BMJ Open

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018593
Article Type:	Research
Date Submitted by the Author:	10-Jul-2017
Complete List of Authors:	Agoritsas, Thomas; University Hospitals of Geneva, Division of General Internal Medicine & Division of Clinical Epidemiology; McMaster University Faculty of Health Sciences, Department of Health Research Methods, Evidence, and Impact Merglen, Arnaud; University Hospitals of Geneva, Division of General Pediatrics Heen, Anja; Innlandet Hospital Trust-division Gjøvik, Department of Internal Medicine Kristiansen, Annette; Inland hospital trust, Internal medicine, Gjøvik Neumann, Ignacio; Pontificia Universidad Catolica de Chile, Department of Internal Medicine; McMaster University, Department of Health Research Methods, Evidence, and Impact, Brito, Juan; Mayo Clinic Minnesota, Department of Medicine and Knowledge and Evaluation Research Unit, Brignardello-Petersen, Romina; McMaster University, Department of Health Research Methods, Evidence, and Impact Alexander, Paul; McMaster University, Department of Health Research Methods, Evidence, and Impact Rind, David; Institute for Clinical and Economic Review Vandvik, Per; Norwegian Knowledge Centre for the Health Services, Guyatt, Gordon; Mcmaster University, Department of Health Research Methods, Evidence, and Impact
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Epidemiology
Keywords:	Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Thomas Agoritsas, MD, PhD [1,2]

Arnaud Merglen, MD, MSc [3]

Anja Fog Heen, MD [4]

Annette Kristiansen, MD, PhD [4]

Ignacio Neumann, MD, PhD [2,5]

Juan P Brito, MD, MSc [6]

Romina Brignardello-Petersen, DDS, MSc, PhD [2, 7]

Paul E Alexander, MSc, PhD [2]

David M Rind, MD, MSc [8]

Per O. Vandvik, MD, PhD [4,9]

Gordon H Guyatt, MD, MSc [2]

Affiliations:

[1] Division of General Internal Medicine & Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland.

[2] Department of Health Research Methods, Evidence, and Impact, McMaster University, Faculty of Health Sciences, Hamilton, Ontario, Canada

[3] Division of General Pediatrics, University Hospitals of Geneva & Faculty of Medicine, University of Geneva, Geneva, Switzerland.

[4] Department of Internal Medicine, Innlandet Hospital Trust-division Gjøvik, Norway

[5] Department of Internal Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

[6] Division of Endocrinology, Diabetes, Metabolism and Nutrition, Department of Medicine and Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, USA

[7] Faculty of Dentistry, University of Chile, Chile.

[8] Institute for Clinical and Economic Review, Boston, MA, USA

[9] Institute of Health and Society, Faculty of Medicine, University of Oslo, Norway

*** Correspondence to**: Thomas Agoritsas, MD, PhD (thomas.agoritsas@gmail.com)

Division of General Internal Medicine & Division of Clinical Epidemiology

Department Internal Medicine, Rehabilitation and Geriatrics (DMIRG)

University Hospitals of Geneva, Rue Gabrielle-Perret-Gentil 4,

1211 Genève 14, Switzerland

Phone: +41 79 55 34 543

Keywords: Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine.

Abstract: 285 words / **Manuscript**: 2955 words

Tables: 5 / Supplementary Files: 2

ABSTRACT

Introduction

UpToDate is widely used by clinicians worldwide and includes more than 9,400 recommendations that apply the GRADE framework. GRADE guidance warns against strong recommendations when certainty of the evidence is low or very low (discordant recommendations), but has identified five paradigmatic situations in which discordant recommendations may be justified.

Objectives

Our objective was to document the strength of recommendations in UpToDate and assess the frequency and appropriateness of discordant recommendations.

Design

Analytic survey of all recommendations in UpToDate

Methods

We identified all GRADE recommendations in UpToDate, and examined their strength (strong or weak) and certainty of the evidence (high, moderate, or low certainty). We identified all discordant recommendations as of January 2015, and pairs of reviewers independently classified them either into one of the five appropriate paradigms or into one of three categories inconsistent with GRADE guidance.

Results

UpToDate included 9451 GRADE recommendations, of which 6501 (68.8%) were formulated as weak recommendations and 2950 (31.2%) as strong. Among the strong,

844 (28.6%) were based on high certainty in effect estimates, 1,740 (59.0%) on moderate certainty, and 366 (12.4%) on low certainty. Of the 349 discordant recommendations 204 (58.5%) were judged appropriate (consistent with one of the five paradigms); we classified 47 (13.5%) as good practice statements; 38 (10.9%) misclassified the evidence as low certainty when it was at least moderate; and 60 (17.2%) warranted a weak rather than a strong recommendation.

Conclusion

The proportion of discordant recommendations in UpToDate is small, and the proportion that is truly problematic (strong recommendations that would best have been weak) very small. Clinicians should nevertheless be cautious, and look for clear explanations – in UpToDate and elsewhere – when guidelines offer strong recommendations based on low certainty evidence.

