Journal of Clinical Epidemiology 110 (2019) 23−33

Journal of Clinical Epidemiology

ORIGINAL ARTICLE

Structured decision-making drives guidelines panels' recommendations "for" but not "against" health interventions

Benjamin Djulbegovic[a,b,*], Tea Reljic[c], Shira Elqayam[d], Adam Cuker[e], Iztok Hozo[f], Qi Zhou[g], Shelly-Anne Li[h], Paul Alexander[g], Robby Nieuwlaat[g], Wojtek Wiercioch[g], Holger Schünemann[g], Gordon Guyatt[g]

[a]*Department of Supportive Care Medicine, City of Hope, 1500 East Duarte Rd, Duarte, CA, USA*
[b]*Department of Hematology, City of Hope, 1500 East Duarte Rd, Duarte, CA, USA*
[c]*Department of Medicine, University of South Florida, 12901 Bruce B Downs Blvd, Tampa, FL, USA*
[d]*Department of Medicine, De Montfort University, Leicester, UK*
[e]*Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA*
[f]*Department of Mathematics, Indiana University, Gary, IN, USA*
[g]*Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada*
[h]*Faculty of Nursing, University of Toronto, Canada*

Accepted 7 February 2019; Published online 16 February 2019

Abstract

Background and Objectives: The determinants of guideline panels' recommendations remain uncertain. The objective of this study was to investigate factors considered by members of 8 panels convened by the American Society of Hematology (ASH) to develop guidelines using Grading of Recommendations Assessment, Development, and Evaluation (GRADE) system.

Study Design and Setting: Web-based survey of the participants in the ASH guidelines panels. Analysis: two-level hierarchical, random-effect, multivariable regression analysis to explore the relation between GRADE and non-GRADE factors and strength of recommendations (SOR).

Results: In the primary analysis, certainty in evidence [OR = 1.83; (95CI% 1.45−2.31)], balance of benefits and harms [OR = 1.49 (95CI% 1.30−1.69)] and variability in patients' values and preferences [OR = 1.47 (95CI% 1.15−1.88)] proved the strongest predictors of SOR. In a secondary analysis, certainty of evidence was associated with a strong recommendation [OR = 3.60 (95% CI 2.16−6.00)] when panel members recommended "for" interventions but not when they made recommendations "against" interventions [OR = 0.98 (95%CI: 0.57−1.8)] consistent with "yes" bias. Agreement between individual members and the group in rating SOR varied (kappa ranged from −0.01 to 0.64).

Conclusion: GRADE's conceptual framework proved, in general, to be highly associated with SOR. Failure of certainty of evidence to be associated with SOR against an intervention, suggest the need for improvements in the process.

Keywords: Practice guidelines; Clinical recommendations; Evidence based medicine; Decision theory; Group decision making; GRADE

1. Introduction

Trustworthy evidence-based clinical practice guidelines (CPGs) [1−3] represent one approach to addressing suboptimal clinical decision-making [4−7]. In fact, measuring adherence to CPGs is one of the key approaches to quality improvement [7,8].

If CPGs are to improve health outcomes, they must be developed using rigorous methodological principles [2] as advanced during the last 20 years through the various systems of rating the certainty of evidence and strength of recommendations (SOR) [9−13]. Of these systems, the Grading

Funding sources have no involvement in any of the activities conducted in this article.

Conflict of interest: Although most authors have worked on development of GRADE system, the authors declare no conflict of interest related to the content and analysis of this article.

Authors' statements: B.D. has conceptualized the study, received funding, and wrote the first draft. G.G. helped with the grant proposal and revised the first draft of the article, which has then be shared and approved by all authors. T.R., S.A.L., P.A., R.N., and W.W. have helped with data collection. H.S. and A.C. have helped with the study logistics and provided the intellectual input from the guidelines process perspective. S.E. provided the perspective from psychology of decision-making. I.H. and Q.Z. helped with data analysis. B.D. serves as a guarantor.

* Corresponding author. Tel.: +626-218-7502; fax: +626-256-8798.
E-mail address: bdjulbegovic@coh.org (B. Djulbegovic).

https://doi.org/10.1016/j.jclinepi.2019.02.009

What is new?

Key findings

- The Grading of Recommendations Assessment, Development, and Evaluation (GRADE) guidelines system specifies factors that guidelines panels ''should'' take into considerations when issuing recommendations. However, many other (non-GRADE) factors may also affect recommendations.
- To what extent GRADE vs. non-GRADE factors influence guidelines panels' decision-making remains uncertain.
- We found that GRADE factors affect guidelines decision-making process more than non-GRADE factors, likely because of the effect of instructions provided within structured GRADE Evidence-to-Decision (EtD) framework. Consistent with principles of evidence-based medicine, we confirmed relation between the certainty of evidence and strength of recommendations (SOR).
- The findings remained robust when panels issued recommendations ''for'' health interventions. However, when the panels generated recommendations ''against'' health interventions, the relation between certainty of evidence and SOR disappeared, pointing to the existence of so called ''yes'' bias (people acquiesce to ''yes'' statements more readily than to ''no'' statements).
- Even within highly structured GRADE process, the panel members demonstrated ''variability'' in their ''individual'' responses (kappa between individual panel members and the group consensus vote for SOR ranged from very poor [−0.01] to moderate [0.64]),
- Depending on the analytical model, some non-GRADE factors were also associated with the SOR issued by the panels. Different non-GRADE factors were associated with recommendations ''for'' vs. ''against'' health interventions. However, age/clinical experience of the panelists remained statistically significant across all models.

What this adds to what was known?

- This quantitative analysis of 8 panels confirms that GRADE instruction given within EtD structured framework results in consideration of GRADE factors as intended by the GRADE system.
- The system does not, however, appear to give consistent results when the panels issue recommendation ''for'' vs. ''against'' health intervention.
- In addition, individual member ''assessment'' often considerably differs from the group, consensus vote.

What is the implication and what should change now?

- Guideline panels that place a high value on adherence to the GRADE system should consider use of EtD framework in developing their recommendations.
- To avoid ''yes'' bias, guidelines developers should, in most instances, express all recommendations as a vote ''for'' instead of ''against'' recommendations.
- Exploration of reasons why panel members are sometimes in agreement and sometimes not may inform the need for additional strategies such as more extensive training in GRADE to reduce variability.

of Recommendations Assessment, Development, and Evaluation (GRADE) approach represents the most transparent, rigorously developed, and documented to date [2,14], which is endorsed by over 100 professional organizations, including the World Health Organization, the Cochrane Collaboration, and a number of leading American organizations [15].

