BMJ Open

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053246
Article Type:	Protocol
Date Submitted by the Author:	12-May-2021
Complete List of Authors:	Morgano, Gian Paolo; McMaster University, Department of Health Research Methods, Evidence and Impact Mbuagbaw, Lawrence; McMaster University, Department of Health Research Methods, Evidence, and Impact Santesso, Nancy; McMaster University, Department of Health Research Methods, Evidence, and Impact Xie, Feng; McMaster University, Department of Health Research Methods, Evidence, and Impact Brozek, Jan; McMaster University, Department of Health Research Methods, Evidence and Impact Siebert, Uwe; UMIT, Institute of Public Health, Medical Decision Making and Health Technology Assessment; Harvard University, Center for Health Decision Science/Dept. of Health Policy and Management, Harvard TH Chan School of Public Health and MGH-ITA/Dept. of Radiology, Harvard Medical School Bognanni, Antonio; McMaster University, Department of Health Research Methods, Evidence and Impact Wiercioch, Wojtek; McMaster University, Department of Health Research Methods, Evidence, and Impact Piggott, Thomas; McMaster University, Department of Health Research Methods, Evidence and Impact Darzi, Andrea; McMaster University, Department of Health Research Methods, Evidence, and Impact Akl, Elie; American University of Beirut, Clinical Research Institute, Faculty of Medicine Verstijnen, Ilse; Dutch National Health Care Institute, Parmelli, Elena; European Commission Joint Research Centre Ispra Sector, Saz-Parkinson, Zuleika; European Commission Joint Research Centre Alonso-Coello, Pablo; Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Schunemann, Holger; McMaster University, Department of Health Research Methods, Evidence, and Impact
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES

	ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, QUALITATIVE RESEARCH, STATISTICS & RESEARCH METHODS

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study

Authors:

Gian Paolo Morgano[1,2], Lawrence Mbuagbaw[1], Nancy Santesso[1,2], Feng Xie[1], Jan L. Brozek[1,2], Uwe Siebert[3], Antonio Bognanni[1,2], Wojtek Wiercioch[1,2], Thomas Piggott[1,2], Andrea J. Darzi[1,2], Elie Akl[4,5], Ilse Verstijnen[6], Elena Parmelli[7], Zuleika Saz-Parkinson[7], Pablo Alonso-Coello[7], Holger J. Schünemann[1,2,5]

Author Affiliations:

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
2. Michael G. DeGroote Cochrane Canada & McMaster GRADE Centres; Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
3. Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Technology Assessment, UMIT – University for Health Sciences, Medical

Informatics and Technology, Hall i.T., Austria; Division of Health Technology Assessment and Bioinformatics, ONCOTYROL - Center for Personalized Cancer Medicine, Innsbruck, Austria; Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA; Institute for Technology Assessment and Department of Radiology and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

4. Clinical Research Institute, Faculty of Medicine, American University of Beirut, Beirut, Lebanon
5. Department of Medicine, McMaster University, Hamilton, ON, Canada
6. Dutch National Health Care Institute (ZIN), Diemen, The Netherlands
7. European Commission, Joint Research Centre (JRC), Ispra, VA, Italy
8. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain

Corresponding author:

Holger J. Schünemann

Department of Health Research Methods, Evidence and Impact, McMaster University,

1280 Main St West, Hamilton,

ON; L8N 3Z5, Canada

Abstract

Introduction

The GRADE Evidence to Decision (EtD) frameworks require its users to judge how substantial the effects of interventions are on desirable and undesirable people-important health outcomes. However, decision thresholds (DTs) that could help differentiate across judgments and serve as reference for interpretation of findings are not yet available. The objective of this study is an approach to derive and use decision-thresholds (DTs) for EtD judgments about the magnitude of health benefits and harms. We hypothesize that approximate DTs could have the ability to discriminate between the existing four categories of EtD judgments (Trivial, Small, Moderate, Large), support panels of decision-makers in their work, and promote consistency and transparency in judgments.

Methods

We will conduct a methodological randomized controlled trial to collect the data that allow deriving the DTs. We will invite clinicians, epidemiologists, decision scientists, health research methodologists, experts in Health Technology Assessment (HTA), members of guideline development groups and the public to participate in the trial. Then, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Ethics and dissemination

The Hamilton Integrated Research Ethics Board reviewed this protocol. The findings from this study will be disseminated through a publication in a peer-reviewed journal.

Strengths and limitations of this study

- Decision Thresholds will be derived based on empirical data. Therefore, these thresholds may be used in the widely used GRADE Evidence to Decision Frameworks to inform judgments by guideline development group.
- We will use structured case-scenarios to present survey participants with the information relevant to make their judgments. The case-scenarios included effective presentation formats such as the GRADE Summary of Findings tables and health outcomes descriptors that can reduce the risk of error due to an inadequate presentation of data.
- We will employ a randomization process that ensured that case-scenarios were equally distributed across survey participants and that ratings were collected through judgments on outcomes having different values.
- We acknowledge that the survey represents a quite challenging exercise and that this could impact test-retest reliability and applicability of the survey results.

Introduction

As advocated by the National Academy of Medicine of the United States (formerly the Institute of Medicine), the assessment of the benefits and harms of alternative care options (i.e., interventions, actions) is an essential component of any decision-making process underlying guideline recommendations.[1] This assessment should be explicit and include considerations around the probability, magnitude, and importance of health related benefits and health related harms, and other desirable and undesirable consequences of the recommendation or decision.[2] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group has developed the Evidence to Decision (EtD) frameworks to help guideline developers use the evidence in a structured and transparent way and to ensure that they consider all the criteria relevant to their decisions.[3,4] The EtD frameworks require decision-makers to evaluate explicitly the benefits and harms of alternative care options through separate judgments based on the two following questions: *"How substantial are the desirable anticipated effects (health benefits)?"*, *"How substantial are the undesirable anticipated effects (health harms)?"*. To facilitate communication, the GRADE Working Group suggests expressing these judgments by assigning the health benefits or health harms of some intervention under evaluation to one of the following four categories: 'Trivial or None', 'Small', 'Moderate' and 'Large'. To be useful, however, this simplification requires that EtD users have a similar understanding of what magnitude of health benefits or health harms belong into which category and are consistent in their judgments. A similar common understanding is also important between those assigning a category and those interpreting the meaning of a category that is communicated to them (i.e. "imagining" how substantial is an effect based on the category). This can be achieved only when people make similar judgments. To direct EtD users on how to make these judgments appropriately, the GRADE Working Group has produced guidance articles that include the description of the underpinning concepts and examples of judgments based on clinical scenarios.[4,5] Despite the popular use of thresholds to support decision-making in various fields of healthcare research,[6-8] and its adoption in some aspects of the GRADE approach[9,10], use of Decision Thresholds (DTs) for EtD judgments about health benefits and harms is not yet established. For continuous outcomes, EtD users are advised to revert to statistical notions such as Cohen's standardized effect sizes or the Minimal Important Difference (MID) to interpret the magnitude of effects.[11,12] However, empirical data supporting judgments on health benefits and harms for dichotomous outcomes are not yet available for the EtDs.

Objectives

The objective of this study is to derive DTs for EtD judgments on the magnitude of health benefits and harms. We hypothesize that DTs could discriminate between the four categories for EtD judgments. Explicit DTs, providing an indication for which could be the appropriate judgment for a given scenario, might have the potential to support panels of decision-makers in their work, facilitate a common understanding, and promote consistency and transparency in judgments.

Conceptual approach

In the proposed DTs approach, we will consider that judgments on how substantial anticipated effects (health benefits and harms) are should be influenced by: 1) the size of the intervention's effects on each outcome (e.g. the probability of people who experience benefit or harm) 2) the value assigned to those outcomes by the people who are affected.[5] Under this assumption, we will collect data about the association between the dyad composed of size of intervention's effects and value of the outcome on one hand, and judgments on the magnitude of the anticipated effects on the other. In accordance with the EtD frameworks, judgments on desirable effects and on undesirable effects will be collected separately and should not account for any potential tradeoff between benefits and harms. We will use this data to estimate the DTs and provide a conceptual framework for their interpretation and use (see supplemental file 1).

Methods

This study will consist of two parts. In the first part, we will conduct a methodological randomized controlled trial to collect the data that will be used to derive the DTs. Second, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Randomized controlled trial

The following description of methods and analysis of this trial follows the latest guidance by the Standard Protocol Items: Recommendations for Interventional Trials.[13]

Design and setting

Study participants will be recruited to complete a randomized electronic survey (see supplemental file 2) designed to elicit ratings on the magnitude of the potential health effects (benefits or harms) of interventions. Ratings on health benefits and health harms will be collected separately. We will organize the survey into three sections: introduction and example, ratings, questions about respondent demographics. Ratings will be based on five outcomes having a different impact on health (death, major ischemic stroke, pulmonary embolism of moderate severity, diarrhea of moderate severity, and mild nausea/vomiting) presented through descriptive case-scenarios. Each case-scenario will include: (1) a GRADE Summary of Finding (SoF) table[14] providing information about the PICO (Population, Intervention, Comparator, Outcome), the relative and absolute anticipated effects of the intervention, and the certainty in the evidence; (2) a Health Outcome Descriptor [15] describing key attributes of the outcome under consideration; (3) a measure of the impact on health of the outcome (also known as 'value' of the outcome or 'health utility' in health economics). This measure will be expressed on a scale from 0 (being dead) and 1 (perfect health) which means that outcomes with a higher value are valued closer to perfect health as compared to outcomes with a lower value. For each outcome, we will include a case-scenario descriptive of desirable health effects and another one descriptive of undesirable health effects, for a total of ten case-scenarios across five outcomes.

Participants

Selection Criteria

The target population of the survey will include clinicians, epidemiologists, decision scientists, health research methodologists, experts in health technology assessment (HTA), and members of guideline working groups but it will be open to the public too. Prior knowledge of the GRADE approach and experience with the EtD frameworks will be not required for participation.

Recruitment

We will distribute the survey through colleagues, the research group's e-mail lists including that of the Cochrane Collaboration, Guidelines International Network (G-I-N), and of the Global Evidence Synthesis Initiative (GESI). Twitter, LinkedIn, and other social medial platforms will be also used for broader distribution.

Intervention and comparison

Participants will be randomized to a set of 4 case-scenarios, written in lay language, that will be used as intervention (or comparison) in this trial. For each case-scenario, we will ask survey participants to consider the intervention's effects and the value of the outcome and rate how substantial the described health benefits or health harms are. We will also ask them to indicate the lower and upper bound for the ranges of magnitudes of absolute risk difference (ARD) that they associate with the judgments of 'Small' and 'Moderate'. Any estimate below the lower bound for 'Small' will be considered as 'Trivial or None', and any estimate above the upper bound of 'Moderate' will be considered as 'Large'.

Outcomes

The primary endpoints of this trial are the three DTs ($T1=DT_{Trivial/Small}$, $T2=DT_{Small/Moderate}$, $T3=DT_{Moderate/Large}$) that would allow discriminating between EtD judgments of 'Trivial or None' and 'Small', 'Small' and 'Moderate', and 'Moderate' and Large', respectively.

Randomization

Randomization will ensure that case-scenarios will be equally distributed across survey participants to get balanced judgments on outcomes. It will reduce potential confounding due to order effects and possible differences between case-scenarios (e.g. clarity). Randomization will also avoid selection bias that could arise if allowing participants to select the case-scenarios more familiar to them.

Sample size calculation

We based our sample size calculation on the data collected during pilot-testing (n=15 participants). Based on this data, we computed the mean thresholds T1, T2, and T3 for each outcome separately and estimated that we need to recruit 1406 survey respondents to demonstrate a difference of 15% of the mean with non-overlapping 95% confidence intervals. These computations were done using Winpepi. [16]

Statistical methods

Calculation of thresholds from survey ratings

We will use the ranges of ARD for judgments of 'Small' and 'Moderate' collected through the survey to calculate the thresholds associated with each rating. The thresholds will be derived through the product between each the ARD indicated as range boundary and the difference in value from perfect health (1 - outcome's value) for the outcome associated with that rating (see supplemental file 3). We will calculate the DTs as the weighted mean of the corresponding thresholds derived from survey ratings. We will use a weighted mean to account for multiple ratings from the same survey respondent.

Primary analysis

We will use frequencies and percentages to describe the characteristics of survey respondents. For each DT, we will calculate mean, standard deviation (SD), and 95%

confidence intervals (C.I.). Since each participant will contribute data to each threshold, we will employ a paired sample t-test to assess if the DTs are different (T1≠T2≠T3). Our a priori hypothesis is that there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms. All statistical tests will be performed at the 0.05 level of significance.

Sensitivity analyses

Subgroup analysis

We will conduct a subgroup analysis based on participants' characteristics (training in epidemiology, familiarity with the EtD framework, previous participation in guideline development groups). Our a-priori hypotheses is that, in each of the identified subgroups, there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms. All statistical tests will be performed at the 0.05 level of significance.

Incoherent ratings and outliers

We expected that, given the complexity of the topic, some responses might not be internally coherent or represent outliers. We define a threshold as incoherent if T1>T2 OR T2>T3. We define thresholds as outliers if they fall more than three interquartile ranges below the first quartile or above the third quartile. We will verify if the primary analysis would differ if incoherent thresholds or data outliers are excluded. The a priori hypothesis for the sensitivity analyses will be the same as for the primary analysis.

Order effects

We will conduct an ANOVA analysis to assess for potential order effects. We will examine whether participants randomized to a case-scenario for a low-value outcome (outcome value <0.5) in the first case-scenario provided different thresholds as compared to participants who were randomized to a high-value outcome first. Similarly, we will examine whether participants who provided a judgment of 'Small' in the first iteration provided different thresholds as compared to participants who provided a judgment of 'Large' in the first iteration. Our a priori hypothesis is that of no differences if comparing each DT between these groups.

Retrospective comparison of judgments

To investigate the validity of our DTs, we will purposively select judgments from existing guidelines developed using the EtD frameworks and measure the agreement between judgments made by guideline panels and the judgments that our DTs approach would suggest. We will consider for inclusion guidelines reporting the value assigned to outcomes during the decision-making process. We will use frequencies and percentages to describe the agreement. We will employ SPSS v26 (IBM Corp., Armonk, NY) to conduct all statistical analyses.

Pilot testing and assessment of feasibility

To ensure usability and clarity of the survey across respondents having different background or expertise, we piloted the survey with study co-investigators as well as complementary representatives of the target population (n=15). Comments on three iterations of the survey were collected either electronically or by voice recordings and discussed during study meetings. Furthermore, to test the feasibility of the study, we recruited 75 participants from the target population. Participants were able to complete the exercise in the majority of cases. Only 7 out of 75 did not complete the survey after they signed up. Participants contributed a total of 295 ratings with only 17 out of 312 expected ratings missing indicating that the approach to obtaining DTs is feasible. This is true for people of varying backgrounds and educational levels. The findings based on the preliminary analysis of the data support our hypothesis that DTs can help discriminate between the judgments (see supplemental file 3).

Discussion

We believe that DTs for judgments on desirable and undesirable health effects can be useful to decision-makers using the EtD frameworks. Guideline panels using the GRADE EtDs often ask what are 'Trivial or None', 'Small', 'Moderate' and 'Large' effects. The proposed DTs approach could provide an answer based on empirical data and be used to initiate and promote discussion. Furthermore, it is simple to apply, and requires only to calculate the product between ARD and the reduction in value associated with the outcome. This endeavor will expand the research on the use of decision thresholds within the GRADE methodology and could be integrated into GRADEpro.

Our work with Hultcrantz et al.[12] suggests that clinical decision thresholds can be used to allow appropriate ratings of the certainty of the evidence, but there is no empirical data. Furthermore, it focuses on the construct of certainty of evidence and targets different degrees of contextualization, while we address judgments on the magnitude of effects and made by users of the EtD frameworks. The joint consideration of the estimate of effect and outcome's importance has been already adopted in another effort of the GRADE Working Group. In a GRADE concept paper[17], Alper et al. aim to define the certainty in the net benefit and suggest calculating the net effect of an intervention by combining importance-adjusted effect estimates calculated from different outcomes. While this strategy is appealing and would allow us to apply our research to EtD judgments on the trade-off between benefits and harms, further research is needed to establish if the estimates to be combined are independent and not correlated with each other. Other quantitative approaches to assess the benefits, harms, and net benefit associated with treatments are available in the literature[18], but none aims to characterize the magnitude of effects into categories (i.e 'Trivial or None', 'Small', 'Moderate', 'Large') as needed to make judgments using the EtD frameworks. Utilitarian frameworks are common in health economic research, where health-utilities elicited from target populations are used to inform modeling techniques such as cost-effectiveness analysis based on quality-adjusted-life-years (QALY). [19,20]

Ethics and dissemination

After review, the Hamilton Integrated Research Ethics Board (HiREB) determined that as a quality improvement project, this study was exempt from formal ethics review as per TCPS2 (2014) Article 2.5. We will inform respondents of this decision and the anonymous nature of the study. The results of this randomized trial will be published in a peer-reviewed journal. We also aim to present the results in national and international conferences.

Availability of data and materials

Data will be available on request from the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Roles and responsibilities

HJS is the principal investigator of the study and together with GPM, NS, FX, and JB designed and established this research project. ML, IV, TP, EP, ZSP, and US, piloted the survey and provided methodological input. GPM was responsible for the ethics application and for registering the protocol on the website, www.ClinicalTrials.gov. HJS, GPM and LM designed the statistical analysis. GPM, AB, WW, AD are responsible for recruitment. HJS and GPM are responsible for the coordination of the study. HJS, GPM, NS, FX, and JB drafted the manuscript. All authors participated in the writing and revision of the manuscript, and approved the its final version.

