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'.