GRADE has identified a number of factors that CPG panels *should* consider when making recommendations, including the certainty of evidence, the balance between benefits and harms, patient values and preferences, resource and cost considerations, as well as issues related to acceptability, feasibility, and health equity [16,17]. Although GRADE provides a normative system for how CPG panels ''ought to'' develop guidelines, in what manner guideline panels *actually* make their judgments remains unclear.

Despite the breadth of GRADE's specified considerations, many additional factors not formally captured in the GRADE system may affect panel judgments. Broadly, these factors include [18] (a) decision features, or characteristics of the decision/recommendation (e.g., high-stake vs. low-stake clinical recommendations, such as developing guidelines for vulnerable populations in a politically charged atmosphere), (b) situational/contextual factors (e.g., time pressure, cognitive load, role in the panel as chair, methodologist, panelists, etc.), and (c) individual characteristics of the decision-maker (e.g., age) [18−21].

How and to what extent these additional factors contribute to the decision-making process remains unclear. In addition, development of CPGs ultimately relies on the ''group judgment'' of the panel. Despite its importance, we know little about how the group consensus relates to the individual judgments of its members. We therefore designed a study addressing the interplay of group and individual processes in

real-life decision-making and provide the first analysis of a guidelines panel's decision-making process.

2. Materials and methods

We studied the process in 8 panels convened by the American Society of Hematology (ASH) to develop guidelines for the management of the following conditions: heparin-induced thrombocytopenia, thrombophilia, venous thromboembolism (VTE) in pregnancy, VTE in pediatric populations, optimal management of anticoagulation therapy, VTE in patients with cancer, treatment of VTE, and management of immune thrombocytopenia.

A series of webinars introduced panel members to the GRADE system. During a number of conference calls, the panel members defined and prioritized the clinical questions, guided a systematic review team in the collection and analysis of the relevant evidence, and in some cases discussed prevoting results for guideline questions. All recommendations included judgments (weak or conditional vs. strong) in favor of intervention (I) or comparator (C) for a given outcome (O) related to the population of interest (P).

Each panel developed final recommendations through "group consensus" during face-to-face meetings using GRADE's structured evidence-to-decision (EtD) framework [16,17]. For each recommendation, the panel made explicit judgments for each factor in the framework; before doing so, panelists reviewed a summary of these judgments (see Appendix for the actual presentation framework). Each panel member completed a survey detailing their judgments related to relevant GRADE factors and the final recommendations during the meeting or shortly thereafter.

We used frequencies and percentages to describe characteristics of panels and panel members participating in the development of the ASH guidelines. We explored GRADE and non-GRADE factors that might influence the panels' recommendations, all defined a priori as per current literature [18–21]. The GRADE factors included certainty of evidence supporting recommended intervention, balance between benefits and harms, assessment of variability or uncertainty in patients' value and preferences, and resources that the panel judged may be needed to implement

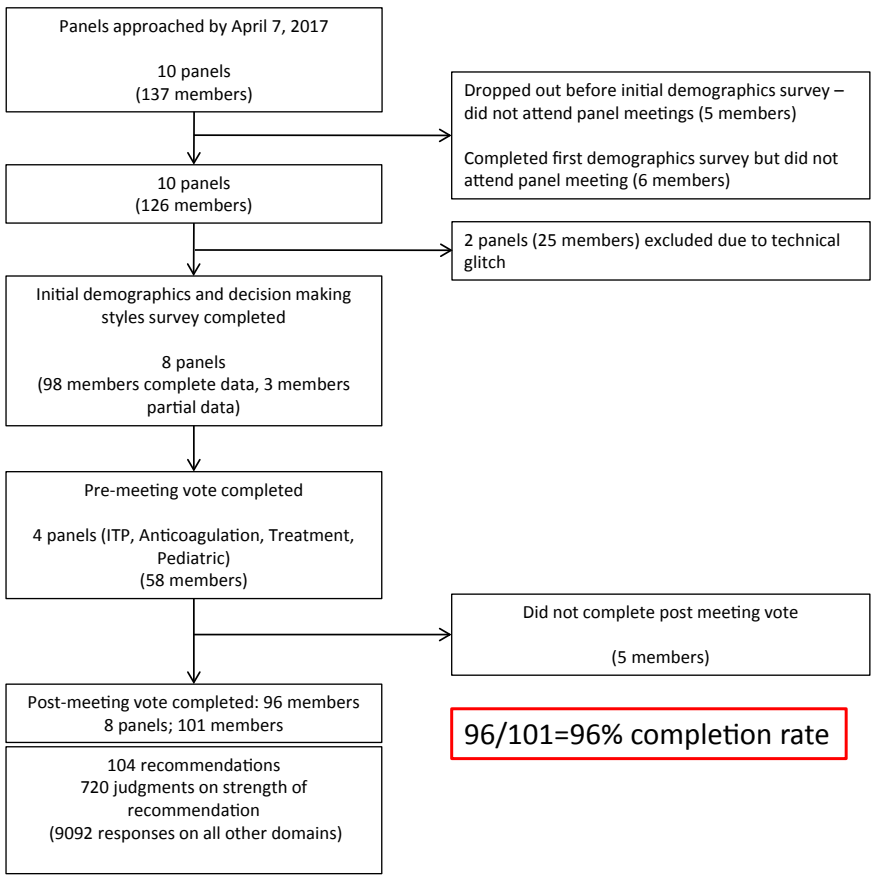

Fig. 1. Overview of data collection process. Data were collected from the guidelines panels convened by the American Society of Hematology (ASH) to develop guidelines for the management of the following conditions: (1) prevention of venous thromboembolism (VTE) in surgical hospitalized patients, (2) prevention of VTE in medical patients, (3) heparin-induced thrombocytopenia (HIT), (4) management of thrombophilia, (5) VTE in the context of pregnancy, (6) VTE in pediatric populations, (7) optimal management of anticoagulation therapy, (8) VTE in patients with cancer, (9) treatment of VTE, and (10) management of immune thrombocytopenia (ITP). Unfortunately, data collection from two panels was not recorded due to technical glitches. The final analysis included data from 8 panels (see text).

recommendations. Certainty of evidence was coded on a 1 to 4 scale, with 1 indicating very low certainty of evidence and 4 high certainty of evidence. Variability or uncertainty in patient's values and preferences (V&P) was coded on a scale 1 to 4 (1 = important, 2 = possibly important, 3 = probably not important, 4 = not important). Judgments on use of resource/costs were coded on a 5-point Likert scale (1 = large costs, 2 = moderate costs, 3 = neither, 4 = moderate savings, 5 = large savings). Judgments on the balance of intervention benefit/harms was coded on 5-point Likert scale (1 = favors the comparison, 2 = probably favors the comparison, 3 = does not favor either, 4 = probably favors the intervention; 5 = favors the intervention). Each of these categorical variables was treated in the analysis as continuous assuming the equivalent interval effects among the consecutive scores.