References

1. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice G. *Clinical Practice Guidelines We Can Trust.* Washington (DC): National Academies Press (US); 2011.
2. Schünemann HJ. Guidelines 2.0: do no net harm-the future of practice guideline development in asthma and other diseases. *Curr Allergy Asthma Rep.* 2011;11(3):261-268.
3. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res.* 2004;4(1):38. doi: 10.1186/1472-6963-1184-1138.
4. Alonso-Coello P, Schunemann 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.* 2016;353:i2016.
5. 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.* 2016;353:i2089.
6. Marcucci M, Sinclair JC. A generalised model for individualising a treatment recommendation based on group-level evidence from randomised clinical trials. *BMJ Open.* 2013;3(8).
7. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics.* 2008;26(9):733-744.
8. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med.* 1975;293(5):229-234.

9. Hultcrantz M, Mustafa RA, Leeflang MMG, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. *J Clin Epidemiol.* 2020;117:138-148.

10. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol.* 2017;87:4-13.

11. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. *P T.* 2008;33(12):700-711.

12. Johnston BC, Ebrahim S, Carrasco-Labra A, et al. Minimally important difference estimates and methods: a protocol. *BMJ Open.* 2015;5(10):e007953.

13. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *Bmj.* 2013;346:e7586.

14. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol.* 2013;66(2):173-183. doi: 110.1016/j.jclinepi.2012.1008.1001. Epub 2012 Oct 1030.

15. Baldeh T, Saz-Parkinson Z, Muti P, et al. Development and use of health outcome descriptors: a guideline development case study. *Health Qual Life Outcomes.* 2020;18(1):167.

16. Abramson JH. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. *Epidemiol Perspect Innov.* 2011;8(1):1.

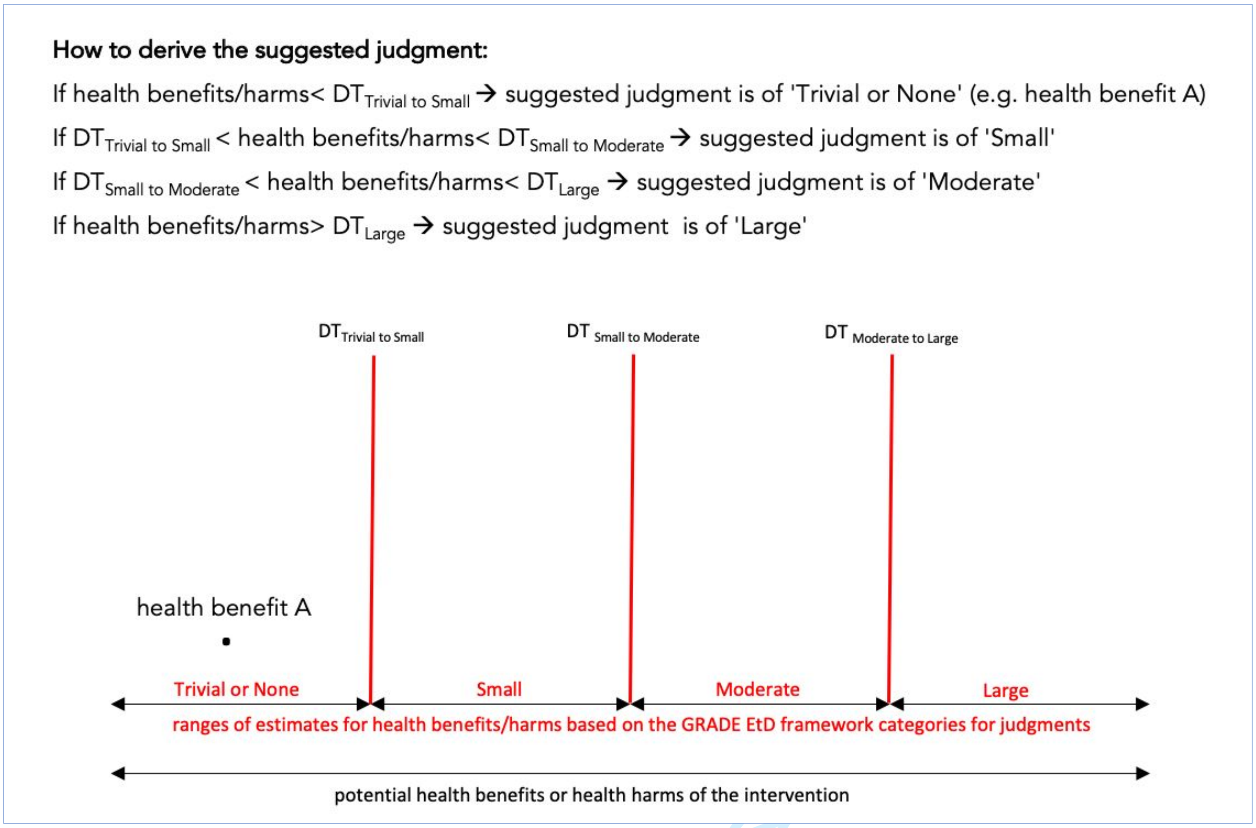
17. Alper BS, Oettgen P, Kunnamo I, et al. Defining certainty of net benefit: a GRADE concept paper. *BMJ Open.* 2019;9(6):e027445.

18. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol.* 2012;12:173.

19. Siebert U. When should decision-analytic modeling be used in the economic evaluation of health care? *The European Journal of Health Economics, formerly: HEPAC.* 2003;4(3):143-150.

20. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;296(13):716-721.

Supplement file 1



How to derive the suggested judgment:

If health benefits/harms< $DT_{\text{Trivial to Small}}$ → suggested judgment is of 'Trivial or None' (e.g. health benefit A)

If $DT_{\text{Trivial to Small}}$ < health benefits/harms< $DT_{\text{Small to Moderate}}$ → suggested judgment is of 'Small'

If $DT_{\text{Small to Moderate}}$ < health benefits/harms< DT_{Large} → suggested judgment is of 'Moderate'

If health benefits/harms> DT_{Large} → suggested judgment is of 'Large'

$DT_{\text{Trivial to Small}}$
$DT_{\text{Small to Moderate}}$
$DT_{\text{Moderate to Large}}$
health benefit A
Trivial or None
Small
Moderate
Large
ranges of estimates for health benefits/harms based on the GRADE EtD framework categories for judgments
potential health benefits or health harms of the intervention

Figure 1

Figure 1 legend: The availability of three DTs ($DT_{\text{Trivial/Small,}}$ $DT_{\text{Small/Moderate,}}$ $DT_{\text{Moderate/Large}}$) would allow to discriminate between the four GRADE EtD framework categories for judgments. For a given health benefit/harm, the suggestion on the judgment would depend on how the estimate of health benefits/harms compares to the DTs. In this example, the health benefit A lies on the left (is smaller) of the $DT_{\text{Trivial/Small}}$ which would suggest that the judgment of 'Trivial or None' would be more appropriate than the others.

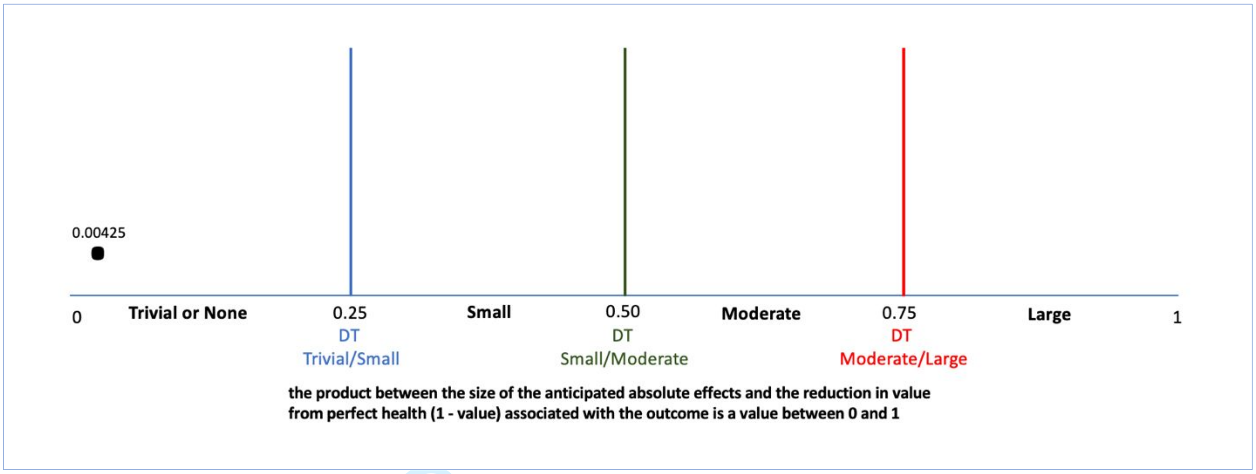


Figure 2

Figure 2 legend: We assumed to have known DTs ($DT_{Trivial/Small}$ = 0.25, $DT_{Small/Moderate}$ = 0.50, $DT_{Moderate/Large}$ = 0.75) and wanted to assign to one of the 4 EtD categories the health benefit of an intervention showing an anticipated absolute effect of 17 fewer per 1000 on an outcome valued 0.75. Following the proposed approach, we calculated the result of the product (score) of the size of anticipated effects (Absolute Risk Difference, ARD) and the reduction in value from perfect health (1 - outcome's value) associated with the outcome under evaluation. In this example, the following approach (ARD * (1 - outcome's value) = (17/1000)*(1 - 0.75)) resulted in the value of 0.00425. We then plotted this value and obtained the suggested judgment according to the DTs approach that, in this case, would be of 'Trivial or None' considering that the calculated value is smaller than the $DT_{Trivial/Small}$. In case of judgments made considering more than one outcome, it would be required to calculate the aforementioned product for each of the outcomes under evaluation and derive an aggregate score defined as the sum of all the individual scores.

Under the assumption that the outcomes are independent, this score could be then compared to the DTs and used to obtain a suggestion on the judgment.

Supplemental file 2

Descriptive statistics

The presented preliminary analysis is based on survey data collected between May 1st and July 21st, 2020. Our dissemination strategy allowed recruitment of 75 participants who contributed a total of 295 ratings. Fifty-six survey participants had a background in research (74.6%) and 36 were healthcare professionals (50.6%). Thirty-four respondents (45.3%) were members of academia. Other major groups were participants from HTA organizations and professional societies (13.3% and 18.6%, respectively). Participants were equally randomized to case-scenarios descriptive of desirable and undesirable health effects (144/295, 49%; 151/295, 51%, respectively) and completed the entire exercise in the majority of cases (68/75, 90.7%). Detailed descriptive characteristics of survey respondents and ratings are shown in Tables 1 and 2, respectively.

Characteristic [a]	**Respondents, n = 75**
Background [a]	n (%)
Clinical/Health Professional	38 (50.6)
Policymaking	6 (8.0)
Research	56 (74.6)
Teaching	18 (24.0)
Administrative	3 (4.0)

Patient representative	2 (2.6)
Other	3 (4.0)
Degree [a]	
Degree in Nursing (RN)	1 (1.3)
Medical School (MD)	30 (4.0)
Master of Sciences (MSc)	17 (22.6)
Master of Public Health (MPH)	9 (0.12)
Doctor of Philosophy (PhD)	25 (33.3)
None	2 (2.6)
Other	5 (6.6)
Formal Training in health research methodology/epidemiology/biostatistics	
Never completed	12 (16.0)
Completed some form of formal training but do not have a graduate degree	30 (40.0)
Earned a MSc degree	16 (21.3)
Earned a PhD degree	16 (21.3)
Not available	1 (1.4)
Organization [a]	
Cochrane collaboration	13 (17.3)
GRADE Working Group	16 (21.3)

World Health Organization	1 (1.4)
Guidelines International Network (G-I-N)	-
Health Technology Assessment (HTA) organization	10 (13.3)
Academia	34 (45.3)
Professional society	14 (18.6)
Familiarity with the Evidence to Decision framework	
Not at all familiar	5 (6.6)
Not so familiar	9 (12.0)
Somewhat familiar	16 (21.3)
Very familiar	30 (40.0)
Extremely familiar	8 (10.6)
Not available	7 (9.5)
Previous participation in guideline development groups	
Yes	52 (69.3)
No	18 (24.0)
Not available	5 (6.6)
Primary role in the guideline development group [a]	
Clinical Chair	5 (6.6)
Chair for methods	15 (19.8)
Guideline methodologist	29 (38.6)
Panel member	15 (19.8)

Topic or content expert	7 (9.5)
Patient representative	2 (2.6)
Systematic review author	26 (34.6)
Expert in Health Technology Assessment	3 (4.0)

Values represent the number and in parentheses the percentage.

[a] Percentages do not add up to 100 because respondents could choose more than one option.

Table 1: Characteristics of survey respondents

Characteristics of ratings collected through the survey	n (%)
Total number of ratings collected	295
Missing data (expected ratings - collected ratings/expected ratings)	17/312 (0.054)[a]
randomized to a scenario showing desirable effects	144/295 (49)
randomized to a scenario showing undesirable effects	151/295 (51)
randomized to the outcome of death	73/295 (25)
randomized to the outcome of major stroke	66/295 (22)
randomized to the outcome of pulmonary embolism	55/295 (19)
randomized to the outcome of moderate diarrhea	63/295 (21)
based on the outcome of mild nausea/vomiting	38/295 (13)

a. 73 participants were randomized to 4 case-scenarios, 2 were mistakenly randomized to 10.

Table 2: Descriptive statistics of survey ratings

Table 3 describes the estimates of DTs that were derived from survey ratings through the joint measure of absolute effects and outcome values. For example, an outcome valued as 0.8, these thresholds would indicate that the effect of an intervention preventing 30 events of that outcome per 1000 should be categorized as trivial (since 0.03*(1-0.8)) =0.006 is smaller than T1). More details about the calculation of the DTs are available in Appendix 1 (Table 1).

Decision Threshold			**95% Confidence Interval**	
	Estimate	Std. Deviation	Lower Bound	Upper Bound
T1: Trivial/Small	0.0165	0.0467	0.0059	0.0271
T2: Small/Moderate	0.0312	0.0601	0.0176	0.0448
T3: Moderate/Large	0.0577	0.0781	0.0400	0.0754

Table 3: Estimates of DTs

Primary analysis

Our analysis showed a difference in the estimates between T1 and T2 (mean difference [MD] -0.0147; 95% CI -0.0201 to -0.0093; p<0.001) and T2 and T3 (mean difference [MD] -0.0264; 95% CI -0.0544 to -0.0062; p<0.001).

Within-participant analyses

The analyses showed that at a respondent level there was no difference between DTs derived from judgments on benefits and from those on harms: $T1_{benefit}=T1_{harms}$ (mean difference [MD] -0.0040 ; 95% CI -0.0195 to 0.0116 ; p=0.615) ; $T2_{benefits}=T2_{harms}$ (mean difference [MD] -0.0124; 95% CI -0.0313 to 0.0064 ; p=0.196); $T3_{benefit}=T3_{harms}$ (mean difference [MD] -0.0209; 95% CI -0.0451 to 0.0033; p=0.090).

Subgroup analyses

Our subgroup analyses showed a difference in the estimates between T1 and T2, and T2 and T3 also in DTs derived from subgroup of ratings identified by outcome, direction of interventions' effects, and prior participation to guideline development groups. No difference was observed in the estimates between T1 and T2 in those with no experience with the EtD (mean difference [MD] -0.0046; 95% CI -0.0100 to 0.0006; p=0.810) and between T2 and T3 in those who had no training in epidemiology (mean difference [MD] -0.0056; 95% CI -0.0218 to 0.0106; p=0.483).

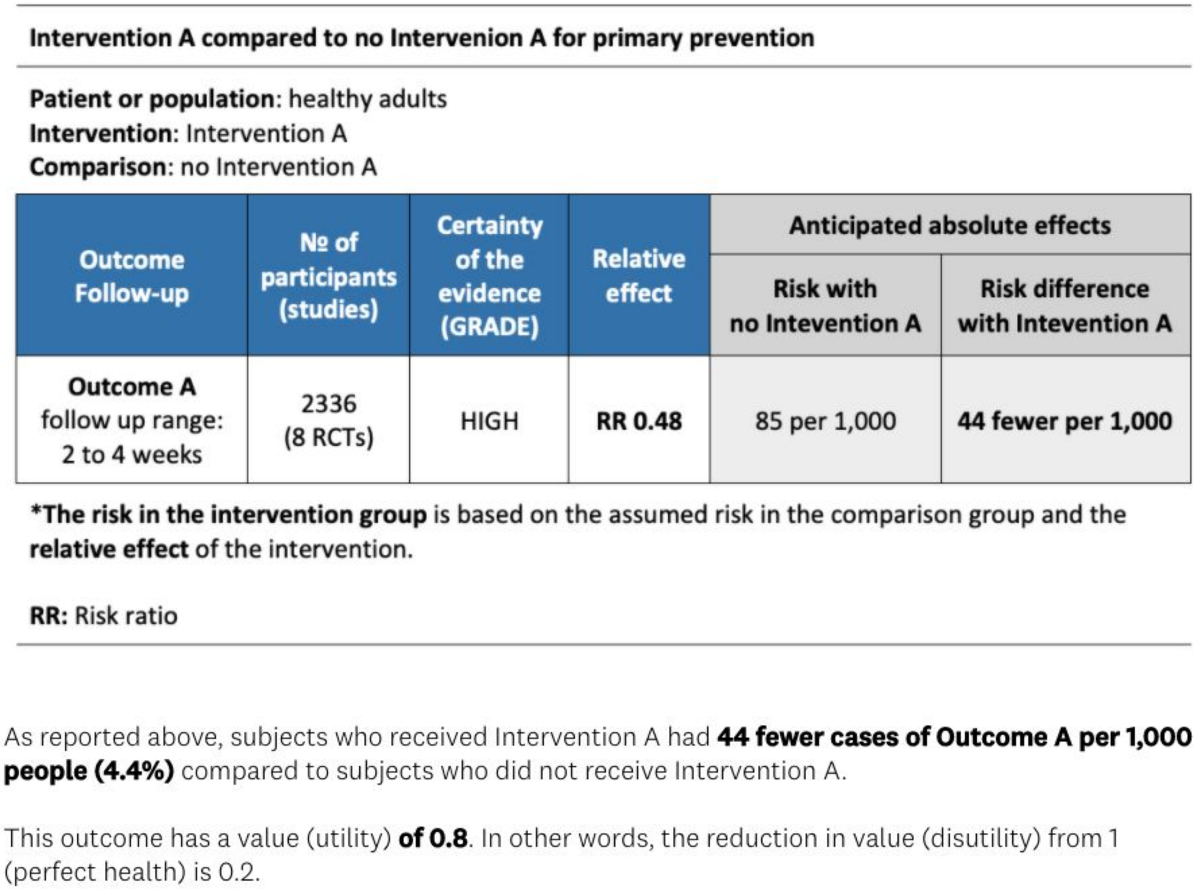
Sensitivity analyses

The findings of the sensitivity analyses conducted by excluding raters who provided incoherent thresholds (n=3; T1/T2 mean difference [MD] -0.0143; 95% CI -0.0192 to -0.0094; p<0.001; T2/T3 mean difference [MD] -0.0291; 95% CI -0.0417 to -0.0165; p<0.001) or who were presumed outliers (n=10; T1/T2 mean difference [MD] -0.0096 ; 95% CI -0.0113 to -0.0078; p<0.001; T2/T3 mean difference [MD] -0.0194; 95% CI -0.0240 to -0.0148; p<0.001) were similar to that of the primary analysis.