Strengths and limitations of this study

- We assessed the strength of recommendations in the largest known sample of recommendations using GRADE (N=9451) addressing a wide array of clinical fields.
- We used a taxonomy to appraise discordant recommendations that has been successfully implemented in two prior assessments of clinical practice guidelines.
- We based our assessment solely on information published in UpToDate, while authors of the topics may have considered other factors in deciding to issue a discordant recommendation.
- UpToDate topics are narrative in nature and do not include formal summary of finding tables. As a result, the comparators were often not clearly stated, which may have influenced the reviewers' inferences about the discordant recommendations.

INTRODUCTION

To ensure that patients receive optimal care, consistent with their values and preferences, clinicians need trustworthy recommendations based on transparent ratings of certainty of evidence and strength of recommendations.[1] The widely adopted GRADE system (Grading of Recommendations Assessment, Development and Evaluation) offers a systematic and transparent framework to rate certainty (also referred to as quality or confidence) of evidence and to move from evidence to recommendations.[2-5]

Using GRADE, guideline-makers issue strong recommendations when they are confident that the desirable consequences clearly outweigh the undesirable consequences.[6 7] Conversely they should issue weak (also called conditional) when the balance of desirable and undesirable consequences between alternatives is close, the certainty in evidence is low, uncertainty or variability in patients' values and preferences is large, or cost-effectiveness is questionable.[6] Strong recommendations represent "just do it" recommendations applicable to almost all patients; weak recommendations are applicable to the majority of patients and include preference-sensitive decisions that require clinicians to ensure, through shared-decision making, that patients' choices are congruent with their values.[8]

GRADE views strong recommendations in the face of low certainty evidence (we will refer to such situations as *discordant recommendations*) as questionable, and often inappropriate. Some guidelines have a clear surfeit of discordant recommendations. For example, of 456 recommendations in 116 WHO guidelines, 160 (35%) proved discordant.[9 10] Similarly 121 of 357 (34%) recommendations in 17 Endocrine Society Guidelines proved discordant.[11 12]

Though discordant recommendations often represent a violation of GRADE guidance,

this is not always the case. GRADE has identified 5 seldom-occurring paradigmatic situations in which a strong recommendation is warranted despite low certainty in the evidence (Table 1).[6 13] Further, there is more than one explanation for an apparent violation of GRADE guidance (a discordant recommendation that fails to meet one of these criteria). First, the discordant recommendation may actually represent a good practice statement, in which indirect evidence justifies an inference that the recommended management option is far superior to the alternative.[14] Second, the panel may have misclassified the certainty of the evidence (it may actually be moderate or high). Third, and most concerning, the optimal management option is in fact value and preference-sensitive and the panel should have issued a weak recommendation (Table 2).[6 13]

Of the 160 discordant recommendations in the WHO guideline, 73 (46%) fell into the most concerning category of those that warranted a weak recommendation.[9 10] Of the 121 discordant recommendations in the Endocrine Society guidelines, 33 (27%) warranted a weak recommendation.[11] These results demonstrate that excessive use of strong recommendations in the face of low certainty evidence is common and concerning.

UpToDate (www.uptodate.com)[15] is an electronic medical textbook that uses GRADE and includes over 9,400 GRADE recommendations[15 16]. UpToDate has instituted intensive training in GRADE methods for their in-house deputy editors who are largely responsible for UpToDate material. Training involves regular large and small group seminars, and individual feedback from in-house methodologists.

Because it is enormously popular and used by clinicians worldwide, and despite the training in GRADE that their deputy editors undergo, the possibility that UpToDate is

issuing misleading strong recommendations on the basis of low certainty evidence constitutes a matter of concern. Therefore, we set out to determine, among all GRADE recommendations in UpToDate, the distribution of strong and weak recommendations, the proportion of discordant recommendations, and to characterize discordant recommendations based on the taxonomy described above (Table 1 & 2).

METHODS

Design and data source

We conducted an analytic survey of all GRADE recommendations included in UpToDate. We collaborated with UpToDate to identify all 9451 included in UpToDate as of June 2014, and determined their strength (strong or weak), and their certainty in evidence (high, moderate, or low – UpToDate does not use GRADE's "very low" category). We abstracted the title of each topic, as well as their corresponding clinical domains and age-group populations. From this database, we identified all discordant recommendations included in UpToDate as of January 2015.

Data abstraction on the discordant recommendations

UpToDate topics summarizing the evidence and rationale supporting the recommendations are mostly in narrative formats, and do not provide summary of finding tables or evidence profiles.[17] To assess the appropriateness of discordant recommendations according to the paradigmatic situation defined in the GRADE framework, we therefore standardized data abstraction to collect relevant information from the main text (detailed instruction Supplementary File 1).

Eight reviewers working in six pairs – all working actively as clinicians and proficient in GRADE methodology – performed data abstraction and assessed the appropriateness of discordant recommendations in duplicate. They abstracted the following information related to each discordant recommendation:

- Patient population (clinical field and age group);
- Type of intervention (drug, procedure, device, etc.) and type of comparator

(existing standard care, no intervention, alternative intervention, etc.);

- The clarity of the comparator, classified as (i) clearly and explicitly stated; (ii) not clearly and explicitly stated, but obvious; (iii) not clearly and explicitly stated or obvious, but relatively easy to infer; (iv) not at all clear - very uncertain;
- Outcomes: whether there was an explicit statement on mortality as well as the balance of benefits and harms;
- Whether there was an explicit statement on the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternative courses of action;
- Whether issues of cost or resources were explicitly discussed;
- The evidence supporting the recommendation, both for systematic reviews and primary study designs (randomized trials, observational studies, etc.)
- Whether the evidence summary suggested large effects in critical outcomes, or that indirect evidence, not incorporated in the grading, seemed to drive the recommendation.

