 (pulmonary or lung$ or airway$ or airflow$ or bronch$ or respirat$)).mp.
70 COPD.mp.
71 COAD.mp.
72 COBD.mp.
73 AECB.mp.
74 66 or 67 or 68 or 69 or 70 or 71 or 72 or 73
75 60 and 74
76 61 and 74
77 62 and 74
78 63 and 74
79 64 and 74
80 65 and 74

4. CINAHL

S99 S94 OR S95 OR S96 OR S97 OR S98
S98 S10 AND S93
S97 S10 AND S87
S96 S10 AND S78
S95 S10 AND S61
S94 S10 AND S49
S93 S88 OR S89 OR S90 OR S91 OR S92
S92 (MH "Quality of Life") OR (MH "Quality of Life (Iowa NOC)") OR (MH "Health and Life Quality (Iowa NOC) (Non-Cinahl)")
S91 TI health related quality of life OR AB health related quality of life
S90 TI HRQol OR AB HRQol
S89 TI SF6D OR AB SF6D OR TI SF12 OR AB SF12 OR TI SF 12 OR AB SF 12
S88 TI SF-36 OR AB SF-36 OR TI SF 36 OR AB SF 36 OR TI SF 6D OR AB SF 6D
S87 S79 OR S80 OR S81 OR S82 OR S83 OR S84 OR S85 OR S86
S86 TI HUI OR AB HUI OR TI Health utilities index OR AB Health utilities index
S85 TI SF6D OR AB SF6D OR TI SF12 OR AB SF12 OR TI SF 12 OR AB SF 12
S84 TI EuroQol OR AB EuroQol OR TI EQ5D OR AB EQ5D OR TI EQ 5D OR AB EQ 5D OR TI SF-36 OR AB SF-36 OR TI SF 36 OR AB SF 36 OR TI SF 6D OR AB SF 6D
S83 TI multi-attribute utility theory OR AB multi-attribute utility theory
S82 TI multi attribute OR AB multi attribute
S81 TI multiattribute OR AB multiattribute
S80 TI preference score* OR AB preference score*
S79 TI preference based OR AB preference based
S78 S62 OR S63 OR S64 OR S65 OR S66 OR S67 OR S68 OR S69 OR S70 OR S71 OR S72 OR S73 OR S74 OR S75 OR S76 OR S77
S77 (MH "Visual Analog Scaling") OR (MH "Behavior Rating Scales")
S76 (MH "Health Status Indicators") OR (MH "Acceptance: Health Status (Iowa NOC)")
S75 TI utility score* OR AB utility score* OR TI utility scale* OR AB utility scale*
S74 TI probability trade off OR AB probability trade off
S73 TI TTO OR AB TTO
S72 TI time trade off OR AB time trade off
S71 TI standard gamble OR AB standard gamble
S70 TI best-worst scaling OR AB best-worst scaling
S69 TI feeling thermometer OR AB feeling thermometer
S68 TI health AND TI utilit*
S67 TI health state utilit* OR AB health state utilit*
S66 TI utility value* OR AB utility value* OR TI utility score* OR AB utility score* OR TI utility estimate* OR AB utility estimate*
S65 TI health utilit* OR AB health utilit*
S64 TI preference elicitation OR AB preference elicitation
S63 TI preference score* OR AB preference score*
S62 TI prospect theory OR AB prospect theory
S61 S52 OR S53 OR S54 OR S55 OR S56 OR S57 OR S58 OR S59 OR S60
S60 (MH "Decision Making") OR (MH "Decision Making, Organizational") OR (MH "Decision Making, Computer Assisted") OR (MH "Decision Making, Patient") OR (MH "Decision Making, Family") OR (MH "Decision Making, Ethical") OR (MH "Decision Making, Clinical")
S59 (MH "Decision Support Systems, Clinical") OR (MH "Decision Support Systems, Management") OR (MH "Decision-Making Support (Iowa NIC)") OR (MH "Decision Support Techniques")
S58 TI decision tool* OR AB decision tool*
S57 TI decision support* OR AB decision support*
S56 TI decision analys* OR AB decision analys*