Assessment of order effects

The analyses suggest no difference between DTs derived from participants who evaluated a high-value outcome (i.e. moderate diarrhea) in the first iteration compared to those who evaluated a low-value outcome (i.e. death) first. Similarly, there was no difference in the DTs depending on whether the first judgment made was 'Small' or 'Large'.

Supplemental file 3



Intervention A compared to no Intervenion A for primary prevention

Patient or population: healthy adults
Intervention: Intervention A
Comparison: no Intervention A

Outcome Follow-up	№ of participants (studies)	Certainty of the evidence (GRADE)	Relative effect	Anticipated absolute effects	
				Risk with no Intevention A	Risk difference with Intevention A
Outcome A follow up range: 2 to 4 weeks	2336 (8 RCTs)	HIGH	**RR 0.48**	85 per 1,000	**44 fewer per 1,000**

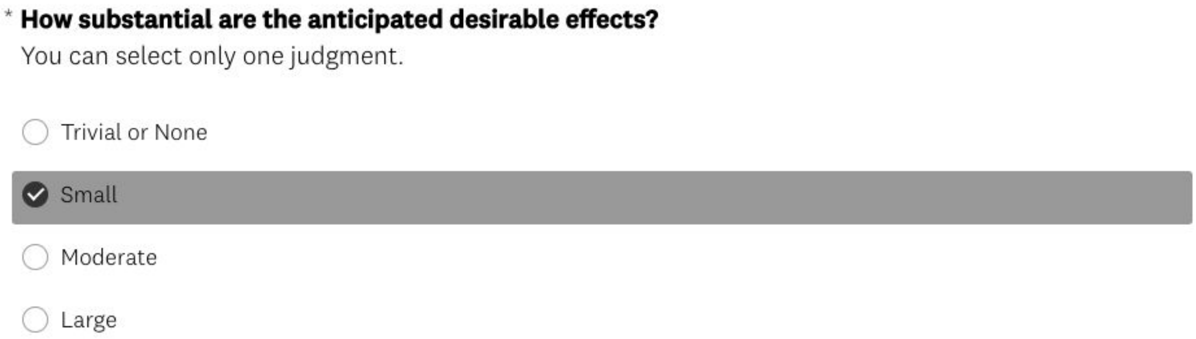
***The risk in the intervention group** is based on the assumed risk in the comparison group and the **relative effect** of the intervention.

RR: Risk ratio

As reported above, subjects who received Intervention A had **44 fewer cases of Outcome A per 1,000 people (4.4%)** compared to subjects who did not receive Intervention A.

This outcome has a value (utility) **of 0.8**. In other words, the reduction in value (disutility) from 1 (perfect health) is 0.2.

Figure 1 - Example of a case-scenario



* **How substantial are the anticipated desirable effects?**

You can select only one judgment.

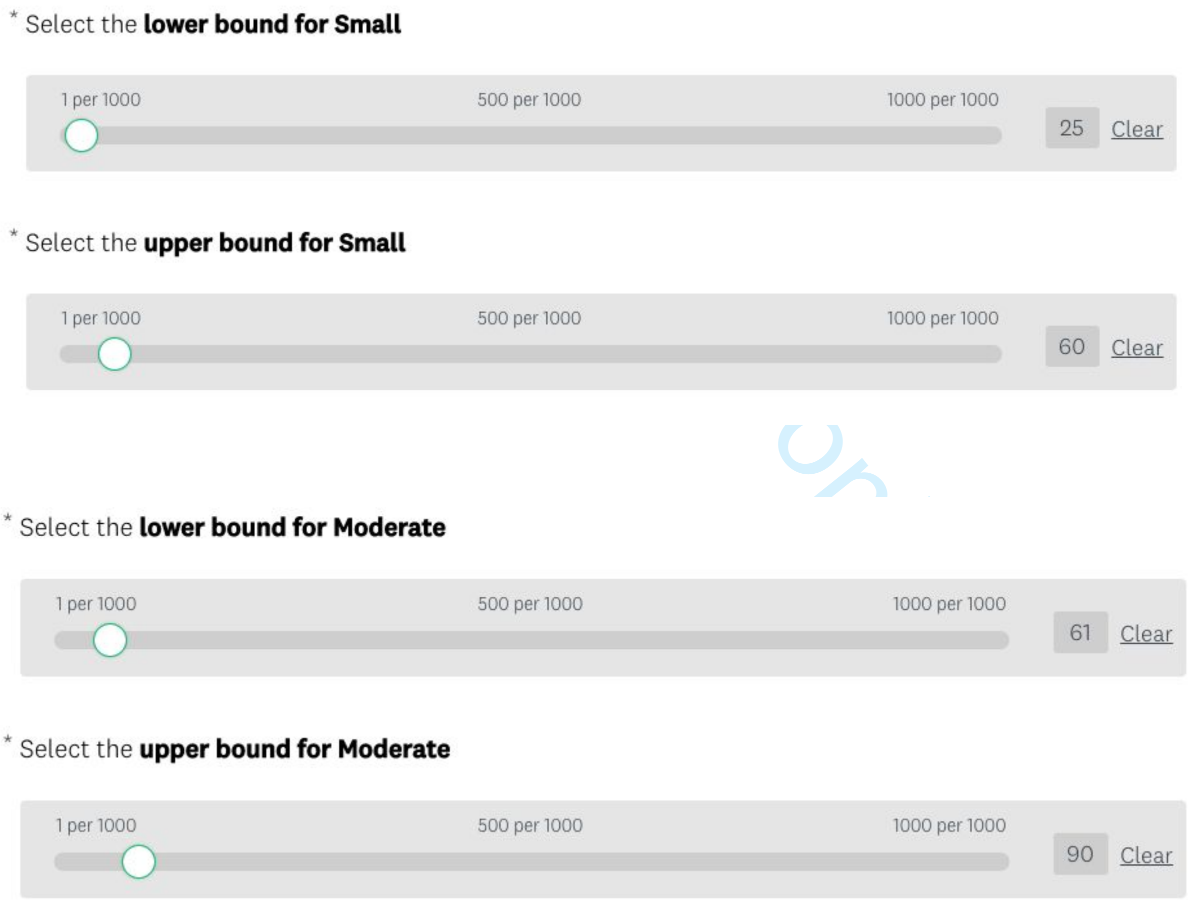
○ Trivial or None

☑ Small

○ Moderate

○ Large

Figure 2 - Judgment on health benefits



* Select the **lower bound for Small**

25

* Select the **upper bound for Small**

60

* Select the **lower bound for Moderate**

61

* Select the **upper bound for Moderate**

90

Figure 3 - Selection of ranges for judgments of Small and Moderate

Example of calculation of DTs based on survey data

In the examples shown in Figures 2 and 3 we assumed that, after having evaluated a given case-scenario (ARD of 44 events fewer per 1000 on an outcome valued 0.8), a survey participant rated the hypothetical ranges of ARD for judgments of 'Small' and 'Moderate' of from 25 fewer per 1000 to 60 fewer per 1000, and of from 61 fewer per 1000 to 90 fewer per 1000, respectively. We used this data to derive the ranges of ARD for judgments of 'Trivial or None' and of 'Large' (table below).

boundaries of ranges described in Figure 3 value of the outcome = 0.8						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	range
0 per 1000	24 per 1000	25 per 1000	60 per 1000	61 per 1000	90 per 1000	more than 90 per 1000

Table 1 - Ranges of sizes of effects (ARD)

For each range of ARD, we calculated the product between range boundaries and the reduction in value from perfect health (1 - outcome's value) for the outcome associated. Then, we derived the DTs as follow: DT$_{Trivial/Small}$ equal to the the product calculated from the lower bound for the judgment of 'Small', DT$_{Small/Moderate}$ equal to average of the products calculated from the upper bound for the judgment of 'Small' and the lower bound for the judgment of 'Moderate', and DT$_{Moderate/Large}$ equal to the smallest number larger than the mean of the products calculated from the upper bound for the judgment of 'Moderate'.

product values = ARD * (1- outcome's value))						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	any value
(0/1000)*0.2 0	(24/1000) *0.2 0.0048	(25 /1000) *0.2 0.005	(60/1000) *0.2 0.012	(61/1000)*0.2 0.0122	(90/1000)*0.2 0.018	bigger than (90/1000)*0.2 >0.018

Table 2 - Ranges of product values

Using the data from Table 2, the DTs would result as follow: DT$_{Trivial/Small}$ = 0.005, DTs$_{mall/Moderate}$ = 0.0121$_{,}$ DTModerate$_{/Large\ =}$ 0.0180001.

BMJ Open

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study (GRADE-THRESHOLD)

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053246.R1
Article Type:	Protocol
Date Submitted by the Author:	14-Dec-2021
Complete List of Authors:	Morgano, Gian Paolo; McMaster University, Department of Health Research Methods, Evidence and Impact Mbuagbaw, Lawrence; McMaster University, Department of Health Research Methods, Evidence, and Impact Santesso, Nancy; McMaster University, Department of Health Research Methods, Evidence, and Impact Xie, Feng; McMaster University, Department of Health Research Methods, Evidence, and Impact Brozek, Jan; McMaster University, Department of Health Research Methods, Evidence and Impact Siebert, Uwe; UMIT, Institute of Public Health, Medical Decision Making and Health Technology Assessment; Harvard University, Center for Health Decision Science/Dept. of Health Policy and Management, Harvard TH Chan School of Public Health and MGH-ITA/Dept. of Radiology, Harvard Medical School Bognanni, Antonio; McMaster University, Department of Health Research Methods, Evidence and Impact Wiercioch, Wojtek; McMaster University, Department of Health Research Methods, Evidence, and Impact Piggott, Thomas; McMaster University, Department of Health Research Methods, Evidence and Impact Darzi, Andrea; McMaster University, Department of Health Research Methods, Evidence, and Impact Akl, Elie; American University of Beirut, Clinical Research Institute, Faculty of Medicine Verstijnen, Ilse; Dutch National Health Care Institute, Parmelli, Elena; European Commission Joint Research Centre Ispra Sector, Saz-Parkinson, Zuleika; European Commission Joint Research Centre Alonso-Coello, Pablo; Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Schünemann, Holger J ; McMaster University, Department of Health Research Methods, Evidence and Impact
<b>Primary Subject Heading</b>:	Evidence based practice

Secondary Subject Heading:	Communication, Epidemiology, Health informatics, Health policy, Health services research
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study (GRADE-THRESHOLD)

Authors:

Gian Paolo Morgano[1,2], Lawrence Mbuagbaw[1], Nancy Santesso[1,2], Feng Xie[1], Jan L. Brozek[1,2], Uwe Siebert[3], Antonio Bognanni[1,2], Wojtek Wiercioch[1,2], Thomas Piggott[1,2], Andrea J. Darzi[1,2], Elie A. Akl[1,4], Ilse Verstijnen[5], Elena Parmelli[6], Zuleika Saz-Parkinson[6], Pablo Alonso-Coello[7], Holger J. Schünemann[1,2,8,9]

Author Affiliations:

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
2. Michael G. DeGroote Cochrane Canada & McMaster GRADE Centres; Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
3. Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Technology Assessment, UMIT – University for Health Sciences, Medical

Informatics and Technology, Hall i.T., Austria; Division of Health Technology Assessment and Bioinformatics, ONCOTYROL - Center for Personalized Cancer Medicine, Innsbruck, Austria; Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA; Institute for Technology Assessment and Department of Radiology and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

4. Department of Internal Medicine, American University of Beirut, Beirut, Lebanon
5. Dutch National Health Care Institute (ZIN), Diemen, The Netherlands
6. European Commission, Joint Research Centre (JRC), Ispra, VA, Italy
7. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
8. Department of Medicine, McMaster University, Hamilton, ON, Canada
9. Dipartimento di Scienze Biomediche Humanitas University, Milan, Italy

Corresponding author:

Prof. Holger J. Schünemann
Department of Health Research Methods, Evidence and Impact, McMaster University,
1280 Main St West, Hamilton,
ON; L8N 3Z5, Canada
schuneh@mcmaster.ca

Abstract

Introduction

The GRADE Evidence to Decision (EtD) frameworks require its users to judge how substantial the effects of interventions are on desirable and undesirable people-important health outcomes. However, decision thresholds (DTs) that could help differentiate across judgments and serve as reference for interpretation of findings are not yet available.

The objective of this study is an approach to derive and use decision-thresholds (DTs) for EtD judgments about the magnitude of health benefits and harms. We hypothesize that approximate DTs could have the ability to discriminate between the existing four categories of EtD judgments (Trivial, Small, Moderate, Large), support panels of decision-makers in their work, and promote consistency and transparency in judgments.

Methods and analysis

We will conduct a methodological randomized controlled trial to collect the data that allow deriving the DTs. We will invite clinicians, epidemiologists, decision scientists, health research methodologists, experts in Health Technology Assessment (HTA), members of guideline development groups and the public to participate in the trial. Then, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Ethics and dissemination

The Hamilton Integrated Research Ethics Board reviewed this study as a quality improvement study and determined that it requires no further consent. Survey participants will be required to read a consent statement in order to participate in this study at the beginning of the trial. This statement reads: You are being invited to participate in a research project which aims to identify indicative decision thresholds that could assist users of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks in making judgments. Your input will be used in determining these indicative thresholds. By completing this survey, you provide consent that the anonymized data collected will be used for the research study and to be summarized in aggregate in publication and electronic tools.

Article summary

Strengths and limitations of this study

- The calculation of the Decision Thresholds will be based on empirical data.
- We will use structured case-scenarios to present survey participants with the information relevant to make their judgments.
- We will employ a randomization process to ensure that case-scenarios will be equally distributed across survey participants
- We acknowledge that the survey requires effort and that this could impact test-retest reliability and applicability of the survey results which we overcome in part by conducting a large trial
-

Introduction

As advocated by the National Academy of Medicine of the United States (formerly the Institute of Medicine), the assessment of the benefits and harms of alternative care options (i.e., interventions, actions) is an essential component of any decision-making process underlying guideline recommendations.[1] This assessment should be explicit and include considerations around the probability, magnitude, and importance of health related benefits and health related harms, and other desirable and undesirable consequences of the recommendation or decision.[2] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group has developed the Evidence to Decision (EtD) frameworks to help guideline developers use the evidence in a structured and transparent way and to ensure that they consider all the criteria relevant to their decisions.[3,4] The EtD frameworks require decision-makers to evaluate explicitly the benefits and harms of alternative care options through separate judgments based on the two following questions: *"How substantial are the desirable anticipated effects (health benefits)?"*, *"How substantial are the undesirable anticipated effects (health harms)?"*. The guidance from the GRADE Working Group includes expressing and facilitating these judgments by assigning the health benefits or health harms of some intervention under evaluation to one of the following four categories: 'Trivial or None', 'Small', 'Moderate' and 'Large'.[3,4] To be useful, however, this simplification requires that EtD users have a similar understanding of what magnitude of health benefits or health harms belong into which category and are consistent in their judgments. A similar common understanding is also important between those assigning a category and those interpreting the meaning of a category that is communicated to them (i.e. “imagining” how substantial is an effect based on the category). This can be achieved only when people make similar judgments. To direct EtD users on how to make these judgments appropriately, the GRADE Working Group has produced guidance articles that include the description of the underpinning concepts and examples of judgments based on clinical scenarios.[4,5] Despite the popular use of thresholds to support decision-making in various fields of healthcare research,[6-8] and its adoption in some aspects of the GRADE approach[9,10], use of Decision Thresholds (DTs) for EtD judgments about health benefits and harms is not yet established. For continuous outcomes, EtD users are advised to revert to statistical notions such as Cohen's standardized effect sizes or the Minimal Important Difference (MID) to interpret the magnitude of effects.[11,12] However, empirical data supporting judgments on health benefits and harms for dichotomous outcomes are not yet available for the EtDs.

Objectives

The objective of this study is to derive DTs for EtD judgments on the magnitude of health benefits and harms. We hypothesize that DTs could discriminate between the four categories for EtD judgments. Explicit DTs, providing an indication for which could be the appropriate judgment for a given scenario, might have the potential to support panels of decision-makers in their work, facilitate a common understanding, and promote consistency and transparency in judgments.