Based on this abstracted information, each reviewer independently classified each of the discordant recommendations as either consistent with one of the five previously identified optimal categories for discordant recommendations (Table 1)[6 10 13] or in one of three categories in which we judged discordant recommendations to be inconsistent with GRADE guidance (Table 2): (i) good practice statements; (ii) a misclassification of the evidence – the evidence warranted moderate or high certainty rather than low; or (iii) uncertainty in the estimates of effect would best lead to a weak recommendation. We assessed agreement for whether recommendations were appropriate (vs. inappropriate) according to GRADE guidance using the chance-corrected kappa statistic.

The reviewers resolved all disagreements by discussion or through referral to an additional reviewer.

Data analysis and reporting

We abstracted data in an MS Excel database (v. 14.4) with pre-specified response categories whenever possible, and exported in SPSS (v. 22.0) for analysis. We analyzed the recommendation and sample characteristics as natural frequencies and proportions.

RESULTS

The 2971 topics in UpToDate that included GRADE recommendations covered a broad spectrum of clinical fields and health care, including 16.1% in oncology, 49.2% topics in other internal medicine specialties or primary care, and 12.5% in pediatrics. These topics included 9451 GRADE recommendations, of which 6501 (68.8%) were formulated as weak recommendations and 2950 (31.2%) as strong recommendations (Table 3). The proportion of strong recommendations varied greatly across clinical fields, ranging from 5.8% (in dermatology) to 42.7% (in cardiovascular medicine) (Supplementary File 2).

Of the 2950 strong recommendations, 844 (28.6%) were based on high certainty evidence, 1740 (59.0%) on moderate certainty, and 366 (12.4%) were discordant strong recommendations based on low certainty evidence (Table 3). Because UpToDate is continuously updated, 17 recommendations were modified in strength and/or certainty between the time all 9451 recommendations were retrieved, and the time all topics were downloaded for abstraction, as of January 2015.[15] The final study cohort therefore comprised a total of 349 discordant recommendations.

The 349 discordant recommendations were issued across 274 individual topics in UpToDate (each including a range of one to five recommendations), and the topics addressed covered a broad spectrum of health care issues within each clinical field, (Supplementary File 2). Interventions included drugs (56.4% of recommendations), surgery (19.8%), medical devices (6.9%), diagnostic or screening tests (20.9%), and other behavioral or multi-disciplinary interventions (10.0%). These interventions were most often compared to another intervention or to standard of care (56.7%) and less often to no intervention or placebo (36.1%).

The 349 discordant recommendations represent 3.7% of all 9451 recommendations. The proportion of discordant recommendations varied from 0% (e.g. in palliative care, dermatology or for recommendations applying specifically to the elderly population), to 7.0% in pediatrics, 8.0% in infectious disease, and 10.9% in hematology (Supplementary File 2).

Evidence supporting the discordant recommendations

The comparator was clearly and explicitly stated in 73 (20.9%) of the 349 recommendations, not clearly but either obvious or relatively easy to infer in 230 (65.9%) and very uncertain in 46 (13.2%). The direction of the recommendation was most often framed in favor of the intervention (78.5%) rather than against it (Table 4).

The full-text of the UpToDate topic often provided a rationale supporting the recommendation. An explicit statement on the balance of benefits and harms was present in 92 (26.4%), and an implicit statement in 157 (45.0%), and no statement in 100 (28.7%). Explicit statements addressing the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternatives were present in 10 (2.9%) of the recommendations; they could be inferred in 171 (49.0%), but not in the remaining 168 (48.1%) of discordant recommendations. Cost or resources considerations were mentioned in 15 (4.3%). The evidence cited to support each discordant recommendation varied substantially, with a median of 4 references cited, range from 0 to 33, with 45 (12.9%) of recommendations without any citation. Observational studies dominated (203, 58.2%); 49 (14.0%) were supported by a systematic review (Table 4).

Appropriateness of the discordant recommendations

Kappa for the initial taxonomic judgment regarding whether the recommendation was appropriate or inappropriate according to GRADE guidance was 0.46 (moderate agreement). The two reviewers required consensus discussions for 43% of the discordant recommendations. Third party adjudication to determine the appropriate classification was required in 12 of the discordant recommendations (3.4%).

Reviewers judged 204 (58.5%) of the 349 discordant recommendations to be consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations (Table 5). The most common paradigm was a "life-threatening or potentially catastrophical situation", followed by "potential similar benefits, one clearly less risky or costly", "potential catastrophic harm", "uncertain benefits, certain harm", and "established similar benefits, one potentially more risky or costly" (Table 5).

Reviewers judged 47 (13.5%) of the 349 discordant recommendations as "good practice statements"; 38 (10.9%) as a "misclassification of certainty (evidence warranted moderate or high certainty)"; and 60 (17.2%) as warranting a weak recommendation (see Table 5).