Although the panels, using the EtD framework, also considered issues of acceptability, feasibility, and health equity (see Appendix), limiting response burden on panelists precluded our considering these issues.

Non-GRADE variables included the following:

(a) Individual characteristics of the decision maker: age, sex, experience, expertise, and cognitive styles, that is, propensities to favor one decision-making or reasoning approach over another [20]. The latter was assessed by administration of instruments to measure objectivism, that is, tendency to seek empirical information to support decision-making; intolerance of uncertainty [22]; maximizing-satisficing, that is, assessment of tendency for individual to use reasoning processes that will lead to making a good vs. best possible decision [23]; propensity to engage in analytical, rational thinking vs. experiential-intuitive thinking [20,24]; and tendency to experience regret about making a decision [21]. These instruments have proved valid and applicable to assessment of physicians' decision-making [20].

(b) Characteristics of the decision/recommendation: recommendations made for vulnerable populations (children, women, inner city, rural, ethnic minority, low-income), reports of feeling pressured to issue certain type of recommendations/to conform with the group because of the potentially politically sensitive nature of guideline recommendations.

(c) Situational/contextual factors related to a given guideline recommendation: individual panel member's conflicts of interest, role in the panel (chair, methodologist, patient representative, panel member). The Supplementary material provides details related to all variables and instruments.

We constructed a model relating these variables to the SOR as either strong, weak, or no recommendation. We repeated the analysis according to the direction of the recommendation (''for'' vs. ''against''), omitting

Table 1. Participant characteristics by panel[a]

Variable	Overall N (%)
Number of participants	101
Age, median (quartile1, quartile 3); range	48.0 (41, 56) (28–78)
Sex	
Male	56 (55.4)
Female	45 (44.6)
Role	
Chair	8 (7.9)
Methodologist	16 (15.8)
Clinician	66 (65.4)
Patient representatives	11 (10.9)
Panel	
Anticoagulation	13 (12.9)
Cancer	15 (14.9)
Heparin-induced thrombocytopenia	11 (10.9)
Immune thrombocytopenia	16 (15.8)
Pediatric	15 (14.9)
Pregnancy	10 (9.9)
Thrombophilia	7 (6.9)
Treatment	14 (13.9)
Country of origin	
United States	52 (51.5)
Canada	25 (24.8)
Netherlands	5 (5.0)
Italy	3 (3.0)
United Kingdom	3 (3.0)
Germany	3 (3.0)
Australia	3 (3.0)
Austria	2 (2.0)
Argentina	1 (1.0)
Belgium	1 (1.0)
Denmark	1 (1.0)
New Zealand	1 (1.0)
Switzerland	1 (1.0)
Years of experience[b], median (quartile 1, quartile 3); (range)	18 (11,26) (2–49)
Self-reported level of experience[b]	
Higher than others	46 (55.4)
About same as others	32 (38.6)
Lower than others	5 (6.0)
How many patients with similar condition do you treat per month[b]	
None	13 (16.3)
1–5	14 (17.5)
6–10	3 (3.8)
11–15	7 (8.8)
More than 15	43 (53.8)

[a] Included in the final analysis; there was no statistically significant difference between these participants and those that were excluded from the analysis (see Fig 1).

[b] Questions regarding professional experience were only answered by clinicians.

questions in which panels did not make a recommendation [25,26].

We used a two-level hierarchical, mixed multivariable logistic, and ordered regression analysis to account both for panel level factors and individual level factors. Thus, judgments of recommendations were clustered within panel members, and these were clustered within panels.

To compare individual panel recommendations with the group consensus, we calculated the agreement (kappa statistics and correlations) between each individual panel member's average judgment (weighted by the number of their responses) and the group consensus recommendation. To account both for sampling error and the variability among the panels, we pooled kappa statistics across all panels by meta-analyzing it under a random-effects model [27]. To estimate the random effects of the panels on the percentage of the total residual variance in each individual member's voting pattern, we estimated intraclass correlations (ICCs) after running the two-level mixed effect logistic regressions. All calculations were performed using STATA, version 15 [28], and verified in SAS, version 9.4, by a second investigator.

3. Results

Fig. 1 presents an overview of the data collection process. Table 1 presents characteristics of the panels and panel members participating in the development of the

Table 2. Association between decision-making factors and the strength of recommendations

Mixed-effect model [ordered logistic regression] Fixed effect	Dependent variable: neither for/against; weak for/weak against; strong for/strong against odds ratio (OR) with 95% confidence interval (CI)
Role: chair (reference category)	
Methodologist	1.47 (95% CI .57–3.80; $P = .43$)
Patient representative	0.38 (95% CI .13–1.18; $P = .09$)
Panel member	1.07 (95% CI .58–1.98; $P = .83$)
Vulnerable population	
Yes vs. no	1.27 (95% CI .81–2.00; $P = .3$)
Pressured to ''vote'' certain way	0.84 (95% CI .31–2.31; $P = .74$)
Recused from ''voting''	
Yes vs. no	1.36 (95% CI .92–2.03; $P = .13$)
Age (per decade)	**1.79 (95% CI 1.2–2.84; $P = 0.005$)**
Sex	
Female vs. male	1.10 (95% CI .75–1.62; $P = .61$)
Experience (years in management of given condition)	0.97 (95% CI .926–1.00; $P = .09$)
Expertise (considers oneself with higher, same or low expertise than most other experts)	0.74 (95% CI .51–1.09; $P = .13$)
Exposure (# of patients per month with given condition)	0.90 (95% CI .770–1.06; $P = .21$)
Objectivism (tendency to seek empirical information)	1.29 (95% CI .78–2.17; $P = .34$)
Tendency toward rational (analytical) thinking	0.66 (95% CI .36–.1.10; $P = .11$)
Tendency toward experiential-intuitive thinking	0.95 (95% CI .67–1.34; $P = .76$)
Satisficing (tendency to accept ''good'' enough solution)	1.13 (95% CI .63–2.02; $P = .68$)
Maximizing (decision difficulty)—degree difficulty experienced when making choices among abundant options	1.05 (95% CI .78–1.41; $P = .766$)
Maximizing (alternative search)—tendency to expand resources in search for best possible solution	1.08 (95% CI .81–1.43; $P = .606$)
Intolerance of uncertainty	**0.57 (95% CI .37–.86;$P = 0.008$)**
Regret of making a wrong recommendation	0.99 (95% CI .98–1.02; $P = .93$)
Certainty in evidence	**1.84 (95% CI 1.46–2.31;$P < 0.0001$)**
Importance of patients' values and preferences	**1.48 (95% CI 1.15–1.89;$P = 0.002$)**
Balance between benefits and harms	**1.49 (95% CI 1.31–1.70;$P < 0.0001$)**
Importance of cost and resources	1.06 (95% CI .86–1.28; $P = .56$)
Random intercepts	
Panel (variance)	0.62 (95% CI .19–2.05)
Participant within panel (variance)	2.18×10^{-34}