Conceptual approach

In the proposed DTs approach, we will consider that judgments on how substantial anticipated effects (health benefits and harms) are should be influenced by: 1) the size of the intervention's effects on each outcome (e.g. the probability of people who experience benefit or harm) 2) the value assigned to those outcomes by the people who are affected.[5] Under this assumption, we will collect data about the association between the dyad composed of size of intervention's effects and value of the outcome on one hand, and judgments on the magnitude of the anticipated effects on the other. In accordance with the EtD frameworks, judgments on desirable effects and on undesirable effects will be collected separately and should not account for any potential tradeoff between benefits and harms. We will use this data to estimate the DTs and provide a conceptual framework for their interpretation and use (see supplemental file 1).

Methods

This study will consist of two parts. In the first part, we will conduct a methodological randomized controlled trial to collect the data that will be used to derive the DTs. Second, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Randomized controlled trial

The following description of methods and analysis of this trial follows the latest guidance by the Standard Protocol Items: Recommendations for Interventional Trials.[13] We submitted this protocol for registration to the Protocols Registration and Results System (clinicaltrials.gov).

Design and setting

Study participants will be recruited to complete a randomized electronic survey (see supplemental file 2) designed to elicit ratings on the magnitude of the potential health effects (benefits or harms) of interventions. Ratings on health benefits and health harms will be collected separately. We will organize the survey into three sections: introduction and example, ratings, questions about respondent demographics. Ratings will be based on five outcomes having a different impact on health (death, major ischemic stroke, pulmonary embolism of moderate severity, diarrhea of moderate severity, and mild nausea/vomiting) presented through descriptive case-scenarios. Each case-scenario will include: (1) a GRADE Summary of Finding (SoF) table[14] providing information about the PICO (Population, Intervention, Comparator, Outcome), the relative and absolute anticipated effects of the intervention, and the certainty in the evidence; (2) a Health Outcome Descriptor [15] describing key attributes of the outcome under consideration including symptoms, time horizon, testing and treatment, and consequences; (3) a measure of the impact on health of the outcome (also known as ‘value’ of the outcome or ‘health utility’ in health economics). This measure will be expressed on a scale from 0 (being dead) and 1 (perfect health) which means that outcomes with a higher value are valued closer to perfect health as compared to outcomes with a lower value. For each outcome, we will include a case-scenario descriptive of desirable health effects and another one descriptive of undesirable health effects, for a total of ten case-scenarios across five outcomes. These scenarios differ in the description of the severity of the outcome and the consequences to represent clearly different values.

Participants

Selection Criteria

The target population of the survey will include clinicians, epidemiologists, decision scientists, health research methodologists, experts in health technology assessment (HTA), and members of guideline working groups, but it will be open to the public too. Prior knowledge of the GRADE approach and experience with the EtD frameworks will not be required for participation.

Patient and Public Involvement statement

There was no direct dedicated patient or public involvement but patients and the public will participate in the survey and can provide feedback.

Recruitment

We will distribute the survey through colleagues, the research group's e-mail lists including that of the Cochrane Collaboration, Guidelines International Network (G-I-N), guideline developers, and of the Global Evidence Synthesis Initiative (GESI). Twitter, LinkedIn, and other social medial platforms will be also used for broader distribution. We will continue recruitment for this trial until reaching our anticipated sample size (see below) or until December 31, 2022 as it is unlikely that we will meet the sample size through additional recruitment efforts beyond then.

Intervention and comparison

Participants will be randomized to a set of 4 case-scenarios, written in lay language, that will be used as intervention (or comparison) in this trial. For each case-scenario, we will ask survey participants to consider the intervention's effects and the value of the outcome and rate how substantial the described health benefits or health harms are. We will also ask them to indicate the lower and upper bound for the ranges of magnitudes of absolute risk difference (ARD) that they associate with the judgments of 'Small' and 'Moderate'. Any estimate below the lower bound for 'Small' will be considered as 'Trivial or None', and any estimate above the upper bound of 'Moderate' will be considered as 'Large'.

Outcomes

The primary endpoints of this trial are the three DTs (T1=$DT_{Trivial/Small}$, T2=$DT_{Small/Moderate}$, T3=$DT_{Moderate/Large}$) that would allow discriminating between EtD judgments of 'Trivial or None' and 'Small', 'Small' and 'Moderate', and 'Moderate' and Large', respectively.

Randomization

Randomization will ensure that case-scenarios will be equally distributed across survey participants to get balanced judgments on outcomes. It will reduce potential confounding due to order effects and possible differences between case-scenarios (e.g. clarity). Randomization will also avoid selection bias that could arise if allowing participants to select the case-scenarios more familiar to them.

Sample size calculation

We based our sample size calculation on the data collected during pilot-testing (n=15 participants). Based on this data, we computed the mean thresholds T1, T2, and T3 for each outcome separately and estimated that we need to recruit 1406 survey respondents to demonstrate a difference of 15% of the mean with non-overlapping 95% confidence intervals. These computations were done using Winpepi. [16]

Statistical methods

Calculation of thresholds from survey ratings

We will use the ranges of ARD for judgments of 'Small' and 'Moderate' collected through the survey to calculate the thresholds associated with each rating. The thresholds will be derived through the product between each the ARD indicated as range boundary and the difference in value from perfect health (1 - outcome's value) for the outcome associated with that rating (see supplemental file 2). We will calculate the DTs as the weighted mean of the corresponding thresholds derived from survey ratings. We will use a weighted mean to account for multiple ratings from the same survey respondent.

Primary analysis

We will use frequencies and percentages to describe the characteristics of survey respondents. For each DT, we will calculate mean, standard deviation (SD), and 95% confidence intervals (C.I.). We will conduct an ANOVA to determine if there are any differences between the thresholds (T1≠T2≠T3). If we identify a difference, since each participant will contribute data to each threshold, we will employ a post-hoc paired sample t-test to assess which of the DTs are different i.e., (T1≠T2; T2≠T3; T1≠ T3). Our a priori hypothesis is that there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms.

Sensitivity analyses

Subgroup analysis

We will conduct explorative subgroup analyses based on participants' characteristics (training in epidemiology, familiarity with the EtD framework, previous participation in guideline development groups, language used). Our a-priori hypotheses is that, in each of the identified subgroups, there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms.

Incoherent ratings and outliers

We expected that, given the complexity of the topic, some responses might not be internally coherent or represent outliers. We define a threshold as incoherent if T1>T2 OR T2>T3. We define thresholds as outliers if they fall more than three interquartile ranges below the first quartile or above the third quartile. We will verify if the primary analysis would differ if incoherent thresholds or data outliers are excluded. The a priori hypothesis for the sensitivity analyses will be the same as for the primary analysis.

Order effects

We will conduct an ANOVA analysis to assess for potential order effects. We will examine whether participants randomized to a case-scenario for a low-value outcome (outcome value <0.5) in the first case-scenario provided different thresholds as compared to participants who were randomized to a high-value outcome first. Similarly, we will examine whether participants who provided a judgment of 'Small' in the first iteration provided different thresholds as compared to participants who provided a judgment of 'Large' in the first iteration. Our a priori hypothesis is that of no differences if comparing each DT between these groups.

Retrospective comparison of judgments

To investigate the validity of our DTs, we will purposively select judgments from existing guidelines developed using the EtD frameworks and measure the agreement between judgments made by guideline panels and the judgments that our DTs approach would suggest. We will consider for inclusion guidelines reporting the value assigned to outcomes during the decision-making process. We will use frequencies and percentages to describe the agreement. We will employ SPSS v26 (IBM Corp., Armonk, NY) to conduct all statistical analyses. We will use the Bonferroni correction for multiple testing in all secondary analyses.[17]

Pilot testing and assessment of feasibility

To ensure usability and clarity of the survey across respondents having different background or expertise, we piloted the survey with study co-investigators as well as complementary representatives of the target population (n=15). Comments on three iterations of the survey were collected either electronically or by voice recordings and discussed during study meetings. Furthermore, to test the feasibility of the study, we recruited 75 participants from the target population. Participants were able to complete the exercise in the majority of cases. Only 7 out of 75 did not complete the survey after they signed up. Participants contributed a total of 295 ratings with only 17 out of 312 expected ratings missing indicating that the approach to obtaining DTs is feasible. This is true for people of varying backgrounds and educational levels. The findings based on the preliminary analysis of the data support our hypothesis that DTs can help discriminate between the judgments (see supplemental file 3). Furthermore, we will use periodic interim results to inform judgments by guideline groups that develop recommendations but will not use these to draw final conclusions about the trial results until it is stopped formally by reaching the calculated sample size or on December 31, 2022. No additional data is available.

Ethics and dissemination

After review, the Hamilton Integrated Research Ethics Board (HiREB) determined that as a quality improvement project, this study was exempt from formal ethics review as per TCPS2 (2014) Article 2.5. We will inform respondents of this decision and the anonymous nature of the study. Survey participants will be required to read a consent statement in order to participate in this study at the beginning of the trial. This statement reads: You are being invited to participate in a research project which aims to identify indicative decision thresholds that could assist users of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks in making judgments. Your input will be used in determining these indicative thresholds. By completing this survey, you provide consent that the anonymized data collected will be used for the research study and to be summarized in aggregate in publication and electronic tools.

The results of this randomized trial will be published in a peer-reviewed journal. We also aim to present the results in national and international conferences.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Discussion

We believe that DTs for judgments on desirable and undesirable health effects can be useful to decision-makers using the EtD frameworks. Guideline panels using the GRADE EtDs often ask what are 'Trivial or None', 'Small', 'Moderate' and 'Large' effects. The proposed DTs approach could provide an answer based on empirical data and be used to initiate and promote discussion. Furthermore, it is simple to apply, and requires only to calculate the product between ARD and the reduction in value associated with the outcome. This endeavor will expand the research on the use of decision thresholds within the GRADE methodology and could be integrated into GRADEpro.

Our work with Hultcrantz et al.[12] suggests that clinical decision thresholds can be used to allow appropriate ratings of the certainty of the evidence, but there is no empirical data. Furthermore, it focuses on the construct of certainty of evidence and targets different degrees of contextualization, while we address judgments on the magnitude of effects and made by users of the EtD frameworks. The joint consideration of the estimate of effect and outcome's importance has been already adopted in another effort of the GRADE Working Group. In a GRADE concept paper[18], Alper et al. aim to define the certainty in the net benefit and suggest calculating the net effect of an intervention by combining importance-adjusted effect estimates calculated from different outcomes. While this strategy is appealing and would allow us to apply our research to EtD judgments on the trade-off between benefits and harms, further research is needed to establish if the estimates to be combined are independent and not correlated with each other. Other quantitative approaches to assess the benefits, harms, and net benefit associated with treatments are available in the literature[19], but none aims to characterize the magnitude of effects into categories (i.e 'Trivial or None', 'Small', 'Moderate', 'Large') as needed to make judgments using the EtD frameworks. Utilitarian frameworks are common in health economic research, where health-utilities elicited from target populations are used to inform modeling techniques such as cost-effectiveness analysis based on quality-adjusted-life-years (QALY). [20,21]

.

Contributor statement

HJS is the principal investigator and conceived of the study and together with GPM, NS, FX, and JB designed and established this research project. ML, IV, TP, EP, ZSP, AB, and US piloted the survey and provided methodological input. GPM was responsible for the ethics application and with HJS for registration of the protocol on clinicaltrials.gov. HJS, GPM, AB, and LM designed the statistical analysis. GPM, AB, WW, AD, HJS are responsible for recruitment. HJS, WW and GPM are responsible for the coordination of the study. GPM and HJS drafted the manuscript. NS, FX, AB, and JB reviewed early drafts. EAA and PA-C provided methodological input in the study design. All listed authors participated in the writing and revision of the manuscript and approved the its final version.

References

1. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice G. *Clinical Practice Guidelines We Can Trust.* Washington (DC): National Academies Press (US); 2011.
2. Schünemann HJ. Guidelines 2.0: do no net harm-the future of practice guideline development in asthma and other diseases. *Curr Allergy Asthma Rep.* 2011;11(3):261-268.
3. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res.* 2004;4(1):38. doi: 10.1186/1472-6963-1184-1138.
4. Alonso-Coello P, Schunemann 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.* 2016;353:i2016.
5. 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.* 2016;353:i2089.
6. Marcucci M, Sinclair JC. A generalised model for individualising a treatment recommendation based on group-level evidence from randomised clinical trials. *BMJ Open.* 2013;3(8).
7. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics.* 2008;26(9):733-744.
8. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med.* 1975;293(5):229-234.
9. Hultcrantz M, Mustafa RA, Leeflang MMG, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. *J Clin Epidemiol.* 2020;117:138-148.
10. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol.* 2017;87:4-13.
11. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. *P T.* 2008;33(12):700-711.
12. Johnston BC, Ebrahim S, Carrasco-Labra A, et al. Minimally important difference estimates and methods: a protocol. *BMJ Open.* 2015;5(10):e007953.

13. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ (Clinical research ed).* 2013;346:e7586.

14. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol.* 2013;66(2):173-183. doi: 110.1016/j.jclinepi.2012.1008.1001. Epub 2012 Oct 1030.

15. Baldeh T, Saz-Parkinson Z, Muti P, et al. Development and use of health outcome descriptors: a guideline development case study. *Health Qual Life Outcomes.* 2020;18(1):167.

16. Abramson JH. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. *Epidemiol Perspect Innov.* 2011;8(1):1.

17. Abdi H. Holm’s sequential Bonferroni procedure. In: Salkin N, ed. *Encyclopedia of research design.* Thousand Oaks, California2010:1-8.

18. Alper BS, Oettgen P, Kunnamo I, et al. Defining certainty of net benefit: a GRADE concept paper. *BMJ Open.* 2019;9(6):e027445.

19. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol.* 2012;12:173.

20. Siebert U. When should decision-analytic modeling be used in the economic evaluation of health care? *The European Journal of Health Economics, formerly: HEPAC.* 2003;4(3):143-150.

21. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;296(13):716-721.

Supplemental file 1

How to derive the suggested judgment:

If health benefits/harms< $\text{DT}_{\text{Trivial to Small}}$ → suggested judgment is of 'Trivial or None' (e.g. health benefit A)

If $\text{DT}_{\text{Trivial to Small}}$ < health benefits/harms< $\text{DT}_{\text{Small to Moderate}}$ → suggested judgment is of 'Small'

If $\text{DT}_{\text{Small to Moderate}}$ < health benefits/harms< DT_{Large} → suggested judgment is of 'Moderate'

If health benefits/harms> DT_{Large} → suggested judgment is of 'Large'

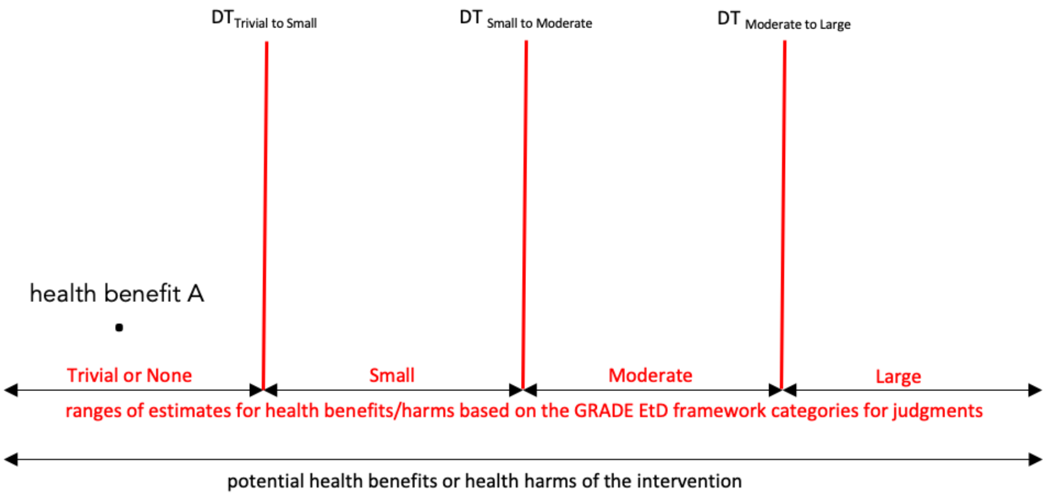

$\text{DT}_{\text{Trivial to Small}}$
$\text{DT}_{\text{Small to Moderate}}$
$\text{DT}_{\text{Moderate to Large}}$
health benefit A
Trivial or None
Small
Moderate
Large
ranges of estimates for health benefits/harms based on the GRADE EtD framework categories for judgments
potential health benefits or health harms of the intervention

Figure 1

Figure 1 legend: The availability of three DTs ($\text{DT}_{\text{Trivial/Small,}}$ $\text{DT}_{\text{Small/Moderate,}}$ $\text{DT}_{\text{Moderate/Large}}$) would allow to discriminate between the four GRADE EtD framework categories for judgments. For a given health benefit/harm, the suggestion on the judgment would depend on how the estimate of health benefits/harms compares to the DTs. In this example, the health benefit A lies on the left (is smaller) of the $\text{DT}_{\text{Trivial/Small}}$ which would suggest that the judgment of 'Trivial or None' would be more appropriate than the others.