DISCUSSION

Among 9451 GRADE recommendations in UpToDate, about two thirds were formulated as weak recommendations and the remainder as strong recommendations. Of all recommendations, only 3.7% (n=349) were strong recommendations based on low certainty in effect estimates (Table 3). Of these discordant recommendations, over half were consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations; approximately 14% represented "good practice statements"; approximately 11% were based on a misclassification of certainty (evidence warranted moderate or high certainty), and approximately 17% were judged to warrant a weak recommendation (Table 5). The proportion of appropriate discordant recommendations varied across intervention types or clinical fields (Supplementary File 2). Although most topics in UpToDate provided a rationale to support the discordant recommendation, 29% lacked statements about benefits and harms and 13% did not provide citations, which points at potential areas of improvement for UpToDate related to standards for trustworthy guidelines.[1]

Strengths and limitations

This study assessed the strength of recommendations in the largest known sample of recommendations developed using GRADE. Indeed, even large guidelines include a few hundred recommendations[18], whereas UpToDate topics have one of the largest known coverage in clinical fields and included 9451 recommendations at the time of this assessment.

The taxonomy that we used has been successfully implemented in two prior studies of clinical guidelines[10 11] (see below: relation to prior work). Our reviewers were clinical

epidemiologists with an in-depth understanding of GRADE methodology and were therefore well equipped to assess judgments on evidence and recommendations. Despite these advanced skills, chance corrected kappa agreement on the appropriateness of recommendations was moderate, albeit satisfactory (0.48). Consensus discussions were needed for 43% of discordant recommendations, although formal adjudication by third parties was required for only 12 discordant recommendations (3.4%).

The necessity for frequent consensus discussions reflects the substantial judgment required in categorizing recommendations. This is in part due to the narrative nature of UpToDate topics, which does not include formal summary of finding tables or evidence profiles[17], often discussing the evidence and rationale for several recommendations in a free-text cross-referenced structure that sometimes omits statements regarding benefits and harms, and lacks citations. A limitation of our study is that decisions were based solely on information published in UpToDate, while authors of the topics my have considered other factors.[19]

Another element contributing to the challenges in making categorizations is the clarity of the comparison on which the recommendation applies. As in previous assessment in guidelines[9], the comparator was clearly and explicitly stated in only 73 (20.9%) of discordant recommendations and was very uncertain in 46 (13.2%). When comparators were not clear and explicit, reviewers' inferences may not always have been correct.[19]

Relation to previous work

Two prior studies provided a formal structured exploration of discordant recommendations using the GRADE approach. An assessment of 357 recommendations

in 17 Endocrine Society Guidelines found that 58% were strong, of which 59% were based on low certainty.[11 12] Only 29% of discordant recommendations were consistent with one of the 5 paradigmatic situations, whereas 36% were good practice statements, 4% were recommendations for additional research; 4% involved misclassification of the certainty; and 27% had no compelling explanation and should have been weak recommendations.[11]

A second study of 456 recommendations in 116 WHO guidelines using GRADE found that 63% were strong, of which 56% were based on low (33%) or even very low (23%) certainty.[10] Of the 160 discordant recommendations, only 15.6% were judged consistent with GRADE guidance, while 18% were good practice statements, 21% involved misclassification of the certainty, and 46% should have been weak rather than strong recommendations.[9]

Our results contrast with these previous two studies. First, the proportion of weak recommendations (more than two thirds of the total) was approximately 30% higher in UpToDate than in WHO and Endocrine Society guidelines. This proportion was however similar to the 9th edition ACCP guideline on Antithrombotic Therapy and Prevention of Thrombosis, after it implemented GRADE.[18 20] Second, the proportion of inappropriate, discordant recommendation was considerably lower. In particular, of the discordant recommendations, the proportion that should have been weak was about 17%, rather than 27% (Endocrine Society)[11] or 46% (WHO guidelines).[9]

A subsequent interview of panel members involved in the WHO guidelines highlighted several reasons contributing to discordant recommendations. These included political considerations around long-established practices, the need for funding and policy formulation, or the fear of pushback from media.[19] Certain panel members also

expressed some skepticism about the value of making weak recommendations, or concerns they may be ignored[19], although another study reported that WHO weak recommendations are frequently adopted in national policies (uptake of 61% for weak recommendations versus 82% for strong recommendations).[21] Finally, both financial and intellectual conflicts of interest among panel members may be an explanation for discordant recommendations; for instance, through the dominance by some panelists over inexperienced panel leader.[19 22] Of these explanations, the last – conflict of interest – is in our view the most likely to apply in the UpToDate context.

Implications and conclusion

For users of UpToDate, our results are generally, though not absolutely, reassuring. The proportion of discordant recommendations is very small – only 3.7% of all recommendations. Furthermore, of the three categories inconsistent with GRADE guidance – good practice statement, misclassification of the certainty, and evidence warranting a weak recommendation (Table 2) – the third is by far the most problematic.[9] Good practice statements are appropriate when indirect evidence that is difficult to collect and summarize warrants high certainty in the impact of a given intervention and when the balance benefits and harms is large.[14] Thus, in terms of implications for clinical practice, good practice statements have the same force as strong recommendation. Similarly with misclassification of certainty: since the certainty is actually moderate or high, a strong recommendation is appropriate. Recommendations that should have been weak instead of strong provide inappropriate "just do it" guidance for clinical practice, although they are actually preference-sensitive and should thus warrant shared-decision making.[8] Of the 349 discordant recommendations in UpToDate, only 60 fall in the category of inappropriate strong recommendations.