Bolded text refers to statistically significant findings.

ASH guidelines. Typical panelists were male hematologists around 50 years of age from the United States with approximately 20 years of clinical experience.

The panel meetings occurred between November 2016 and August 2017 in Washington, D.C., and lasted between 15 and 26 hours across 2 days (median = 10 hours per day).

Of 21 variables potentially associated with the SOR, 3 GRADE and 2 non-GRADE factors displayed statistically significant association at the conventional $P < 0.05$ levels (Table 2). Panel members' judgment of certainty of the evidence [OR = 1.84 (95%CI 1.46–2.31)] proved the strongest predictor—the more confident the panel members were regarding the certainty of the evidence, the more inclined they were to issue strong recommendations.

Other factors associated with strong recommendations included age (per decade) [OR = 1.79 (95CI% 1.2–2.84)] (older panel members were more inclined to make strong recommendations), followed by balance of benefits and harms [OR = 1.49 (95CI% 1.30–1.69)] (when balance favors intervention, the panelists are more likely to issue a strong recommendation), the uncertainty or variability in patients' V&P [OR = 1.47 (95CI% 1.15–1.88)] (the less uncertainty or variability, more likely the panel was to issue a strong recommendation), and intolerance of uncertainties [OR = 0.57 (95CI% 0.37–0.86] (more intolerance, less likely a strong recommendations).

Table 3 showed the logistic regression analysis when the panel members issued recommendations "strong for" vs.

Table 3. Association between decision-making factors and the strength of recommendations

Mixed-effect model [logistic regression] **Fixed effect**	**Dependent variable: weak for; strong for odds ratio (OR) with 95% confidence interval (CI)**
Role: chair (reference category)	
Methodologist	**0.06 (95% CI 0–.85; *P* = .04)**
Patient representative	0.64 (95% CI .03–12.4; *P* = .77)
Panel member	1.15 (95% CI .23–5.89; *P* = .87)
Vulnerable population	
Yes vs. no	1.83 (95% CI .56–5.91; *P* = .31)
Pressured to "vote" certain way	2.27 (95% CI .21–24.9; *P* = .5)
Recused from "voting"	
Yes vs. no	1.18 (95% CI .4–3.45; *P* = .76)
Age (per decade)	**2.6 (95% CI .99–7.93; *P* = .079)**
Sex	
Female vs. male	0.55 (95% CI .20–1.48; *P* = .24)
Experience (years in management of given condition)	**0.89 (95% CI .76–.99; *P* = .047)**
Expertise (consider oneself with higher, same or low expertise than most other experts)	1.40 (95% CI .566–3.49; *P* = .47)
Exposure (# of patients per month with given condition)	0.89 (95% CI .61–1.31; *P* = .57)
Objectivism (tendency to seek empirical information)	1.64 (95% CI .44–6.10; *P* = .46)
Tendency toward rational (analytical) thinking	0.51 (95% CI .15–1.76; *P* = .23)
Tendency toward experiential-intuitive thinking	0.89 (95% CI .37–2.15; *P* = .79)
Satisficing (tendency to accept "good" enough solution	2.23 (95% CI .61–8.1; *P* = .23)
Maximizing(decision difficulty)-degree difficulty experienced when making choices among abundant options	**2.14 (95% CI .99–4.65; *P*=.05)**
Maximizing(alternative search)-tendency to expand resources in search for best possible solution	0.65 (95% CI .31–1.35; *P* = .25)
Intolerance of uncertainty	0.4 (95% CI .13–1.21; *P* = .11)
Regret of making a wrong recommendation	0.99 (95% CI .95–1.05; *P* = .89)
Certainty in Evidence	**3.61 (95% CI 2.17–6.01; *P* < 0.001)**
Importance of patients' values and preferences	**2.33 (95% CI 1.34–4.07; *P* = .003)**
Balance between benefits and harms	**18.3 (95% CI 7.68–43.7; *P* = .000)**
Importance of cost and resources	**1.83 (95% CI 1.25–2.67; *P* = .002)**
Random intercepts	
Panel (variance)	3.14×10^{-32}
Participant within panel (variance)	1.00×10^{-33}

Bolded text refers to statistically significant findings.

"weak for" in favor of a given intervention. In this analysis, judgment about balance between benefits and harms was associated with an OR of 18.3 [95% CI 7.67–43.7] for recommendations in favor of the intervention. The second strongest predictor was certainty of evidence [OR = 3.61 (95%CI 2.17–6.01)]. When panels judged that certainty in evidence is high and benefits outweigh harms in favor of intervention over the comparator, the predicted probability of issuing strong recommendation in favor of the intervention exceeded 90% (Fig. 2).

Assessment of patients' values and preferences as well as consideration of costs/resources were also highly statistically significant but at a somewhat lower odds ratio (Table 3). Methodologists, in comparison with panel chairs, were less likely to issue strong recommendations [OR = 0.06 (95CI% 0.04–0.85)]. Three non-GRADE factors also show statistically significant or borderline significant associations (Table 3). As in the main analysis (Table 2), older panel members were more inclined to make strong recommendations [OR = 2.6 (95% CI 0.99–7.93)]. More experienced panel members tended to issue weaker recommendations [OR = 0.891 (95CI% 0.795–0.98)] (see Discussion), whereas the tendency to use a maximizing cognitive style when faced with decision difficulties was associated with OR = 2.14 (95% CI .99–4.65) (Table 3).