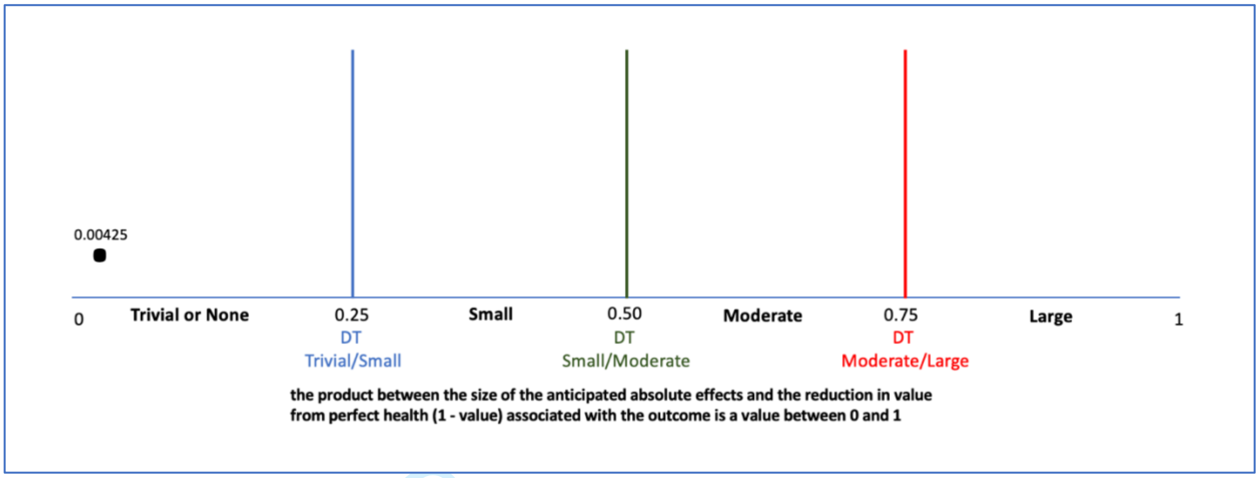
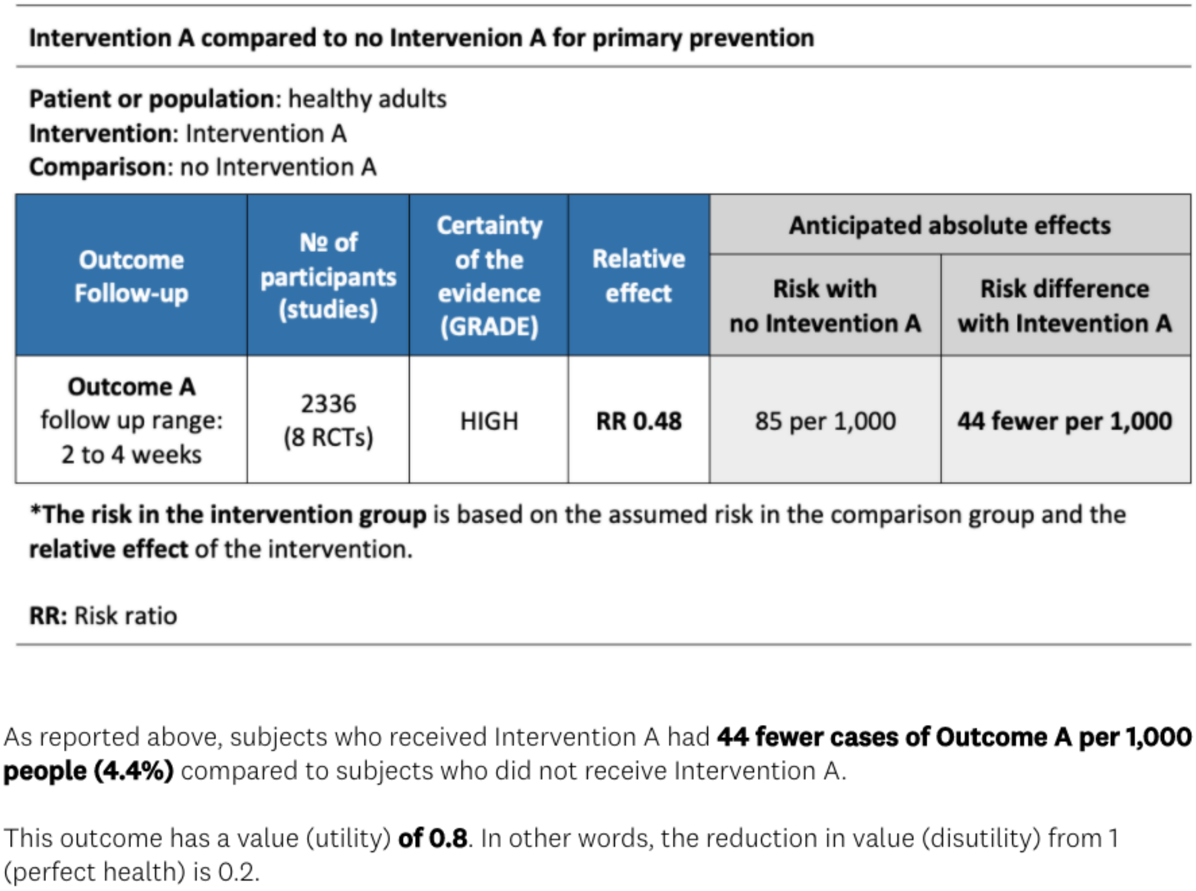

0.00425
0
Trivial or None
0.25
DT
Trivial/Small
Small
0.50
DT
Small/Moderate
Moderate
0.75
DT
Moderate/Large
Large
1
the product between the size of the anticipated absolute effects and the reduction in value from perfect health (1 - value) associated with the outcome is a value between 0 and 1

Figure 2

Figure 2 legend: We assumed to have known DTs ($DT_{Trivial/Small}$ = 0.25, $DT_{Small/Moderate}$ = 0.50, $DT_{Moderate/Large}$ = 0.75) and wanted to assign to one of the 4 EtD categories the health benefit of an intervention showing an anticipated absolute effect of 17 fewer per 1000 on an outcome valued 0.75. Following the proposed approach, we calculated the result of the product (score) of the size of anticipated effects (Absolute Risk Difference, ARD) and the reduction in value from perfect health (1 - outcome's value) associated with the outcome under evaluation. In this example, the following approach (ARD * (1 - outcome's value) = (17/1000)*(1 - 0.75)) resulted in the value of 0.00425. We then plotted this value and obtained the suggested judgment according to the DTs approach that, in this case, would be of 'Trivial or None' considering that the calculated value is smaller than the $DT_{Trivial/Small}$. In case of judgments made considering more than one outcome, it would be required to calculate the aforementioned product for each of the outcomes under evaluation and derive an aggregate score defined as the sum of all the individual scores.

Under the assumption that the outcomes are independent, this score could be then compared to the DTs and used to obtain a suggestion on the judgment.

Supplemental file 2



Intervention A compared to no Intervenion A for primary prevention

Patient or population: healthy adults
Intervention: Intervention A
Comparison: no Intervention A

Outcome Follow-up	№ of participants (studies)	Certainty of the evidence (GRADE)	Relative effect	Anticipated absolute effects	
				Risk with no Intevention A	Risk difference with Intevention A
Outcome A follow up range: 2 to 4 weeks	2336 (8 RCTs)	HIGH	**RR 0.48**	85 per 1,000	**44 fewer per 1,000**

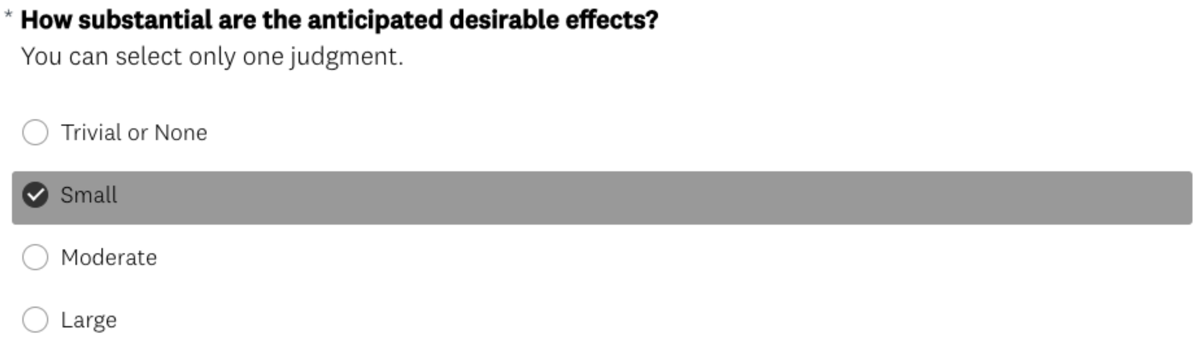
***The risk in the intervention group** is based on the assumed risk in the comparison group and the **relative effect** of the intervention.

RR: Risk ratio

As reported above, subjects who received Intervention A had **44 fewer cases of Outcome A per 1,000 people (4.4%)** compared to subjects who did not receive Intervention A.

This outcome has a value (utility) **of 0.8**. In other words, the reduction in value (disutility) from 1 (perfect health) is 0.2.

Figure 1 - Example of a case-scenario



* **How substantial are the anticipated desirable effects?**

You can select only one judgment.

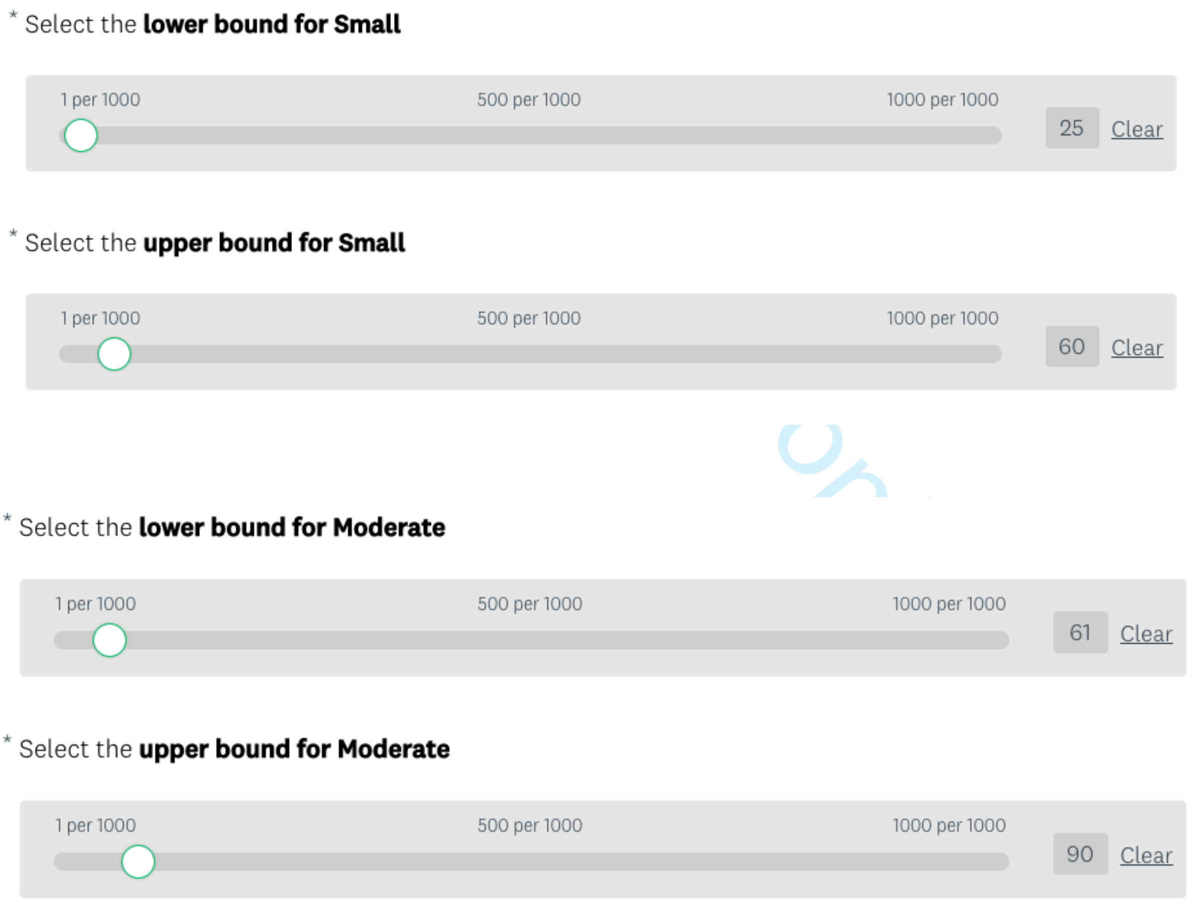
○ Trivial or None

☑ Small

○ Moderate

○ Large

Figure 2 - Judgment on health benefits



* Select the **lower bound for Small**

* Select the **upper bound for Small**

* Select the **lower bound for Moderate**

* Select the **upper bound for Moderate**

Figure 3 - Selection of ranges for judgments of Small and Moderate

Example of calculation of DTs based on survey data

In the examples shown in Figures 2 and 3 we assumed that, after having evaluated a given case-scenario (ARD of 44 events fewer per 1000 on an outcome valued 0.8), a survey participant rated the hypothetical ranges of ARD for judgments of 'Small' and 'Moderate' of from 25 fewer per 1000 to 60 fewer per 1000, and of from 61 fewer per 1000 to 90 fewer per 1000, respectively. We used this data to derive the ranges of ARD for judgments of 'Trivial or None' and of 'Large' (table below).

boundaries of ranges described in Figure 3 value of the outcome = 0.8						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	range
0 per 1000	24 per 1000	25 per 1000	60 per 1000	61 per 1000	90 per 1000	more than 90 per 1000

Table 1 - Ranges of sizes of effects (ARD)

For each range of ARD, we calculated the product between range boundaries and the reduction in value from perfect health (1 - outcome's value) for the outcome associated. Then, we derived the DTs as follow: DT$_{\text{Trivial/Small}}$ equal to the the product calculated from the lower bound for the judgment of 'Small', DT$_{Small/Moderate}$ equal to average of the products calculated from the upper bound for the judgment of 'Small' and the lower bound for the judgment of 'Moderate', and DT$_{Moderate/Large}$ equal to the smallest number larger than the mean of the products calculated from the upper bound for the judgment of 'Moderate'.

product values = ARD * (1- outcome's value))						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	any value
(0/1000)*0.2 0	(24/1000) *0.2 0.0048	(25 /1000) *0.2 0.005	(60/1000) *0.2 0.012	(61/1000)*0.2 0.0122	(90/1000)*0.2 0.018	bigger than (90/1000)*0.2 >0.018

Table 2 - Ranges of product values

Using the data from Table 2, the DTs would result as follow: DT$_{Trivial/Small}$ = 0.005, DTs$_{mall/Moderate}$ = 0.0121, DTModerate$_{/Large}$ = 0.0180001.

Supplemental file 3

Descriptive statistics

The presented preliminary analysis is based on survey data collected between May 1st and July 21st, 2020. Our dissemination strategy allowed recruitment of 75 participants who contributed a total of 295 ratings. Fifty-six survey participants had a background in research (74.6%) and 36 were healthcare professionals (50.6%). Thirty-four respondents (45.3%) were members of academia. Other major groups were participants from HTA organizations and professional societies (13.3% and 18.6%, respectively). Participants were equally randomized to case-scenarios descriptive of desirable and undesirable health effects (144/295, 49%; 151/295, 51%, respectively) and completed the entire exercise in the majority of cases (68/75, 90.7%). Detailed descriptive characteristics of survey respondents and ratings are shown in Tables 1 and 2, respectively.

Characteristic [a]	**Respondents, n = 75**
Background [a]	n (%)
Clinical/Health Professional	38 (50.6)
Policymaking	6 (8.0)
Research	56 (74.6)
Teaching	18 (24.0)
Administrative	3 (4.0)

Patient representative	2 (2.6)
Other	3 (4.0)
Degree [a]	
Degree in Nursing (RN)	1 (1.3)
Medical School (MD)	30 (4.0)
Master of Sciences (MSc)	17 (22.6)
Master of Public Health (MPH)	9 (0.12)
Doctor of Philosophy (PhD)	25 (33.3)
None	2 (2.6)
Other	5 (6.6)
Formal Training in health research methodology/epidemiology/biostatistics	
Never completed	12 (16.0)
Completed some form of formal training but do not have a graduate degree	30 (40.0)
Earned a MSc degree	16 (21.3)
Earned a PhD degree	16 (21.3)
Not available	1 (1.4)
Organization [a]	
Cochrane collaboration	13 (17.3)
GRADE Working Group	16 (21.3)

World Health Organization	1 (1.4)
Guidelines International Network (G-I-N)	-
Health Technology Assessment (HTA) organization	10 (13.3)
Academia	34 (45.3)
Professional society	14 (18.6)
Familiarity with the Evidence to Decision framework	
Not at all familiar	5 (6.6)
Not so familiar	9 (12.0)
Somewhat familiar	16 (21.3)
Very familiar	30 (40.0)
Extremely familiar	8 (10.6)
Not available	7 (9.5)
Previous participation in guideline development groups	
Yes	52 (69.3)
No	18 (24.0)
Not available	5 (6.6)
Primary role in the guideline development group [a]	
Clinical Chair	5 (6.6)
Chair for methods	15 (19.8)
Guideline methodologist	29 (38.6)
Panel member	15 (19.8)

Topic or content expert	7 (9.5)
Patient representative	2 (2.6)
Systematic review author	26 (34.6)
Expert in Health Technology Assessment	3 (4.0)

Values represent the number and in parentheses the percentage.

[a] Percentages do not add up to 100 because respondents could choose more than one option.

Table 1: Characteristics of survey respondents

Characteristics of ratings collected through the survey	n (%)
Total number of ratings collected	295
Missing data (expected ratings - collected ratings/expected ratings)	17/312 (0.054)[a]
randomized to a scenario showing desirable effects	144/295 (49)
randomized to a scenario showing undesirable effects	151/295 (51)
randomized to the outcome of death	73/295 (25)
randomized to the outcome of major stroke	66/295 (22)
randomized to the outcome of pulmonary embolism	55/295 (19)
randomized to the outcome of moderate diarrhea	63/295 (21)
based on the outcome of mild nausea/vomiting	38/295 (13)

a. 73 participants were randomized to 4 case-scenarios, 2 were mistakenly randomized to 10.