Thus, clinicians using UpToDate can anticipate that they will be misleadingly instructed to take a "just do it" rather than an "it depends" approach to clinical decision making in 0.6% (6 of 1,000) UpToDate recommendations.[15] This seems close to a threshold in which one might ignore the problem. Nevertheless, we would still encourage clinicians to be alert to the possibility of an inappropriate strong recommendation – in UpToDate or elsewhere – whenever the recommendation is based on low certainty evidence and authors fail to provide an explicit rationale corresponding to one of the categories in Table 1.

A likely explanation for UpToDate's success in avoiding inappropriate discordant recommendations is the training and feedback that their deputy editors receive. For organizations using GRADE, our results suggest the desirability of such training for those involved in formulating recommendations to optimize use of GRADE.

Finally our results highlight the need for authors of trustworthy recommendations or guidelines[1] to provide clear and explicit comparators, as well as transparent and systematic reports of the key ingredients of their rationale when moving from evidence to recommendation.[17 23 24] Future avenues for research should also look at optimal presentation formats of EBM textbooks and guidelines, to ensure clinicians actually understand both the rationale and potential implications of all recommendations for clinical practice.[8 25-28]

LIST OF ABBREVIATIONS

GRADE: Grading of Recommendations Assessment, Development and Evaluation

WHO: World Health Organization

ACCP: American College of Chest Physicians

COMPETING INTERESTS

TA, AK, IN, RBP, PEA, DR, POV, and GHG are active members of the GRADE working group.

DR, at the time the data on graded recommendations was extracted from UpToDate and until 2016, was an employee of UpToDate – he reports personal fees from UpToDate, outside the submitted work.

GHG contributes to the training in GRADE methods for UpToDate in-house deputy editors, for which he reports personal fees from UpToDate, outside the submitted work.

CONTRIBUTORS

Thomas Agoritsas (TA) and Gordon H. Guyatt (GHG) designed the study. David Rind (DR) provided the list of all recommendations and grading from UpToDate. Paul E. Alexander (PEA) helped structuring data abstraction. Thomas Agoritsas (TA), Arnaud Merglen (AM), Anja F. Heen (AFH), Annette Kristiansen (AK), Ignacio Neumann (IN), Juan P. Brito (JPB), Romina Brignardello-Petersen (RBP), and Per O. Vandvik (POV) reviewed the recommendations in duplicate and classified them according to GRADE taxonomy. Thomas Agoritsas (TA) and Gordon Guyatt (GG) wrote the first draft of the manuscript. All authors have read the manuscripts and made improvements of the content and wording.

FUNDING / ACKNOWLEDGMENTS

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

DATA SHARING STATEMENT

There were no additional unpublished data from this study.

REFERENCES

1. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. . In: Graham R, Mancher M, Miller Wolman D, et al., eds. Clinical Practice Guidelines We Can Trust. Washington (DC), 2011.
2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;**336**(7650):924-6.
3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;**64**(4):383-94.
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. Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013.
7. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;**64**(4):401-6.
8. Agoritsas T, Heen AF, Brandt L, et al. Decision aids that really promote shared decision making: the pace quickens. BMJ 2015;**350**:g7624.
9. Alexander PE, Brito JP, Neumann I, et al. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. J Clin Epidemiol 2016;**72**:98-106.
10. Alexander PE, Bero L, Montori VM, et al. World Health Organization recommendations are often strong based on low confidence in effect estimates. J Clin Epidemiol 2014;**67**(6):629-34.
11. Brito JP, Domecq JP, Murad MH, et al. The Endocrine Society guidelines: when the confidence cart goes before the evidence horse. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3246-52.
12. Vigersky RA, Bhasin S, Martin KA. The Endocrine Society Clinical Practice Guidelines: a self-assessment. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3174-7.
13. Neumann I, Santesso N, Akl EA, et al. A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach. J Clin Epidemiol 2016;**72**:45-55.
14. Guyatt GH, Schunemann HJ, Djulbegovic B, et al. Guideline panels should not GRADE good practice statements. J Clin Epidemiol 2014.
15. UpToDate, Waltham, MA. http://www.uptodate.com (Accessed on July 7, 2017).
16. Agoritsas T, T Vandvik PO, Neumann I, Rochwerg B, Jaeschke R, Hayward R, Guyatt GH, McKibbon A. Chapter 5. Finding Current Best Evidence, in JAMA Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3nd Edition, McGraw-Hill Medical, 2015.

17. Guyatt G, Oxman AD, Akl E, et al. GRADE guidelines 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2010.

18. Agoritsas T, Neumann I, Mendoza C, et al. Guideline conflict of interest management and methodology heavily impacts on the strength of recommendations: comparison between two iterations of the American College of Chest Physicians Antithrombotic Guidelines. J Clin Epidemiol 2017;**81**:141-43.

19. Alexander PE, Gionfriddo MR, Li SA, et al. A number of factors explain why WHO guideline developers make strong recommendations inconsistent with GRADE guidance. J Clin Epidemiol 2016;**70**:111-22.

20. Guyatt GH, Norris SL, Schulman S, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;**141**(2 Suppl):53S-70S.

21. Nasser SM, Cooke G, Kranzer K, et al. Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. J Clin Epidemiol 2015;**68**(6):703-7.