Table 4 outlines the analysis when the panel members issued recommendations "strong against" vs. "weak against" health interventions (Fig. 3). In this analysis, only one formal GRADE factor (the importance of patients' V&P) had an effect, whereas 5 non-GRADE factors displayed statistically significant association. More experienced panel members, those with higher intolerance of uncertainty and those with propensity toward analytical thinking tended to issue weaker recommendations against the intervention. On other hand, being a methodologist, older, recused from voting due to a conflict of interest, or issuing guidelines for a vulnerable population were associated with strong recommendation against intervention.

Agreement between individual panel members and the group regarding SOR ranged from poor (kappa ranging from: −0.01 to 0.03; 2 panels) to fair (kappa range: 0.21–0.47; 4 panels) to moderate (kappa = 0.64; 1 panel) (Fig. 4). Agreement of judgments related to voting "for" and "against" the intervention was somewhat better 0.37 (95% CI 0.16–0.58) and 0.42 (95% CI: 0.19–0.64), respectively (data not shown).

Finally, we calculated ICC to determine the extent of the overall variation in the response of the panel members. The results varied with the analyses: in the main analysis (Table 2), we found negligible correlation between individual vote and the panel voting pattern (ICC = 0.06), but in the analysis that omitted the recommendations in which the panel did not issue a recommendation determining the strength of association and direction of the vote, ICC was 0.50 in recommendations "for" and 0.48 in recommendations "against".

4. Discussion

We report the first study evaluating the impact of the GRADE system and non-GRADE factors that could impact on guidelines panel members' decision-making. Overall, we showed that factors associated with GRADE's conceptual framework were, in general, highly associated with SOR. A secondary analysis suggested, however, that certainty of evidence may have little or no influence on SOR when a panel makes recommendations against an

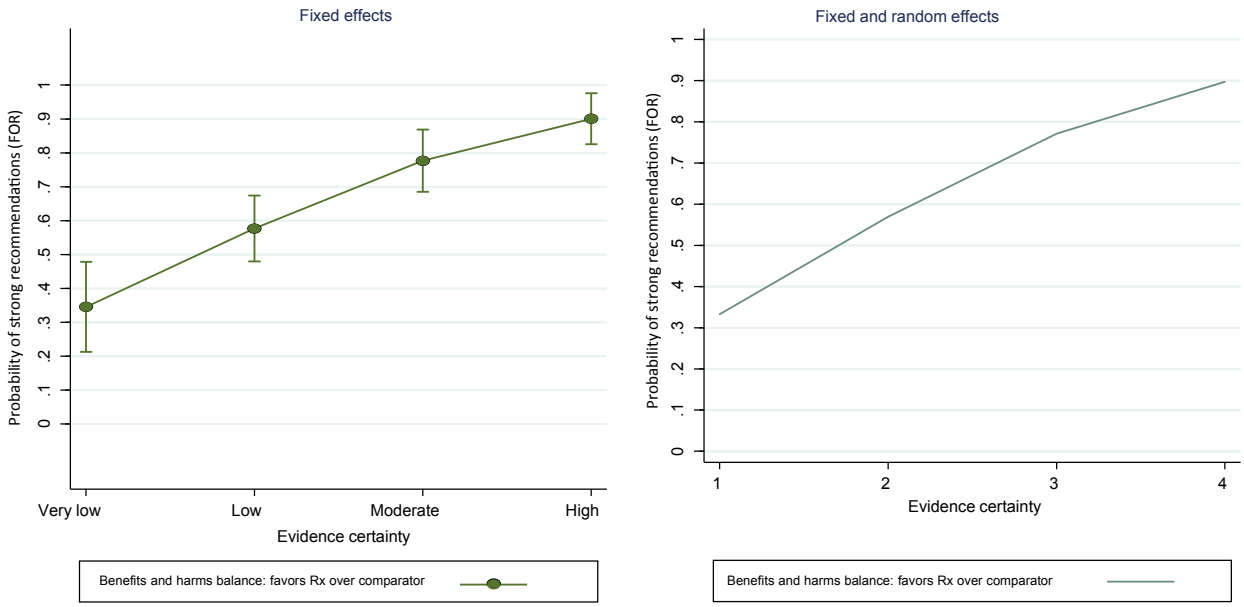

Fig. 2. Effect of certainty of evidence and judgments about the balance of benefits and harms (in favor of intervention over comparator). The vertical line around the each point denotes a 95% confidence interval.

Table 4. Association between decision-making factors and the strength of recommendations

Mixed-effect model [logistic regression] Fixed effect	Dependent variable: weak against; strong against odds ratio (OR) with 95% confidence interval (CI)
Role: chair (reference category)	
methodologist	**11.7 (95% CI 1.50–91.4; $P = .019$)**
Patient representative	1.28 (95% CI .070–23.4; $P = .87$)
Panel member	2.34 (95% CI .558–9.83; $P = .25$)
Vulnerable population	
Yes vs. no	**3.97 (95% CI 1.30–12.1; $P = .015$)**
Pressured to ''vote'' certain way	0.34 (95% CI .036–3.28; $P = .35$)
Recused from "voting"	
Yes vs. no	**3.34 (95% CI 1.29–8.65;$P = .013$)**[a]
Age (per decade)	**4.8 (95% CI 1.8–12.76; $P = .002$)**
Sex	
Female vs. male	0.56 (95% CI .247–1.43; $P = .25$)
Experience (years in management of given condition)	**0.87 (95% CI .79–.96;$P = .007$)**
Expertise (consider oneself with higher, same or low expertise than most other experts)	0.46 (95% CI .18–1.15; $P = .09$)
Exposure (# of patients per month with given condition)	0.84 (95% CI .58–1.20; $P = .33$)
Objectivism (tendency to seek empirical information)	1.15 (95% CI .28–4.63; $P = .84$)
Tendency toward rational (analytical) thinking	**0.215 (95% CI .063–.729; $P = .014$)**
Tendency toward experiential-intuitive thinking	1.32 (95% CI .60–2.90; $P = .49$)
Satisficing (tendency to accept ''good'' enough solution	1.31 (95% CI .36–4.72; $P = .68$)
Maximizing(decision difficulty)-degree difficulty experienced when making choices among abundant options	1.45 (95% CI .74–2.85; $P = .28$)
Maximizing(alternative search)-tendency to expand resources in search for best possible solution	1.00 (95% CI .55–1.84; $P = .99$)
Intolerance of uncertainty	**0.16 (95% CI .05–.49; $P = .001$)**
Regret of making a wrong recommendation	OR = 1.01 (95% CI .97–1.06; $P = .62$)
Certainty in Evidence	OR = 0.98 (95% CI .57–1.68; $P = .94$)
Importance of patients' values and preferences	**2.26 (95% CI 1.32–3.87; $P = .003$)**
Balance between benefits and harms	0.78 (95% CI .52–1.23; $P = .31$)
Importance of cost and resources	1.01 (95% CI .60–1.69; $P = .97$)
Random intercepts	
Panel (variance)	1.54 (95% CI .380–6.23)
Participant within panel (variance)	1.47×10^{-36}