Table 2: Descriptive statistics of survey ratings

Table 3 describes the estimates of DTs that were derived from survey ratings through the joint measure of absolute effects and outcome values. For example, an outcome valued as 0.8, these thresholds would indicate that the effect of an intervention preventing 30 events of that outcome per 1000 should be categorized as trivial (since 0.03*(1-0.8)) =0.006 is smaller than T1). More details about the calculation of the DTs are available in Appendix 1 (Table 1).

Decision Threshold			**95% Confidence Interval**	
	Estimate	Std. Deviation	Lower Bound	Upper Bound
T1: Trivial/Small	0.0165	0.0467	0.0059	0.0271
T2: Small/Moderate	0.0312	0.0601	0.0176	0.0448
T3: Moderate/Large	0.0577	0.0781	0.0400	0.0754

Table 3: Estimates of DTs

Primary analysis

Our analysis showed a difference in the estimates between T1 and T2 (mean difference [MD] -0.0147; 95% CI -0.0201 to -0.0093; p<0.001) and T2 and T3 (mean difference [MD] -0.0264; 95% CI -0.0544 to -0.0062; p<0.001).

Within-participant analyses

The analyses showed that at a respondent level there was no difference between DTs derived from judgments on benefits and from those on harms: $T1_{benefit}$=$T1_{harms}$ (mean difference [MD] -0.0040 ; 95% CI -0.0195 to 0.0116 ; p=0.615) ; $T2_{benefits}$=$T2_{harms}$ (mean difference [MD] -0.0124; 95% CI -0.0313 to 0.0064 ; p=0.196); $T3_{benefit}$=$T3_{harms}$ (mean difference [MD] -0.0209; 95% CI -0.0451 to 0.0033; p=0.090).

Subgroup analyses

Our subgroup analyses showed a difference in the estimates between T1 and T2, and T2 and T3 also in DTs derived from subgroup of ratings identified by outcome, direction of interventions' effects, and prior participation to guideline development groups. No difference was observed in the estimates between T1 and T2 in those with no experience with the EtD (mean difference [MD] -0.0046; 95% CI -0.0100 to 0.0006; p=0.810) and between T2 and T3 in those who had no training in epidemiology (mean difference [MD] -0.0056; 95% CI -0.0218 to 0.0106; p=0.483).

Sensitivity analyses

The findings of the sensitivity analyses conducted by excluding raters who provided incoherent thresholds (n=3; T1/T2 mean difference [MD] -0.0143; 95% CI -0.0192 to -0.0094; p<0.001; T2/T3 mean difference [MD] -0.0291; 95% CI -0.0417 to -0.0165; p<0.001) or who were presumed outliers (n=10; T1/T2 mean difference [MD] -0.0096 ; 95% CI -0.0113 to -0.0078; p<0.001; T2/T3 mean difference [MD] -0.0194; 95% CI -0.0240 to -0.0148; p<0.001) were similar to that of the primary analysis.

Assessment of order effects

The analyses suggest no difference between DTs derived from participants who evaluated a high-value outcome (i.e. moderate diarrhea) in the first iteration compared to those who evaluated a low-value outcome (i.e. death) first. Similarly, there was no difference in the DTs depending on whether the first judgment made was 'Small' or 'Large'.

BMJ Open

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study (GRADE-THRESHOLD)

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-053246.R2
Article Type:	Protocol
Date Submitted by the Author:	31-Jan-2022
Complete List of Authors:	Morgano, Gian Paolo; McMaster University, Department of Health Research Methods, Evidence and Impact Mbuagbaw, Lawrence; McMaster University, Department of Health Research Methods, Evidence, and Impact Santesso, Nancy; McMaster University, Department of Health Research Methods, Evidence, and Impact Xie, Feng; McMaster University, Department of Health Research Methods, Evidence, and Impact Brozek, Jan; McMaster University, Department of Health Research Methods, Evidence and Impact Siebert, Uwe; UMIT, Institute of Public Health, Medical Decision Making and Health Technology Assessment; Harvard University, Center for Health Decision Science/Dept. of Health Policy and Management, Harvard TH Chan School of Public Health and MGH-ITA/Dept. of Radiology, Harvard Medical School Bognanni, Antonio; McMaster University, Department of Health Research Methods, Evidence and Impact Wiercioch, Wojtek; McMaster University, Department of Health Research Methods, Evidence, and Impact Piggott, Thomas; McMaster University, Department of Health Research Methods, Evidence and Impact Darzi, Andrea; McMaster University, Department of Health Research Methods, Evidence, and Impact Akl, Elie; American University of Beirut, Clinical Research Institute, Faculty of Medicine Verstijnen, Ilse; Dutch National Health Care Institute, Parmelli, Elena; European Commission Joint Research Centre Ispra Sector, Saz-Parkinson, Zuleika; European Commission Joint Research Centre Alonso-Coello, Pablo; Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Schünemann, Holger J ; McMaster University, Department of Health Research Methods, Evidence and Impact
<b>Primary Subject Heading</b>:	Evidence based practice

Secondary Subject Heading:	Communication, Epidemiology, Health informatics, Health policy, Health services research
Keywords:	Health policy < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Protocols & guidelines < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, Quality in health care < HEALTH SERVICES ADMINISTRATION & MANAGEMENT, PUBLIC HEALTH, STATISTICS & RESEARCH METHODS, EPIDEMIOLOGY

Defining decision thresholds for judgments on health benefits and harms using the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks: a protocol for a randomized methodological study (GRADE-THRESHOLD)

Authors:

Gian Paolo Morgano[1,2], Lawrence Mbuagbaw[1], Nancy Santesso[1,2], Feng Xie[1], Jan L. Brozek[1,2], Uwe Siebert[3], Antonio Bognanni[1,2], Wojtek Wiercioch[1,2], Thomas Piggott[1,2], Andrea J. Darzi[1,2], Elie A. Akl[1,4], Ilse Verstijnen[5], Elena Parmelli[6], Zuleika Saz-Parkinson[6], Pablo Alonso-Coello[7], Holger J. Schünemann[1,2,8,9]

Author Affiliations:

1. Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
2. Michael G. DeGroote Cochrane Canada & McMaster GRADE Centres; Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada
3. Institute of Public Health, Medical Decision Making and Health Technology Assessment, Department of Public Health, Health Services Research and Technology Assessment, UMIT – University for Health Sciences, Medical

Informatics and Technology, Hall i.T., Austria; Division of Health Technology Assessment and Bioinformatics, ONCOTYROL - Center for Personalized Cancer Medicine, Innsbruck, Austria; Center for Health Decision Science, Departments of Epidemiology and Health Policy & Management, Harvard T.H. Chan School of Public Health, Boston, MA, USA; Institute for Technology Assessment and Department of Radiology and Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

4. Department of Internal Medicine, American University of Beirut, Beirut, Lebanon
5. Dutch National Health Care Institute (ZIN), Diemen, The Netherlands
6. European Commission, Joint Research Centre (JRC), Ispra, VA, Italy
7. Iberoamerican Cochrane Centre, CIBERESP-IIB Sant Pau, Barcelona, Spain
8. Department of Medicine, McMaster University, Hamilton, ON, Canada
9. Dipartimento di Scienze Biomediche Humanitas University, Milan, Italy

Corresponding author:

Prof. Holger J. Schünemann
Department of Health Research Methods, Evidence and Impact, McMaster University,
1280 Main St West, Hamilton,
ON; L8N 3Z5, Canada
schuneh@mcmaster.ca

Abstract

Introduction

The GRADE Evidence to Decision (EtD) frameworks require its users to judge how substantial the effects of interventions are on desirable and undesirable people-important health outcomes. However, decision thresholds (DTs) that could help differentiate across judgments and serve as reference for interpretation of findings are not yet available. The objective of this study is an approach to derive and use decision-thresholds (DTs) for EtD judgments about the magnitude of health benefits and harms. We hypothesize that approximate DTs could have the ability to discriminate between the existing four categories of EtD judgments (Trivial, Small, Moderate, Large), support panels of decision-makers in their work, and promote consistency and transparency in judgments.

Methods and analysis

We will conduct a methodological randomized controlled trial to collect the data that allow deriving the DTs. We will invite clinicians, epidemiologists, decision scientists, health research methodologists, experts in Health Technology Assessment (HTA), members of guideline development groups and the public to participate in the trial. Then, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Ethics and dissemination

The Hamilton Integrated Research Ethics Board reviewed this study as a quality improvement study and determined that it requires no further consent. Survey participants will be required to read a consent statement in order to participate in this study at the beginning of the trial. This statement reads: You are being invited to participate in a research project which aims to identify indicative decision thresholds that could assist users of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks in making judgments. Your input will be used in determining these indicative thresholds. By completing this survey, you provide consent that the anonymized data collected will be used for the research study and to be summarized in aggregate in publication and electronic tools.

Article summary

Strengths and limitations of this study

- The calculation of the Decision Thresholds will be based on empirical data.
- We will use structured case-scenarios to present survey participants with the information relevant to make their judgments.
- We will employ a randomization process to ensure that case-scenarios will be equally distributed across survey participants
- We acknowledge that the survey requires effort and that this could impact test-retest reliability and applicability of the survey results which we overcome in part by conducting a large trial

Introduction

As advocated by the National Academy of Medicine of the United States (formerly the Institute of Medicine), the assessment of the benefits and harms of alternative care options (i.e., interventions, actions) is an essential component of any decision-making process underlying guideline recommendations.[1] This assessment should be explicit and include considerations around the probability, magnitude, and importance of health related benefits and health related harms, and other desirable and undesirable consequences of the recommendation or decision.[2] The Grading of Recommendations Assessment, Development and Evaluation (GRADE) Working Group has developed the Evidence to Decision (EtD) frameworks to help guideline developers use the evidence in a structured and transparent way and to ensure that they consider all the criteria relevant to their decisions.[3,4] The EtD frameworks require decision-makers to evaluate explicitly the benefits and harms of alternative care options through separate judgments based on the two following questions: *"How substantial are the desirable anticipated effects (health benefits)?"*, *"How substantial are the undesirable anticipated effects (health harms)?"*. The guidance from the GRADE Working Group includes expressing and facilitating these judgments by assigning the health benefits or health harms of some intervention under evaluation to one of the following four categories: 'Trivial or None', 'Small', 'Moderate' and 'Large'.[3,4] To be useful, however, this simplification requires that EtD users have a similar understanding of what magnitude of health benefits or health harms belong into which category and are consistent in their judgments. A similar common understanding is also important between those assigning a category and those interpreting the meaning of a category that is communicated to them (i.e. "imagining" how substantial is an effect based on the category). This can be achieved only when people make similar judgments. To direct EtD users on how to make these judgments appropriately, the GRADE Working Group has produced guidance articles that include the description of the underpinning concepts and examples of judgments based on clinical scenarios.[4,5] Despite the popular use of thresholds to support decision-making in various fields of healthcare research,[6-8] and its adoption in some aspects of the GRADE approach[9,10], use of Decision Thresholds (DTs) for EtD judgments about health benefits and harms is not yet established. For continuous outcomes, EtD users are advised to revert to statistical notions such as Cohen's standardized effect sizes or the Minimal Important Difference (MID) to interpret the magnitude of effects.[11,12] However, empirical data supporting judgments on health benefits and harms for dichotomous outcomes are not yet available for the EtDs.

Objectives

The objective of this study is to derive DTs for EtD judgments on the magnitude of health benefits and harms. We hypothesize that DTs could discriminate between the four categories for EtD judgments. Explicit DTs, providing an indication for which could be the appropriate judgment for a given scenario, might have the potential to support panels of decision-makers in their work, facilitate a common understanding, and promote consistency and transparency in judgments.

Conceptual approach

In the proposed DTs approach, we will consider that judgments on how substantial anticipated effects (health benefits and harms) are should be influenced by: 1) the size of the intervention's effects on each outcome (e.g. the probability of people who experience benefit or harm) 2) the value assigned to those outcomes by the people who are affected.[5] Under this assumption, we will collect data about the association between the dyad composed of size of intervention's effects and value of the outcome on one hand, and judgments on the magnitude of the anticipated effects on the other. In accordance with the EtD frameworks, judgments on desirable effects and on undesirable effects will be collected separately and should not account for any potential tradeoff between benefits and harms. We will use this data to estimate the DTs and provide a conceptual framework for their interpretation and use (see supplemental file 1).

Methods

This study will consist of two parts. In the first part, we will conduct a methodological randomized controlled trial to collect the data that will be used to derive the DTs. Second, we will investigate the validity of our DTs by measuring the agreement between judgments that were made in the past by guideline panels and the judgments that our DTs approach would suggest if applied on the same guideline data.

Randomized controlled trial

The following description of methods and analysis of this trial follows the latest guidance by the Standard Protocol Items: Recommendations for Interventional Trials.[13] We submitted this protocol for registration to the Protocols Registration and Results System (clinicaltrials.gov).

Design and setting

Study participants will be recruited to complete a randomized electronic survey (see supplemental file 2) designed to elicit ratings on the magnitude of the potential health effects (benefits or harms) of interventions. Ratings on health benefits and health harms will be collected separately. We will organize the survey into three sections: introduction and example, ratings, questions about respondent demographics. Ratings will be based on five outcomes having a different impact on health (death, major ischemic stroke, pulmonary embolism of moderate severity, diarrhea of moderate severity, and mild nausea/vomiting) presented through descriptive case-scenarios. Each case-scenario will include: (1) a GRADE Summary of Finding (SoF) table[14] providing information about the PICO (Population, Intervention, Comparator, Outcome), the relative and absolute anticipated effects of the intervention, and the certainty in the evidence; (2) a Health Outcome Descriptor [15] describing key attributes of the outcome under consideration including symptoms, time horizon, testing and treatment, and consequences; (3) a measure of the impact on health of the outcome (also known as ‘value’ of the outcome or ‘health utility’ in health economics). This measure will be expressed on a scale from 0 (being dead) and 1 (perfect health) which means that outcomes with a higher value are valued closer to perfect health as compared to outcomes with a lower value. For each outcome, we will include a case-scenario descriptive of desirable health effects and another one descriptive of undesirable health effects, for a total of ten case-scenarios across five outcomes. These scenarios differ in the description of the severity of the outcome and the consequences to represent clearly different values.

Participants

Selection Criteria

The target population of the survey will include clinicians, epidemiologists, decision scientists, health research methodologists, experts in health technology assessment (HTA), and members of guideline working groups, but it will be open to the public too. Prior knowledge of the GRADE approach and experience with the EtD frameworks will not be required for participation.

Patient and Public Involvement statement

There was no direct dedicated patient or public involvement but patients and the public will participate in the survey and can provide feedback.

Recruitment

We will distribute the survey through colleagues, the research group's e-mail lists including that of the Cochrane Collaboration, Guidelines International Network (G-I-N), guideline developers, and of the Global Evidence Synthesis Initiative (GESI). Twitter, LinkedIn, and other social medial platforms will be also used for broader distribution. We will continue recruitment for this trial until reaching our anticipated sample size (see below) or until December 31, 2022 as it is unlikely that we will meet the sample size through additional recruitment efforts beyond then.

Intervention and comparison

Participants will be randomized to a set of 4 case-scenarios, written in lay language, that will be used as intervention (or comparison) in this trial. For each case-scenario, we will ask survey participants to consider the intervention's effects and the value of the outcome and rate how substantial the described health benefits or health harms are. We will also ask them to indicate the lower and upper bound for the ranges of magnitudes of absolute risk difference (ARD) that they associate with the judgments of 'Small' and 'Moderate'. Any estimate below the lower bound for 'Small' will be considered as 'Trivial or None', and any estimate above the upper bound of 'Moderate' will be considered as 'Large'.

Outcomes

The primary endpoints of this trial are the three DTs (T1=$DT_{Trivial/Small}$, T2=$DT_{Small/Moderate}$, T3=$DT_{Moderate/Large}$) that would allow discriminating between EtD judgments of 'Trivial or None' and 'Small', 'Small' and 'Moderate', and 'Moderate' and Large', respectively.

Randomization

Randomization will ensure that case-scenarios will be equally distributed across survey participants to get balanced judgments on outcomes. It will reduce potential confounding due to order effects and possible differences between case-scenarios (e.g. clarity). Randomization will also avoid selection bias that could arise if allowing participants to select the case-scenarios more familiar to them.

Sample size calculation

We based our sample size calculation on the data collected during pilot-testing (n=15 participants). Based on this data, we computed the mean thresholds T1, T2, and T3 for each outcome separately and estimated that we need to recruit 1406 survey respondents to demonstrate a difference of 15% of the mean with non-overlapping 95% confidence intervals. These computations were done using Winpepi. [16]

Statistical methods

Calculation of thresholds from survey ratings

We will use the ranges of ARD for judgments of 'Small' and 'Moderate' collected through the survey to calculate the thresholds associated with each rating. The thresholds will be derived through the product between each the ARD indicated as range boundary and the difference in value from perfect health (1 - outcome's value) for the outcome associated with that rating (see supplemental file 2). We will calculate the DTs as the weighted mean of the corresponding thresholds derived from survey ratings. We will use a weighted mean to account for multiple ratings from the same survey respondent.

Primary analysis

We will use frequencies and percentages to describe the characteristics of survey respondents. For each DT, we will calculate mean, standard deviation (SD), and 95% confidence intervals (C.I.). We will conduct an ANOVA to determine if there are any differences between the thresholds (T1≠T2≠T3). If we identify a difference, since each participant will contribute data to each threshold, we will employ a post-hoc paired sample t-test to assess which of the DTs are different i.e., (T1≠T2; T2≠T3; T1≠ T3). Our a priori hypothesis is that there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms.

Sensitivity analyses

Subgroup analysis

We will conduct explorative subgroup analyses based on participants' characteristics (training in epidemiology, familiarity with the EtD framework, previous participation in guideline development groups, language used). Our a-priori hypotheses is that, in each of the identified subgroups, there will be a difference between the DTs and no difference between the magnitude of DTs for benefits and harms.