22. Alexander PE, Li SA, Gionfriddo MR, et al. Senior GRADE methodologists encounter challenges as part of WHO guideline development panels: an inductive content analysis. J Clin Epidemiol 2016;**70**:123-8.

23. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;**64**(4):395-400.

24. Guyatt GH, Oxman AD, Kunz R, et al. Going from evidence to recommendations. BMJ 2008;**336**(7652):1049-51.

25. Vandvik PO, Brandt L, Alonso-Coello P, et al. Creating clinical practice guidelines we can trust, use, and share: a new era is imminent. Chest 2013;**144**(2):381-9.

26. Kristiansen A, Brandt L, Alonso-Coello P, et al. Development of a novel multilayered presentation format for clinical practice guidelines. Chest 2014.

27. Treweek S, Oxman AD, Alderson P, et al. Developing and evaluating communication strategies to support informed decisions and practice based on evidence (DECIDE): protocol and preliminary results. Implement Sci 2013;**8**:6.

28. Siemieniuk RA, Agoritsas T, Macdonald H, et al. Introduction to BMJ Rapid Recommendations. BMJ 2016;**354**:i5191.

Table 1. Paradigmatic situations in which a strong recommendation may be warranted despite low or very low certainty in effect estimates *(appropriate strength, consistent with GRADE)*

Situation	Certainty in Estimates (Quality of Evidence)		Balance of Benefits and Harms	Values and Preferences	Resource Considerations	Recommendation	Example
	Benefits	Harms					
1. Life-threatening (or catastrophical) situation	Low or very low	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation; adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits justify the intervention	Strong recommendation in favor of the intervention	Indirect evidence from seasonal influenza suggests that patients with avian influenza may benefit from the use of oseltamivir (low certainty in effect estimates). Given the high mortality of the disease and the absence of effective alternatives, the WHO made a strong recommendation in favor of the use of oseltamivir rather than no treatment in patients with avian influenza.
2. Uncertain benefit, certain harm	Low or very low	High or moderate	Possible but uncertain benefit; substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	High incremental cost (or resource use) relative to the benefits may not justify the intervention	Strong recommendation against the intervention	In patients with idiopathic pulmonary fibrosis, treatment with azathioprine plus prednisone offers a possible but uncertain benefit in comparison with no treatment. The intervention, however, is associated with a substantial established harm. An international guideline made a recommendation against the combination of corticosteroids plus azathioprine in patients with idiopathic pulmonary fibrosis.
3. Potential equivalence, one option clearly less risky or costly	Low or very low	High or moderate	Magnitude of benefit apparently similar—though uncertain—for alternatives; we are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation for less harmful/less expensive	Low-quality evidence suggests that initial Helicobacter pylori eradication in patients with early stage extranodal marginal zone (MALT) B-cell lymphoma results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy, but with high certainty of less harm, morbidity, and cost. Consequently, UpToDate made a strong recommendation in favor of H pylori eradication rather than radiotherapy in patients with MALT lymphoma.
4. High certainty in similar benefits, one option potentially more risky or costly	High or moderate	Low or very low	Established that magnitude of benefit is similar for alternative management strategies; best (though uncertain) estimate is that one alternative has appreciably greater harm	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation against the intervention with possible greater harm	In women requiring anticoagulation and planning conception or in pregnancy, high certainty estimates suggest similar effects of different anticoagulants. However, indirect evidence (low certainty in effect estimates) suggests potential harm to the unborn infant with oral direct thrombin (eg, dabigatran) and factor Xa inhibitors (eg, rivaroxaban, apixaban). The AT9 guidelines recommended against the use of such anticoagulants in women planning conception or in pregnancy.
5. Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) relative to the benefits, may not justify the intervention	Strong recommendation against the intervention	In males with androgen deficiency, testosterone supplementation likely improves quality of life. Low- certainty evidence suggests that testosterone increases cancer spread in patients with prostate cancer. The US Endocrine Society made a recommendation against testosterone supplementation in patients with prostate cancer.

Reproduced and adapted from Neumann & al., [13].

Table 2. Inappropriate reasons for issuing strong recommendation based on low or certainty in effect estimates *(inconsistent with GRADE)*

Situation	Example
Best practice recommendation (for which sensible alternatives do not exist)	"For patients with congenital adrenal hyperplasia, we recommend monitoring patients for signs of glucocorticoid excess, as well as for signs of inadequate androgen suppression." This statement should not have been GRADEed as sensible alternatives do not exist
Misclassification of certainty was warranted (typically because indirect evidence of moderate certainty, driving the strong recommendation, was not acknowledged)	"We recommend intensive lifestyle modification to the entire family and to the patient, and as the prerequisite for all overweight and obesity treatments for children and adolescents." GRADE as low quality when there is moderate quality evidence for benefits
Lack of compelling explanation (the recommendation should have been weak)	"If a patient is unable or unwilling to undergo surgery, we recommend medical treatment with mineralocorticoids" Lack of evidence of mineralocorticoids being superior to other medical treatment (eg, anti-hypertensive medications)

Elements adapted from Brito et al.[11]

Table 3. Distribution of the strength of the recommendations in UpToDate according to the certainty in evidence