Bolded text refers to statistically significant findings.

[a] when interactions with the certainty of evidence was taken into account, OR = 4.63 (95% CI 0.814 to 26.38; $P = 0.084$).

intervention. We also detected statistical association between SOR and non-GRADE factors but, aside from age/clinical experience, these varied across statistical models.

The main findings likely reflect the effect of instructions [29,30] because of use of the highly structured GRADE EtD framework [31]. Adherence to structure is typically seen with high-ability participants (such as expert panelists) who can follow instructions that require cognitive effort and suppress the influence of other factors and prior beliefs [29,32]. The findings extend the observations from our qualitative analysis [33] that policy-makers and users of guidelines who apply GRADE methods may expect that the guideline panels will not only rely on GRADE factors but use the cognitive processes that facilitate decision-making according to the GRADE instructions. Nevertheless, individual characteristics such as age, experience, intolerance of uncertainty, and propensity toward analytical thinking were also, in some models, associated with SOR. Theoretically, a type of a task and instructions can activate cognitive processes to align them toward accomplishing stated goals [34]. For example, the importance of intolerance of uncertainty can be seen as a response to the underlying clinical uncertainties that activate analytical reasoning processes that prompted development of guidelines in the first place [35].

Our results regarding the importance of certainty of evidence are consistent with observations in two smaller

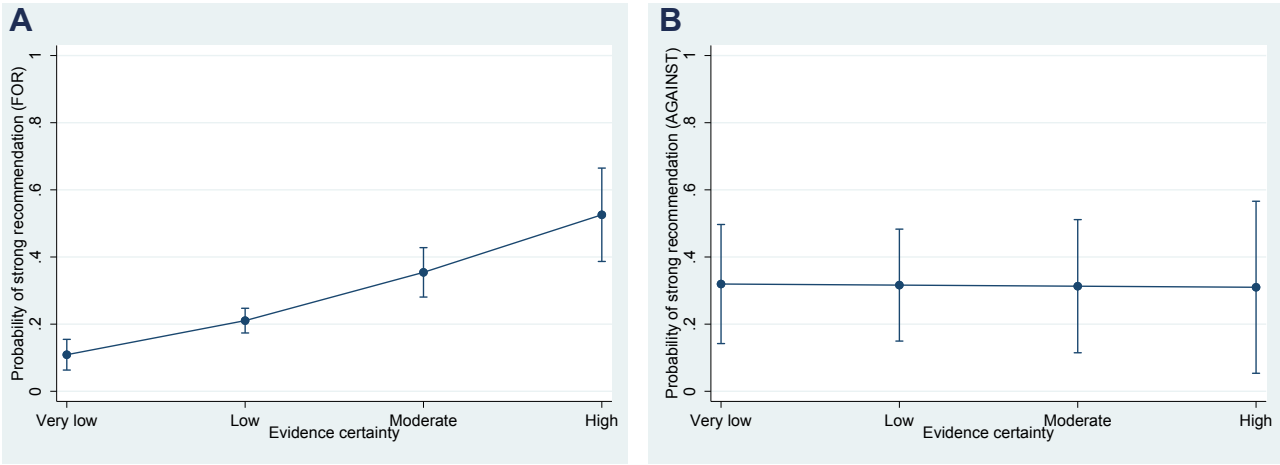

Fig. 3. A relationship between the quality (certainty) of underlying evidence and the probability of issuing of a strong recommendation FOR (A) vs. AGAINST (B) a given health intervention. The vertical line around each point denotes a 95% confidence interval. The results remained the same even though panelists were instructed to align strength of recommendations with direction of recommendations (the reminders were originally issued orally, but it was included in the survey for the last 5 panels).

studies [36,37]. The results provide empirical verification of the key EBM normative principle regarding the relationship between the credibility of underlying evidence and willingness to endorse a health intervention [1]: when certainty of evidence is high, we can expect that most panelists will issue strong recommendations.

However, this relationship disappeared when the panel members "voted" "against" health intervention. Potential explanations for this finding include (1) the "yes/for" bias [25,26,38–40]: according to dual process theory of acquiescence, people are overall slower to respond to "no" than to "yes". "Yes" ("for") responses tap into "feeling of rightness" heuristic that the answer is correct: is automatic, effortless (type 1 process), which is activated much faster than effortful (type 2 processes) associated with processing of "no" ("against") responses. [38–40]; (2) Voting "against" an intervention is cognitively more challenging because people need to mentally simulate the consequences of two contradictory assessments—certainty of evidence, which moves from very low to high in "positive" direction and strength of recommendation "against" the intervention, which goes in the opposite direction. This often occurs when cognitive resources are depleted [41,42] as when people are tired and decision-making occurs in time-constraint settings, which characterize most human engagements including guidelines development process; (3) in a number of cases, the question was formulated as a "vote" against intervention without explicit description of a comparator, which may have introduced a reference class problem (i.e., when reference category is not well specified, people's estimates are often incorrect) [43,44]; (4) GRADE paradigmatic situations that justify strong recommendations despite low certainty evidence, may have occurred more in recommendations "against" than "for" interventions [45].

As in all research, we cannot exclude the possibility that some associations we observed may be simply due to chance. For example, as in our earlier study [20], we detected that effect of age and experience went in the opposite direction, which we judged to be a spurious association. This occurred because in medicine, as in many professions, age and experience are positively correlated ($r = 0.85$ in this study) across individuals, making it difficult to isolate the unique influence of a given variable on the third variable.