S55 TI decision aid* OR AB decision aid*
S54 TI decision board* OR AB decision board*
S53 TI discrete choice* OR AB discrete choice*
S52 S50 AND S51
S51 TI patient* OR AB patient* OR TI user* OR AB user* OR TI men OR AB men OR TI women OR AB women OR TI man OR AB man OR TI woman OR AB woman
S50 TI decision* mak* OR AB decision* mak*
S49 S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27 OR S28 OR S29 OR S30 OR S31 OR S32 OR S33 OR S34 OR S35 OR S36 OR S37 OR S38 OR S39 OR S40 OR S41 OR S42 OR S43 OR S44 OR S45 OR S46 OR S47 OR S48
S48 (MH "Consumer Participation")
S47 "patients views or experiences or perceptions" OR (MH "Patient Attitudes")
S46 (MH "Patient Attitudes") OR (MH "Patient Satisfaction")
S45 "patient preference"
S44 TI patient* view* OR AB patient* view*
S43 TI user view* OR AB user view*
S42 TI health perception* OR AB health perception*
S41 TI patient* perception* OR AB patient* perception*
S40 TI patient* perspective OR AB patient* perspective
S39 TI patient* participation OR AB patient* participation
S38 TI user* participation OR AB user* participation
S37 TI point of view OR AB point of view
S36 TI knowledge OR AB knowledge
S35 TI acceptabilit* OR AB acceptabilit*
S34 TI attitude* OR AB attitude*
S33 TI expectation* OR AB expectation*
S32 TI valuation* OR AB valuation*
S31 TI health state value OR AB health state value
S30 TI value*
S29 TI choice
S28 TI preference*
S27 TI patient* view* OR AB patient* view*
S26 TI user view* OR AB user view*
S25 TI health perception* OR AB health perception*
S24 TI patient* perception* OR AB patient* perception*
S23 TI patient* perspective OR AB patient* perspective
S22 TI patient* participation OR AB patient* participation
S21 TI user* participation OR AB user* participation
S20 TI point of view OR AB point of view
S19 TI knowledge OR AB knowledge
S18 TI acceptabilit* OR AB acceptabilit*
S17 TI attitude* OR AB attitude*
S16 TI expectation* OR AB expectation*
S15 TI valuation* OR AB valuation*
S14 TI health state value OR AB health state value
S13 TI value*
S12 TI choice
S11 TI preference*
S10 S1 OR S4 OR S7 OR S8 OR S9
S9 TI emphysema OR AB emphysema
S8 (MH "Emphysema")
S7 S5 AND S6
S6 TI (pulmonary* or lung* or airway* or airflow* or bronch* or respirat*) OR AB (pulmonary* or lung* or airway* or airflow* or bronch* or respirat*)
S5 TI obstruct* OR AB obstruct*

S4 S2 AND S3
S3 TI bronchiti* OR AB bronchiti*
S2 TI chronic* OR AB chronic*
S1 TI COPD OR AB COPD OR TI COAD OR AB COAD OR TI COBD OR AB COBD OR TI AECB OR AB AECB OR TI chronic obstructive pulmonary disease OR AB chronic obstructive pulmonary disease

Appendix Table 5.1. Study characteristics

https://www.dropbox.com/s/s7rpa58cwnt5xrq/3.%20Appendix%20Table%201.%20Study%20characteristics.pdf?dl=0

Appendix Table 5.2. Risk of bias assessment

https://www.dropbox.com/s/p86rglcht9jq0ul/3.%20Appendix%20Table%202.%20Summary%20of%20RoB.xlsx?dl=0

Appendix Table 5.3. Quantitative results

https://www.dropbox.com/s/zuqjodm92dv0kqw/3.%20Appendix%20Table%203.%20Quantitative%20results.pdf?dl=0