Incoherent ratings and outliers

We expected that, given the complexity of the topic, some responses might not be internally coherent or represent outliers. We define a threshold as incoherent if T1>T2 OR T2>T3. We define thresholds as outliers if they fall more than three interquartile ranges below the first quartile or above the third quartile. We will verify if the primary analysis would differ if incoherent thresholds or data outliers are excluded. The a priori hypothesis for the sensitivity analyses will be the same as for the primary analysis.

Order effects

We will conduct an ANOVA analysis to assess for potential order effects. We will examine whether participants randomized to a case-scenario for a low-value outcome (outcome value <0.5) in the first case-scenario provided different thresholds as compared to participants who were randomized to a high-value outcome first. Similarly, we will examine whether participants who provided a judgment of 'Small' in the first iteration provided different thresholds as compared to participants who provided a judgment of 'Large' in the first iteration. Our a priori hypothesis is that of no differences if comparing each DT between these groups.

Retrospective comparison of judgments

To investigate the validity of our DTs, we will purposively select judgments from existing guidelines developed using the EtD frameworks and measure the agreement between judgments made by guideline panels and the judgments that our DTs approach would suggest. We will consider for inclusion guidelines reporting the value assigned to outcomes during the decision-making process. We will use frequencies and percentages to describe the agreement. We will employ SPSS v26 (IBM Corp., Armonk, NY) to conduct all statistical analyses. We will use the Bonferroni correction for multiple testing in all secondary analyses.[17]

Pilot testing and assessment of feasibility

To ensure usability and clarity of the survey across respondents having different background or expertise, we piloted the survey with study co-investigators as well as complementary representatives of the target population (n=15). Comments on three iterations of the survey were collected either electronically or by voice recordings and discussed during study meetings. Furthermore, to test the feasibility of the study, we recruited 75 participants from the target population. Participants were able to complete the exercise in the majority of cases. Only 7 out of 75 did not complete the survey after they signed up. Participants contributed a total of 295 ratings with only 17 out of 312 expected ratings missing indicating that the approach to obtaining DTs is feasible. This is true for people of varying backgrounds and educational levels. The findings based on the preliminary analysis of the data support our hypothesis that DTs can help discriminate between the judgments (see supplemental file 3). Furthermore, we will use periodic interim results to inform judgments by guideline groups that develop recommendations but will not use these to draw final conclusions about the trial results until it is stopped formally by reaching the calculated sample size or on December 31, 2022. No additional data is available.

Ethics and dissemination

After review, the Hamilton Integrated Research Ethics Board (HiREB) determined that as a quality improvement project, this study was exempt from formal ethics review as per TCPS2 (2014) Article 2.5. We will inform respondents of this decision and the anonymous nature of the study. Survey participants will be required to read a consent statement in order to participate in this study at the beginning of the trial. This statement reads: You are being invited to participate in a research project which aims to identify indicative decision thresholds that could assist users of the Grading of Recommendations Assessment, Development and Evaluation (GRADE) Evidence to Decision (EtD) frameworks in making judgments. Your input will be used in determining these indicative thresholds. By completing this survey, you provide consent that the anonymized data collected will be used for the research study and to be summarized in aggregate in publication and electronic tools.

The results of this randomized trial will be published in a peer-reviewed journal. We also aim to present the results in national and international conferences.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Competing interests:

Holger J. Schünemann is the co-chair of the GRADE working group. Decision thresholds will be used in the GRADEpro app and for other projects. Currently no financial interests. Holger J. Schünemann and Jan L. Brozek are co-developer of the GRADEpro app. Uwe Siebert is an unpaid member of: Working Group for the German Clinical S3 Guideline Prevention of Cervical Cancer; Committee for Cancer Screening of the Austrian Federal Ministry of Health; Oncology Advisory Council of the Federal Ministry of Health, Austria.

Discussion

We believe that DTs for judgments on desirable and undesirable health effects can be useful to decision-makers using the EtD frameworks. Guideline panels using the GRADE EtDs often ask what are 'Trivial or None', 'Small', 'Moderate' and 'Large' effects. The proposed DTs approach could provide an answer based on empirical data and be used to initiate and promote discussion. Furthermore, it is simple to apply, and requires only to calculate the product between ARD and the reduction in value associated with the outcome. This endeavor will expand the research on the use of decision thresholds within the GRADE methodology and could be integrated into GRADEpro.

Our work with Hultcrantz et al.[12] suggests that clinical decision thresholds can be used to allow appropriate ratings of the certainty of the evidence, but there is no empirical data.

Furthermore, it focuses on the construct of certainty of evidence and targets different degrees of contextualization, while we address judgments on the magnitude of effects and made by users of the EtD frameworks. The joint consideration of the estimate of effect and outcome's importance has been already adopted in another effort of the GRADE Working Group. In a GRADE concept paper[18], Alper et al. aim to define the certainty in the net benefit and suggest calculating the net effect of an intervention by combining importance-adjusted effect estimates calculated from different outcomes. While this strategy is appealing and would allow us to apply our research to EtD judgments on the trade-off between benefits and harms, further research is needed to establish if the estimates to be combined are independent and not correlated with each other. Other quantitative approaches to assess the benefits, harms, and net benefit associated with treatments are available in the literature[19], but none aims to characterize the magnitude of effects into categories (i.e 'Trivial or None', 'Small', 'Moderate', 'Large') as needed to make judgments using the EtD frameworks. Utilitarian frameworks are common in health economic research, where health-utilities elicited from target populations are used to inform modeling techniques such as cost-effectiveness analysis based on quality-adjusted-life-years (QALY). [20,21] However, our trial will not be free of limitations. Generalizability of the findings may be limited by the use of the case scenarios we chose and the limited number of effect sizes we include in the trial. Generalizability may also be limited by the type of participants we will be able to recruit. Therefore, we plan, following the completion of this trial, to conduct further research with additional case scenarios and different target populations.

.

Contributor statement

HJS is the principal investigator and conceived of the study and together with GPM, NS, FX, and JB designed and established this research project. ML, IV, TP, EP, ZSP, AB, and US piloted the survey and provided methodological input. GPM was responsible for the ethics application and with HJS for registration of the protocol on clinicaltrials.gov. HJS, GPM, AB, and LM designed the statistical analysis. GPM, AB, WW, AD, HJS are responsible for recruitment. HJS, WW and GPM are responsible for the coordination of the study. GPM and HJS drafted the manuscript. NS, FX, AB, and JB reviewed early drafts. EAA and PA-C provided methodological input in the study design. All listed authors participated in the writing and revision of the manuscript and approved the its final version.

References

1. Institute of Medicine Committee on Standards for Developing Trustworthy Clinical Practice G. *Clinical Practice Guidelines We Can Trust.* Washington (DC): National Academies Press (US); 2011.
2. Schünemann HJ. Guidelines 2.0: do no net harm-the future of practice guideline development in asthma and other diseases. *Curr Allergy Asthma Rep.* 2011;11(3):261-268.

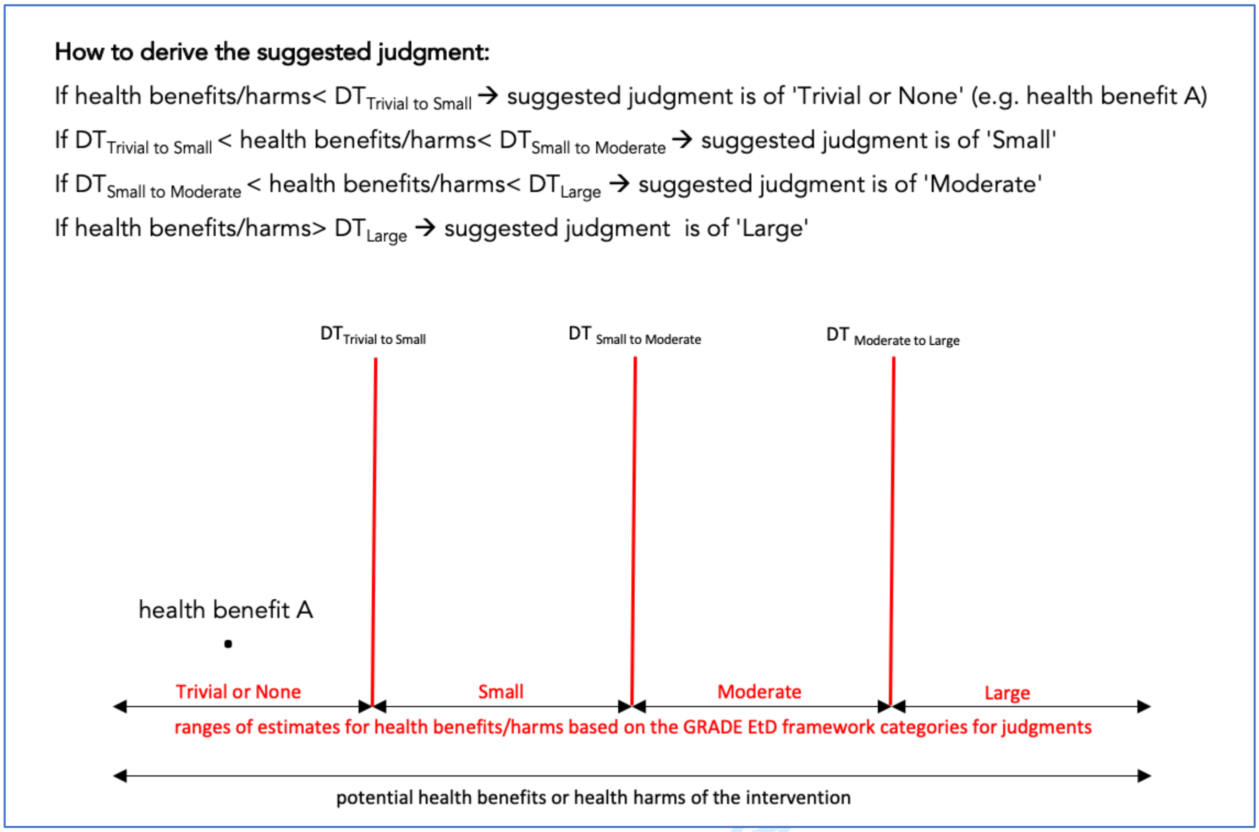
3. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. *BMC Health Serv Res.* 2004;4(1):38. doi: 10.1186/1472-6963-1184-1138.
4. Alonso-Coello P, Schunemann HJ, Moberg J, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ.* 2016;353:i2016.
5. 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.* 2016;353:i2089.
6. Marcucci M, Sinclair JC. A generalised model for individualising a treatment recommendation based on group-level evidence from randomised clinical trials. *BMJ Open.* 2013;3(8).
7. McCabe C, Claxton K, Culyer AJ. The NICE cost-effectiveness threshold: what it is and what that means. *Pharmacoeconomics.* 2008;26(9):733-744.
8. Pauker SG, Kassirer JP. Therapeutic decision making: a cost-benefit analysis. *N Engl J Med.* 1975;293(5):229-234.
9. Hultcrantz M, Mustafa RA, Leeflang MMG, et al. Defining ranges for certainty ratings of diagnostic accuracy: a GRADE concept paper. *J Clin Epidemiol.* 2020;117:138-148.
10. Hultcrantz M, Rind D, Akl EA, et al. The GRADE Working Group clarifies the construct of certainty of evidence. *J Clin Epidemiol.* 2017;87:4-13.
11. Faraone SV. Interpreting estimates of treatment effects: implications for managed care. *P T.* 2008;33(12):700-711.
12. Johnston BC, Ebrahim S, Carrasco-Labra A, et al. Minimally important difference estimates and methods: a protocol. *BMJ Open.* 2015;5(10):e007953.
13. Chan AW, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ (Clinical research ed).* 2013;346:e7586.
14. Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. *J Clin Epidemiol.* 2013;66(2):173-183. doi: 110.1016/j.jclinepi.2012.1008.1001. Epub 2012 Oct 1030.
15. Baldeh T, Saz-Parkinson Z, Muti P, et al. Development and use of health outcome descriptors: a guideline development case study. *Health Qual Life Outcomes.* 2020;18(1):167.
16. Abramson JH. WINPEPI updated: computer programs for epidemiologists, and their teaching potential. *Epidemiol Perspect Innov.* 2011;8(1):1.
17. Abdi H. Holm’s sequential Bonferroni procedure. In: Salkin N, ed. *Encyclopedia of research design.* Thousand Oaks, California2010:1-8.
18. Alper BS, Oettgen P, Kunnamo I, et al. Defining certainty of net benefit: a GRADE concept paper. *BMJ Open.* 2019;9(6):e027445.

19. Puhan MA, Singh S, Weiss CO, Varadhan R, Boyd CM. A framework for organizing and selecting quantitative approaches for benefit-harm assessment. *BMC Med Res Methodol.* 2012;12:173.

20. Siebert U. When should decision-analytic modeling be used in the economic evaluation of health care? *The European Journal of Health Economics, formerly: HEPAC.* 2003;4(3):143-150.

21. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;296(13):716-721.

Supplemental file 1



How to derive the suggested judgment:

If health benefits/harms< $DT_{\text{Trivial to Small}}$ → suggested judgment is of 'Trivial or None' (e.g. health benefit A)

If $DT_{\text{Trivial to Small}}$ < health benefits/harms< $DT_{\text{Small to Moderate}}$ → suggested judgment is of 'Small'

If $DT_{\text{Small to Moderate}}$ < health benefits/harms< DT_{Large} → suggested judgment is of 'Moderate'

If health benefits/harms> DT_{Large} → suggested judgment is of 'Large'

$DT_{\text{Trivial to Small}}$
$DT_{\text{Small to Moderate}}$
$DT_{\text{Moderate to Large}}$
health benefit A
Trivial or None
Small
Moderate
Large
ranges of estimates for health benefits/harms based on the GRADE EtD framework categories for judgments
potential health benefits or health harms of the intervention

Figure 1

Figure 1 legend: The availability of three DTs ($DT_{\text{Trivial/Small}}$, $DT_{\text{Small/Moderate}}$, $DT_{\text{Moderate/Large}}$) would allow to discriminate between the four GRADE EtD framework categories for judgments. For a given health benefit/harm, the suggestion on the judgment would depend on how the estimate of health benefits/harms compares to the DTs. In this example, the health benefit A lies on the left (is smaller) of the $DT_{\text{Trivial/Small}}$ which would suggest that the judgment of 'Trivial or None' would be more appropriate than the others.

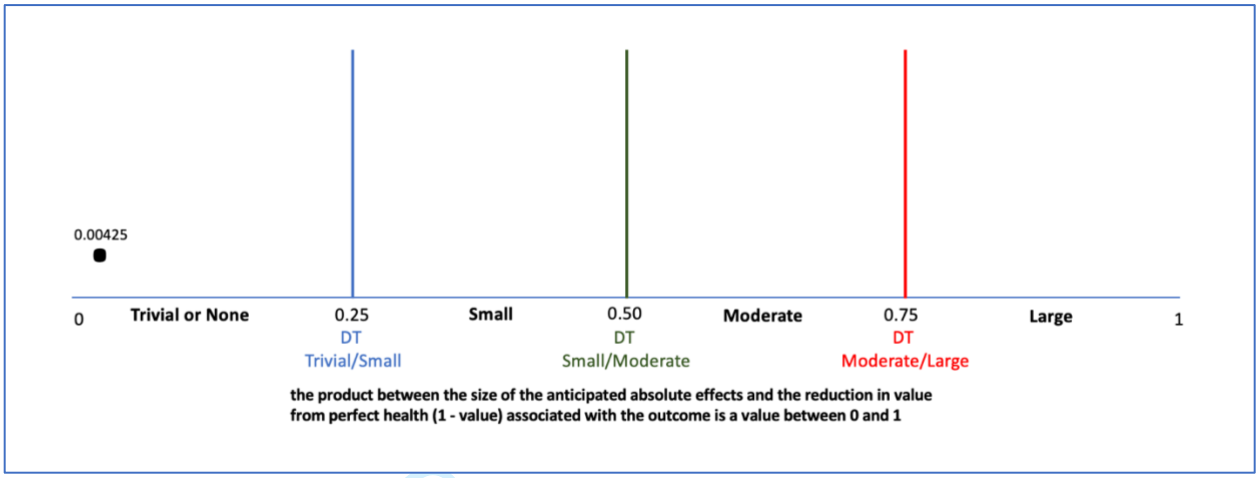
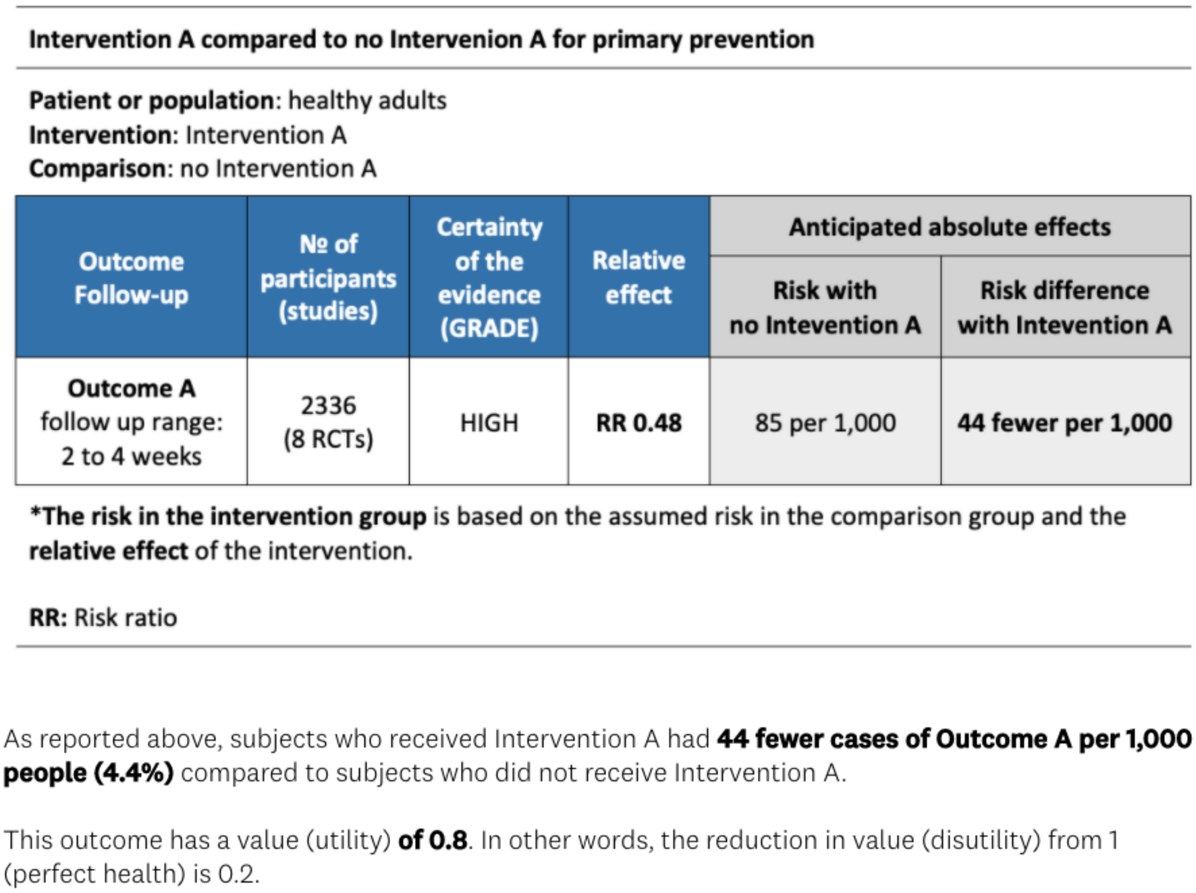