Our results also provide empirical support for the importance of managing conflict of interests [46,47]—the panel members who were required to recuse themselves, had they

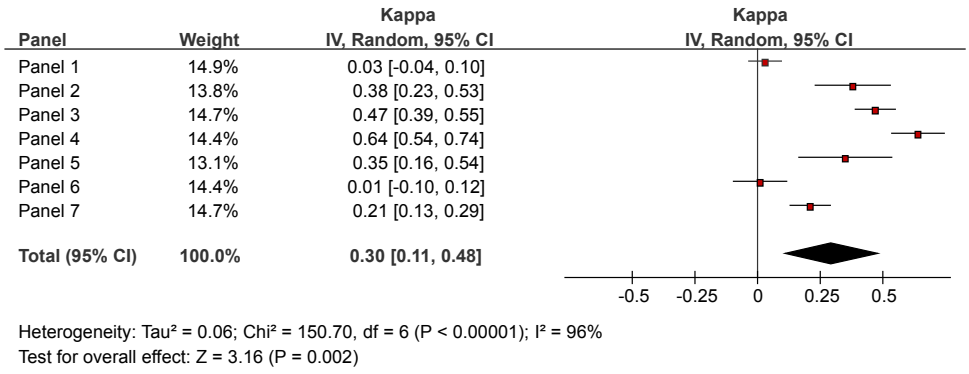

Fig. 4. Agreement in judgments related to strength of recommendations between individual panel members and the group judgments.

been allowed to vote, would have registered different views from those of their colleagues.

The frequent low agreement between judgments of individual panel members' and the group consensus related to SOR raises the possibility that the apparent consensus represents individual panel members' conforming to the group [48,49], particularly since more than 50% of discussion was dominated by chairs and cochairs [33]. In a classic article on opinions and social pressure, Asch warned that "Consensus is an indispensable condition in a complex society, but consensus, to be productive, requires that each individual contribute independently out of experience and insight. When consensus is produced by conformity, the social process is polluted" [48].

Nevertheless, fewer than 2% of participants (Table 2) reported that they felt any pressure to conform to the group vote. Earlier studies suggested that when instructions that clearly operationalize procedures are provided, agreement on assessments such as the certainty of evidence becomes high [50]. Lack of familiarity with the GRADE system (despite introductory lectures about GRADE) and the complexity of the judgment inherent in making recommendations may explain the low agreement we observed. An alternative explanation is that many of the decisions were close calls in the panels where agreement was low—and fewer when agreement was high.

Another explanation for low agreement arises from our observation that variability in V&P was associated with SOR despite, as we have reported previously, only 1% of the discussion was devoted to this issue [33]. It is possible that panelists had different views of the extent of diversity and uncertainty in V&P, views that they did not express in group discussion. This suggests that chairs of guideline using GRADE should insist on repeated discussion of V&Ps issues.

4.1. Strengths and limitations

A strength of our study is that it is the first to assess the decision-making of guidelines panels in natural, real-life setting. At the same time, the observational design precluded experimental control of the variables that may allow drawing stronger inferences. Hence, future studies will be necessary to establish the generalizability of our findings. Nevertheless, our findings represent first initial insights into how guideline decision-making works in real life, and suggests possible improvements in the process.

In conclusion, we found that policy-makers and users of guidelines who apply GRADE methods may expect that the guideline panels will rely on GRADE factors. However, low agreement between individual panel members and group consensus suggests that the process can be improved, perhaps by further operationalization of GRADE criteria, by better training of the panel members in GRADE methodology, and by framing, as far as is possible, all recommendations in terms of voting "for" instead of "against" a given health intervention.

CRediT authorship contribution statement

Benjamin Djulbegovic: Conceptualization, Funding acquisition, Writing - original draft. **Tea Reljic:** Investigation. **Shira Elqayam:** Data curation, Formal analysis, Investigation, Methodology, Writing - review & editing. **Adam Cuker:** Resources. **Iztok Hozo:** Formal analysis. **Qi Zhou:** Formal analysis. **Shelly-Anne Li:** Investigation. **Paul Alexander:** Investigation. **Robby Nieuwlaat:** Investigation. **Wojtek Wiercioch:** Investigation. **Holger Schünemann:** Resources. **Gordon Guyatt:** Writing - review & editing, Funding acquisition.

Acknowledgments

This project was supported by grant number R01HS024917 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. PI: Dr. Djulbegovic. The authors thank Robert Kunkle and his staff from the American Society of Hematology (ASH) for helping facilitate their project and Robert Hamm of the University of Oklahoma Health Sciences Center for useful feedback on an earlier version of the manuscript. They also wish to thank all panelists participating in development of the ASH guidelines on venous thromboembolism and immune thrombocytopenia for agreeing to take part in our study. They also thank the ASH leadership for supporting this project.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jclinepi.2019.02.009.

References

[1] Djulbegovic B, Guyatt GH. Progress in evidence-based medicine: a quarter century on. Lancet 2017;390:415−23.
[2] Graham R, Mancher M, Wolman DM, Greenfield S, Steinberg E, editors. Clinical practice guidelines we can trust. Washington, DC: Institute of Medicine, National Academies Press; 2011.
[3] Vandvik P, Brandt L, Alonso-Coello P, Treweek S, Akl E, Kristiansen A. Creating clinical practice guidelines we can trust, use, and share: a new era is imminent. Chest 2013;144:381−9.
[4] Keeney R. Personal decisions are the leading cause of death. Oper Res 2008;56(6):1335−47.
[5] Medicine IIo. Variation in health care spending: Target decision making, not geography. Washington, DC: The National Academies Press.; 2013.
[6] Pronovost PJ. Enhancing physicians' use of clinical guidelines. JAMA 2013;310:2501−2.
[7] Djulbegovic B. A framework to bridge the gaps between evidence-based medicine, health outcomes, and improvement and implementation science. J Oncol Pract 2014;10(3):200−2.
[8] Rosenthal MB. Physician payment after the SGR — the new meritocracy. N Engl J Med 2015;373(13):1187−9.