	Weak Recmendations	Strong Recommendations	All Recommendations
	N (%)	N (%)	N (%)
Low certainty	4335 (66.7%)	366 (12.4%)	4701 (49.7%)
Moderate certainty	2019 (31.1%)	1740 (59.0%)	3759 (39.8%)
High certainty	147 (2.3%)	844 (28.6%)	991 (10.5%)
Total	**6501** (68.8% of all rec)	**2950** (31.2% of all rec)	**9451** (100%)

Table 4. Characteristics of all 349 discordant recommendations in UpToDate, and proportion of appropriate discordant recommendations

	N (%)	% of appropriate discordant *(p-value)*
Clinical Specialtes		*(p = 0.160)*
Primary Care and General Internal Medicine	15 (4.3)	53.3
Emergency Medicine	16 (4.6)	81.3
Critical Care	5 (1.4)	80.0
Internal Medicine specialties	158 (45.3)	57.6
Oncology (including hemato-oncology)	43 (12.3)	55.8
Pediatrics	73 (20.9)	47.9
Obstetrics, Gynecology and Women Health	19 (5.4)	73.7
General Surgery	13 (3.7)	69.2
Anesthesiology	3 (0.9)	100.0
Psychiatry	4 (1.1)	75.0
Intervention type		*(p = 0.010)*
Drug intervention	197 (56.4)	61.4
Surgical interventions	69 (19.8)	59.4
Medical device	24 (6.9)	62.5
Behavioural or multi-disciplinary intervention	35 (10.0)	57.1
Diagnostic test, screening programms	24 (6.9)	29.2
Clarity of the comparator		*(p <0.001)*
Comparator not at all clear – very uncertain	46 (13.2)	37.0
Comparator not clearly and explicitly stated or obvious, but relatively easy to infer	120 (34.4)	48.3
Comparator not clearly and explicitly stated, but obvious	110 (31.5)	68.2
Comparator clearly and explicitly stated	73 (20.9)	74.0
Type of comparator		*(p = 0.083)*
Too unclear	25 (7.2)	44.0
No intervention (or placebo)	126 (36.1)	54.0
Other intervention(s) (standard of care or alternative(s))	198 (56.7)	63.1
Direction of the recommendation		*(p <0.001)*
For the intervention (i.e. against the comparator)	274 (78.5)	51.1
Against the intervention (i.e. for the comparator)	75 (21.5)	85.3
Mortality		*(p <0.001)*
No statement about mortality	189 (54.2)	47.1
Implicit statement about mortality	47 (13.5)	68.1
Explicit statement about mortality	113 (32.4)	73.5
Balance of benefits and harms		*(p <0.001)*
No statement about the balance of outcomes	100 (28.7)	28.0
Implicit statement about the balance of outcomes	157 (45.0)	66.9
Explicit statement about the balance of outcomes	92 (26.4)	77.2

Relative Importance of outcomes - Values & Preferences			*(p <0.001)*
No statement about the relative importance of outcomes	168	(48.1)	42.9
Implicit statement about the relative importance of outcomes	171	(49.0)	73.1
Explicit statement about the relative importance of outcomes	10	(2.9)	70.0
Cost of resources			*(p = 0.023)*
No statement about cost or resources	334	(95.7)	57.2
Cost or resources clearly and explicitly stated	15	(4.3)	86.7
Supporting systematic review (SR)			*(p = 0.175)*
No SR is cited	300	(86.0)	56.3
SR of Observational studies	22	(6.3)	63.6
SR of both RCT and Observational studies	13	(3.7)	76.9
SR of Randomized Trials (RCT)	14	(4.0)	78.6
Design of primary studies			*(p = 0.002)*
No reference cited	45	(12.9)	35.6
Other type (eg narrative review, book chapter)	48	(13.8)	54.2
Observational studies	203	(58.2)	61.1
Randomized Trials (RCT)	53	(15.2)	71.7
Total	**349**	**(100)**	**58.5**

Table 5. Summary judgments on the appropriateness of 349 discordant strong recommendation based on low certainty in effect in UpToDate

	N	(%)
Appropriate discordant recommendations (consistent with GRADE)		
1. Life-threatening (or catastrophical) situation	70	(20.1)
2. Uncertain Benefit, Certain Harm	28	(8.0)
3. Potential similar benefits, One clearly less risky (or costly)	56	(16.0)
4. Established similar benefits, One potentially more risky (or costly)	18	(5.2)
5. Potential catastrophic harm	32	(9.2)
Total	**204**	**(58.5)**
Inappropriate discordant recommendations (inconsistent with GRADE)		
6. Good Practice Statement	47	(13.5)
7. Misclassification of certainty (judged moderate or high)	38	(10.9)
8. Lack of explanation, should have been weak recommendation (GRADE 2C)	60	(17.2)
Total	**145**	**(41.5)**

Supplementary File 1. INSTRUCTION FOR ABSTRACTION

Please read carefully before starting abstraction.
For any questions please contact me ASAP (thomas.agoritsas@gmail.com)