Figure 2

Figure 2 legend: We assumed to have known DTs ($DT_{Trivial/Small}$ = 0.25, $DT_{Small/Moderate}$ = 0.50, $DT_{Moderate/Large}$ = 0.75) and wanted to assign to one of the 4 EtD categories the health benefit of an intervention showing an anticipated absolute effect of 17 fewer per 1000 on an outcome valued 0.75. Following the proposed approach, we calculated the result of the product (score) of the size of anticipated effects (Absolute Risk Difference, ARD) and the reduction in value from perfect health (1 - outcome's value) associated with the outcome under evaluation. In this example, the following approach (ARD * (1 - outcome's value) = (17/1000)*(1 - 0.75)) resulted in the value of 0.00425. We then plotted this value and obtained the suggested judgment according to the DTs approach that, in this case, would be of 'Trivial or None' considering that the calculated value is smaller than the $DT_{Trivial/Small}$. In case of judgments made considering more than one outcome, it would be required to calculate the aforementioned product for each of the outcomes under evaluation and derive an aggregate score defined as the sum of all the individual scores.

Under the assumption that the outcomes are independent, this score could be then compared to the DTs and used to obtain a suggestion on the judgment.

Supplemental file 2



Intervention A compared to no Intervenion A for primary prevention

Patient or population: healthy adults
Intervention: Intervention A
Comparison: no Intervention A

Outcome Follow-up	№ of participants (studies)	Certainty of the evidence (GRADE)	Relative effect	Anticipated absolute effects	
				Risk with no Intevention A	Risk difference with Intevention A
Outcome A follow up range: 2 to 4 weeks	2336 (8 RCTs)	HIGH	**RR 0.48**	85 per 1,000	**44 fewer per 1,000**

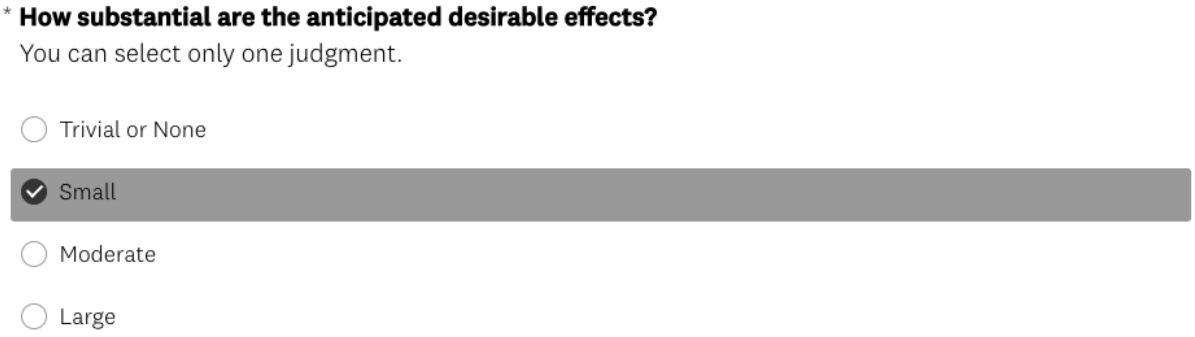
***The risk in the intervention group** is based on the assumed risk in the comparison group and the **relative effect** of the intervention.

RR: Risk ratio

As reported above, subjects who received Intervention A had **44 fewer cases of Outcome A per 1,000 people (4.4%)** compared to subjects who did not receive Intervention A.

This outcome has a value (utility) **of 0.8**. In other words, the reduction in value (disutility) from 1 (perfect health) is 0.2.

Figure 1 - Example of a case-scenario



* **How substantial are the anticipated desirable effects?**

You can select only one judgment.

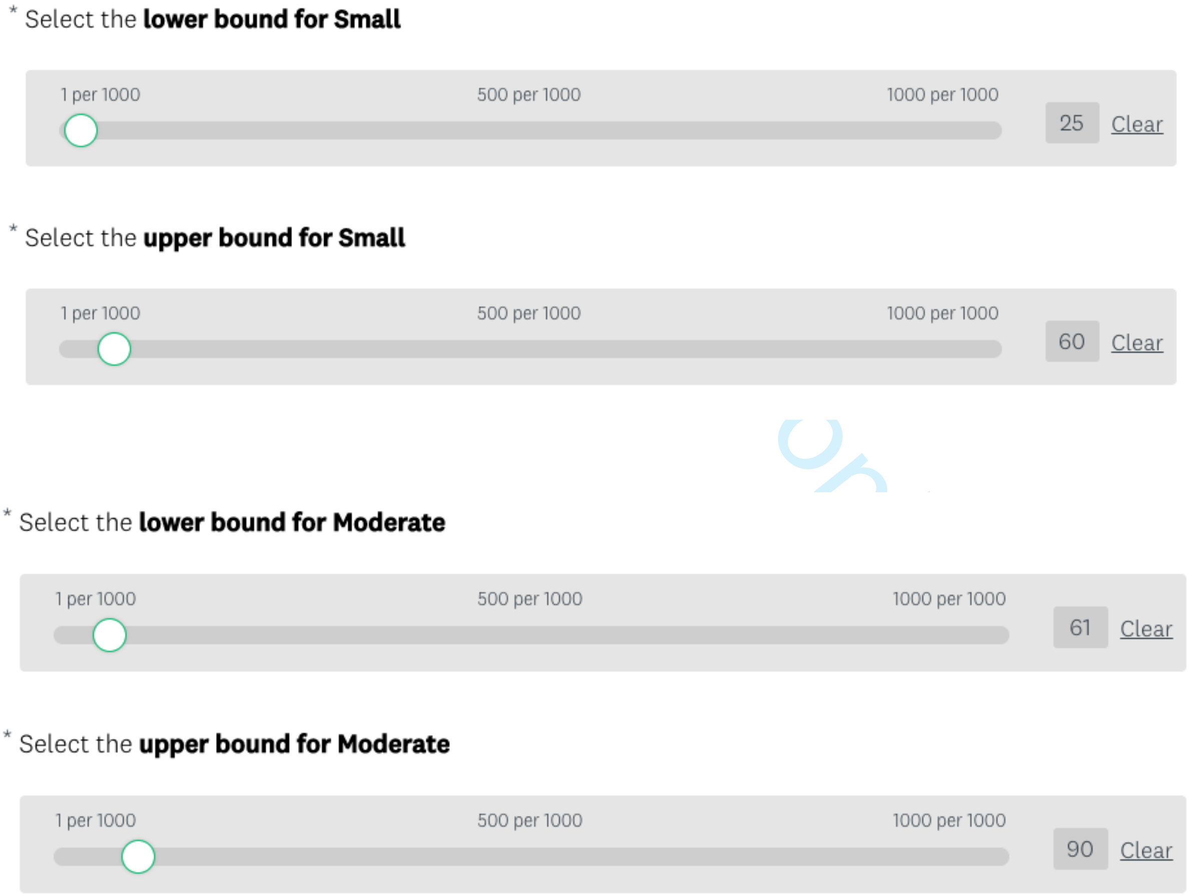
○ Trivial or None

☑ Small

○ Moderate

○ Large

Figure 2 - Judgment on health benefits



* Select the **lower bound for Small**

25 Clear

* Select the **upper bound for Small**

60 Clear

* Select the **lower bound for Moderate**

61 Clear

* Select the **upper bound for Moderate**

90 Clear

Figure 3 - Selection of ranges for judgments of Small and Moderate

Example of calculation of DTs based on survey data

In the examples shown in Figures 2 and 3 we assumed that, after having evaluated a given case-scenario (ARD of 44 events fewer per 1000 on an outcome valued 0.8), a survey participant rated the hypothetical ranges of ARD for judgments of 'Small' and 'Moderate' of from 25 fewer per 1000 to 60 fewer per 1000, and of from 61 fewer per 1000 to 90 fewer per 1000, respectively. We used this data to derive the ranges of ARD for judgments of 'Trivial or None' and of 'Large' (table below).

boundaries of ranges described in Figure 3 value of the outcome = 0.8						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	range
0 per 1000	24 per 1000	25 per 1000	60 per 1000	61 per 1000	90 per 1000	more than 90 per 1000

Table 1 - Ranges of sizes of effects (ARD)

For each range of ARD, we calculated the product between range boundaries and the reduction in value from perfect health (1 - outcome's value) for the outcome associated. Then, we derived the DTs as follow: $DT_{Trivial/Small}$ equal to the the product calculated from the lower bound for the judgment of 'Small', $DT_{Small/Moderate}$ equal to average of the products calculated from the upper bound for the judgment of 'Small' and the lower bound for the judgment of 'Moderate', and $DT_{Moderate/Large}$ equal to the smallest number larger than the mean of the products calculated from the upper bound for the judgment of 'Moderate'.

product values = ARD * (1- outcome's value))						
Trivial or None		Small		Moderate		Large
range of values						
lower bound	upper bound	lower bound	upper bound	lower bound	upper bound	any value
(0/1000)*0.2 0	(24/1000) *0.2 0.0048	(25 /1000) *0.2 0.005	(60/1000) *0.2 0.012	(61/1000)*0.2 0.0122	(90/1000)*0.2 0.018	bigger than (90/1000)*0.2 >0.018

Table 2 - Ranges of product values

Using the data from Table 2, the DTs would result as follow: $DT_{Trivial/Small}$ = 0.005, $DTs_{mall/Moderate}$ = 0.0121, DTModerate$_{/Large =}$ 0.0180001.

Supplemental file 3

Descriptive statistics

The presented preliminary analysis is based on survey data collected between May 1st and July 21st, 2020. Our dissemination strategy allowed recruitment of 75 participants who contributed a total of 295 ratings. Fifty-six survey participants had a background in research (74.6%) and 36 were healthcare professionals (50.6%). Thirty-four respondents (45.3%) were members of academia. Other major groups were participants from HTA organizations and professional societies (13.3% and 18.6%, respectively). Participants were equally randomized to case-scenarios descriptive of desirable and undesirable health effects (144/295, 49%; 151/295, 51%, respectively) and completed the entire exercise in the majority of cases (68/75, 90.7%). Detailed descriptive characteristics of survey respondents and ratings are shown in Tables 1 and 2, respectively.

Characteristic [a]	**Respondents, n = 75**
Background [a]	n (%)
Clinical/Health Professional	38 (50.6)
Policymaking	6 (8.0)
Research	56 (74.6)
Teaching	18 (24.0)
Administrative	3 (4.0)

Patient representative	2 (2.6)
Other	3 (4.0)
Degree [a]	
Degree in Nursing (RN)	1 (1.3)
Medical School (MD)	30 (4.0)
Master of Sciences (MSc)	17 (22.6)
Master of Public Health (MPH)	9 (0.12)
Doctor of Philosophy (PhD)	25 (33.3)
None	2 (2.6)
Other	5 (6.6)
Formal Training in health research methodology/epidemiology/biostatistics	
Never completed	12 (16.0)
Completed some form of formal training but do not have a graduate degree	30 (40.0)
Earned a MSc degree	16 (21.3)
Earned a PhD degree	16 (21.3)
Not available	1 (1.4)
Organization [a]	
Cochrane collaboration	13 (17.3)
GRADE Working Group	16 (21.3)

World Health Organization	1 (1.4)
Guidelines International Network (G-I-N)	-
Health Technology Assessment (HTA) organization	10 (13.3)
Academia	34 (45.3)
Professional society	14 (18.6)
Familiarity with the Evidence to Decision framework	
Not at all familiar	5 (6.6)
Not so familiar	9 (12.0)
Somewhat familiar	16 (21.3)
Very familiar	30 (40.0)
Extremely familiar	8 (10.6)
Not available	7 (9.5)
Previous participation in guideline development groups	
Yes	52 (69.3)
No	18 (24.0)
Not available	5 (6.6)
Primary role in the guideline development group [a]	
Clinical Chair	5 (6.6)
Chair for methods	15 (19.8)
Guideline methodologist	29 (38.6)
Panel member	15 (19.8)

Topic or content expert	7 (9.5)
Patient representative	2 (2.6)
Systematic review author	26 (34.6)
Expert in Health Technology Assessment	3 (4.0)

Values represent the number and in parentheses the percentage.

[a] Percentages do not add up to 100 because respondents could choose more than one option.

Table 1: Characteristics of survey respondents

Characteristics of ratings collected through the survey	n (%)
Total number of ratings collected	295
Missing data (expected ratings - collected ratings/expected ratings)	17/312 (0.054)[a]
randomized to a scenario showing desirable effects	144/295 (49)
randomized to a scenario showing undesirable effects	151/295 (51)
randomized to the outcome of death	73/295 (25)
randomized to the outcome of major stroke	66/295 (22)
randomized to the outcome of pulmonary embolism	55/295 (19)
randomized to the outcome of moderate diarrhea	63/295 (21)
based on the outcome of mild nausea/vomiting	38/295 (13)

a. 73 participants were randomized to 4 case-scenarios, 2 were mistakenly randomized to 10.

Table 2: Descriptive statistics of survey ratings

Table 3 describes the estimates of DTs that were derived from survey ratings through the joint measure of absolute effects and outcome values. For example, an outcome valued as 0.8, these thresholds would indicate that the effect of an intervention preventing 30 events of that outcome per 1000 should be categorized as trivial (since 0.03*(1-0.8)) =0.006 is smaller than T1). More details about the calculation of the DTs are available in Appendix 1 (Table 1).

Decision Threshold			**95% Confidence Interval**	
	Estimate	Std. Deviation	Lower Bound	Upper Bound
T1: Trivial/Small	0.0165	0.0467	0.0059	0.0271
T2: Small/Moderate	0.0312	0.0601	0.0176	0.0448
T3: Moderate/Large	0.0577	0.0781	0.0400	0.0754

Table 3: Estimates of DTs

Primary analysis

Our analysis showed a difference in the estimates between T1 and T2 (mean difference [MD] -0.0147; 95% CI -0.0201 to -0.0093; p<0.001) and T2 and T3 (mean difference [MD] -0.0264; 95% CI -0.0544 to -0.0062; p<0.001).

Within-participant analyses

The analyses showed that at a respondent level there was no difference between DTs derived from judgments on benefits and from those on harms: $T1_{benefit}=T1_{harms}$ (mean difference [MD] -0.0040 ; 95% CI -0.0195 to 0.0116 ; p=0.615) ; $T2_{benefits}=T2_{harms}$ (mean difference [MD] -0.0124; 95% CI -0.0313 to 0.0064 ; p=0.196); $T3_{benefit}=T3_{harms}$ (mean difference [MD] -0.0209; 95% CI -0.0451 to 0.0033; p=0.090).

Subgroup analyses

Our subgroup analyses showed a difference in the estimates between T1 and T2, and T2 and T3 also in DTs derived from subgroup of ratings identified by outcome, direction of interventions' effects, and prior participation to guideline development groups. No difference was observed in the estimates between T1 and T2 in those with no experience with the EtD (mean difference [MD] -0.0046; 95% CI -0.0100 to 0.0006; p=0.810) and between T2 and T3 in those who had no training in epidemiology (mean difference [MD] -0.0056; 95% CI -0.0218 to 0.0106; p=0.483).

Sensitivity analyses

The findings of the sensitivity analyses conducted by excluding raters who provided incoherent thresholds (n=3; T1/T2 mean difference [MD] -0.0143; 95% CI -0.0192 to -0.0094; p<0.001; T2/T3 mean difference [MD] -0.0291; 95% CI -0.0417 to -0.0165; p<0.001) or who were presumed outliers (n=10; T1/T2 mean difference [MD] -0.0096 ; 95% CI -0.0113 to -0.0078; p<0.001; T2/T3 mean difference [MD] -0.0194; 95% CI -0.0240 to -0.0148; p<0.001) were similar to that of the primary analysis.

Assessment of order effects

The analyses suggest no difference between DTs derived from participants who evaluated a high-value outcome (i.e. moderate diarrhea) in the first iteration compared to those who evaluated a low-value outcome (i.e. death) first. Similarly, there was no difference in the DTs depending on whether the first judgment made was 'Small' or 'Large'.