[9] Berkman NDLK, Ansari M, McDonagh M, Balk E, Whitlock E, Reston J, et al. AHRQ publication No. 13(14)-EHC130-EF. In: Grading the strength of a body of evidence when assessing health care interventions for the effective health care program of the agency for healthcare research and quality: an update. Methods guide for comparative effectiveness reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). Rockville, MD: Agency for Healthcare Research and Quality; 2013:2013.
[10] Ebell MH, Siwek J, Weiss BD, Woolf SH, Susman J, Ewigman B, et al. Strength of recommendation taxonomy (SORT): a patient-centered approach to grading evidence in the medical literature. Am Fam Physician 2004;69(3):548−56.
[11] Harbour R, Miller J, Group fSIGNGR. A new system for grading recommendations in evidence-based guidelines. BMJ 2001;323:334−6.
[12] Petitti DB, Teutsch SM, Barton MB, Sawaya GF, Ockene JK, DeWitt T, et al. Update on the methods of the U.S. preventive services task force: insufficient Evidence. Ann Intern Med 2009;150:199−205.
[13] Hill J, Bullock I, Alderson P. A summary of the methods that the national clinical guideline centre uses to produce clinical guidelines for the national bnstitute for health and clinical gxcellence. Ann Intern Med 2011;154:752−7.
[14] Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;336:924−6.
[15] GRADE Working Group. Organizations that have endorsed or that are using GRADE. Available at http://www.gradeworkinggroup.org/society/index.htm;. Accessed April 21, 2016.
[16] Alonso-Coello P, Oxman AD, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, 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.
[17] Alonso-Coello P, Schünemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, 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.
[18] Appelt KC, Milch KF, Handgraaf MJJ, Weber EU. The decision making individual differences inventory and guidelines for the study of individual differences in judgment and decision-making research. Judgm Decis Making 2011;6:252−62.
[19] Hastie R, Dawes RM. Rational choice in an uncertain world. 2nd ed. Los Angeles: Sage Publications, Inc.; 2010.
[20] Djulbegovic B, Beckstead JW, Elqayam S, Reljic T, Hozo I, Kumar A, et al. Evaluation of physicians' cognitive styles. Med Decis Making 2014;34:627−37.
[21] Djulbegovic M, Beckstead J, Elqayam S, Reljic T, Kumar A, Paidas C, et al. Thinking styles and regret in physicians. PLoS One 2015;10:e0134038.
[22] Budner S. Intolerance of ambiguity as a personality variable. J Pers 1962;30:29−50.
[23] Turner BM, Rim HB, Betz NE, Nygren TE. The Maximization Inventory. Judgm Decis Making 2012;7(1):48−60.
[24] Pacini R, Epstein S. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. J Pers Soc Psychol 1999;76:972−87.
[25] Gilbert DT. How mental systems selieve. Am Psychol 1991;46(2):107−19.
[26] Tversky A, Koehler DJ. Support theory: a nonextensional representation of subjective probability. Psychol Rev 1994;101:547−67.
[27] Sun S. Meta-analysis of Cohen's kappa. Health Serv Outcomes Res Method 2011;11:145−63.
[28] STATA, ver. 14 [computer program]. College Station, TX: Stata Corporation; 2013.
[29] Evans JS. Logic and human reasoning: an assessment of the deduction paradigm. Psychol Bull 2002;128(6):978−96.
[30] Heit E, Rotello CM. Traditional difference-score analyses of reasoning are flawed. Cognition 2014;131(1):75−91.
[31] Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ 2008;336:1049−51.
[32] Evans JSBT, Handley SJ, Neilens H. The influence of cognitive ability and instructional set on causal conditional inference. Q J Exp Psychol (Hove) 2010;63(5):892−909.
[33] Li S-A, Alexander PE, Reljic T, Cuker A, Nieuwlaat R, Wiercioch W, et al. Evidence to decision framework provides a structured "roadmap" for making GRADE guidelines recommendations. J Clin Epidemiol 2018;104:103−12.
[34] Phillips WJ, Fletcher JM, Marks AD, Hine DW. Thinking styles and decision making: a meta-analysis. Psychol Bull 2016;142(3):260−90.
[35] Djulbegovic B, Hozo I, Greenland S. Uncertainty in clinical medicine. In: Gifford F, editor. Philosophy of Medicine (Handbook of the Philosophy of Science). London: Elsevier; 2011:299−356.
[36] Djulbegovic B, Trikalinos TA, Roback J, Chen R, Guyatt G. Impact of quality of evidence on the strength of recommendations: an empirical study. BMC Health Serv Res 2009;9:120.
[37] Djulbegovic B, Kumar A, Kaufman RM, Tobian A, Guyatt GH. Quality of evidence is a key determinant for making a strong guidelines recommendation. J Clin Epidemiol 2015;68:727−32.
[38] Thompson VA, Evans JSBT, Campbell JID. Matching bias on the selection task: it's fast and feels good. Think Reason 2013;19(3):431−52.
[39] Shynkaruk JM, Thompson VA. Confidence and accuracy in deductive reasoning. Mem Cognit 2006;34(3):619−32.
[40] Knowles ES, Condon CA. Why people say "yes": a dual-process theory of acquiescence. J Pers Soc Psychol 1999;77(2):379−86.
[41] De Neys W. Dual processing in reasoning: two systems but one reasoner. Psychol Sci 2006;17(5):428−33.
[42] De Neys W, Verschueren N. Working memory capacity and a notorious brain teaser: the case of the Monty Hall Dilemma. Exp Psychol 2006;53(2):123−31.
[43] Gigerenzer G. Content-blind norms, no norms, or good norms?A reply to Vranas. Cognition 2001;81:93−103.
[44] Gigerenzer G. What are natural frequencies? BMJ 2011;343:d6386.
[45] Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines 15: ioing from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol 2013;66:726−35.
[46] George JN, Vesely SK, Woolf SH. Conflicts of interest and clinical recommendations: comparison of two concurrent clinical practice guidelines for primary immune thrombocytopenia developed by different methods. Am J Med Qual 2014;29(1):53−60.
[47] Institute of Medicine (US). Committee on conflict of interest in medical research E, and practice. In: Lo B, Field MJ, editors. Conflict of Interest in Medical Research, Education, and Practice. Washington (DC): National Academies Press (US); 2009.
[48] Asch SE. Opinions and social pressure. Sci Am 1955;193(5):31−5.
[49] Asch SE. Studies of independence and conformity. I A minority of one against unanimous majority. Psychol Monogr Gen Appl 1956;70(9):1−70.
[50] Kumar A, Miladinovic B, Guyatt GH, Schünemann H, Djulbegovic B. GRADE guidelines system is reproducible when instructions are clearly operationalized even among the guidelines panel members with limited experience with GRADE. J Clin Epidemiol 2016;75:115−8.