Background

- In a first phase of the project, we characterized the strength and confidence of the 9451 GRADE recommendations in UTD.
- In this last phase, we are focusing on
  - The **349 strong recommendations based on low confidence (GRADE 1C)**
  - Which are included in a total of **274 topics** (=chapters in UpToDate).
- The main objective is to categorize them based according the following taxonomy
  - Appropriate grading: recommendation consistent with one of the five paradigmatic situations defined the GRADE framework (see examples in Table 1 below):
    - *[App#1] Life-threatening situation*
    - *[App#2] Uncertain Benefit, Certain Harm*
    - *[App#3] Potential similar benefits, one clearly less risky (or costly)*
    - *[App#4] Established similar benefits, one potentially more risky (or costly)*
    - *[App#5] Potential catastrophic harm of one option*
  - Inappropriate grading: recommendation inconsistent with GRADE (see examples in Table 2 below), in short:
    - *[Inapp#1] Good Practice Statement*
    - *[Inapp#2] Misclassification of confidence (should have been GRADE 1B or 1A)*
    - *[Inapp#3] Lack of explanation, should have been a weak rec (GRADE 2C)*

→ Before starting, please read the examples and Appendix Tables 1 & 2, they are also embedded in separate tabs in the abstraction excel file. Do not focus on memorizing them, as data abstraction will guide you in your judgment.

Abstraction Excel File & Variables

- We will conduct the whole abstraction process in the attached standardized excel file
- We have kept the variables to abstract to the minimum necessary, most with **pre-defined response categories in drop-down menus** (and infrequently as free-text for copy-pasting).
- **Each recommendation has a separate row in the file**. There are sometimes more than GRADE 1C per UTD topic (the number are indicated).
- **As a guiding principle, keep in mind that the main objective is to assess the most appropriate taxonomy (first as appropriate or inappropriate, then subcategory).**
  These are the last variables in the file.
- This requires judgment based on what is reported in the UpToDate topic, mostly in narrative form. Indeed, there are typically no "summary of findings tables" of "evidence profiles" to explicit GRADE assessment. Absolute certainty in taxonomy is sometimes hard to achieve, but try and assign the best fit you can.
- The abstraction form is organized as follows to guide your final judgment:

UpToDate TOPIC & RECOMMENDATION	Pre-entered data to help you identify the relevant topic in the UpToDate topic in the dropbox
POPULATION	Check pre-entered clinical field, document age-group
INTERVENTION	Intervention type
COMPARATOR	Clarity and type of the comparator, direction of rec
OUTCOME – Benefits & Harms	Statement about: mortality, balance of benefits & harms, their relative importance (ie. values/preferences), cost
EVIDENCE supporting the rec	Date of literature review & updates number and type of supporting evidence Indication regarding the potential role of indirect evidence, Presence of large effects.
Conflict of interest (COI)	Copy paste statement, presence of financial COI
TAXONOMY – APPROPRIATE	**Separate judgement on each of the 5 paradigmatic situation defined by GRADE** (clear, possible, no)
TAXONOMY – INAPPROPRIATE	**Separate judgmenet on each of the 3 inappropriate situations**
CONFLICT RESOLUTION	**TAXONOMY DECISION** (for kappa), confidence in the decision (to document), assessing agreement and **RECONCILED TAXONOMY** within each pair of abstractor.
ADJUDICATION	Recording adjudication third reviewer if this was necessary

- **Specific guidance for each variable is found in the GREY BOXES on top of each column.**
- **Please read and select best option from DROP-DOWN menus within each cell**
- A few cells are for free-text to copy paste from the topic. Be sure to double click in the cell before pasting content, to keep the format intact.
- Most variables are followed by a column labelled **"additional comments"** or "rationale". These are for your personal notes to guide conflict resolution.
- A few examples already abstracted are shown in the first rows as an indication.

Abstraction: STEP-BY-STEP

Start abstracting a few first recommendations to get familiar with the process and contact me (thomas.agoritsas@gmail.com) for any question. I'm happy to have a quick skype if and as often as necessary.

- Go to the recommendation in the next row in the excel file.
- The [topic_#] and [topic_title] correspond to the name of the **PDF file** for the corresponding UTD topic → Open it.
- **Search automatically (ctrl-R or command-F) for "1C" → This will directly lead you to the GRADE 1C recommendation(s) at the end of the topic under the "Summary and Recommendations" section.**
- Read it carefully, and take a few seconds to try and get some first rough impression re: potential taxonomy
- **Then, find the corresponding paragraph in the topic that supports the recommendation** (it is often indicated soon after the recommendation (e.g. "See treatment…"). If not, try and find which paragraph(s) discuss(es) the recommendation.
- Abstract all variables in the order of the file as this will guide your formal judgment.
- Then judge each of the 5 appropriate and 3 inappropriate categories in the taxonomy.
- Then decide which one fits best and document the confidence you have in your assessment.
- Go to the next recommendation in the next row. (if this is in the same topic, you'll be able to copy several or your answers.

Conflict resolution and adjudication

- **You've been assigned with a paired reviewer. Schedule a first conflict resolution in the following days to ensure you are on the same page.**
- Record the agreed taxonomy in the specific column ("RECONCILED TAXONOMY"). Do NOT modify your initial judgement ("TAXONOMY DECISION") - as this will use to calculate kappa.
- **If you cannot resolve conflict, send us your questions for adjudication by a third reviewer.**
- Record adjudication in the final column.

➢ **Thanks again for your help. Do not hesitate to contact me for any questions (thomas.agoritsas@gmail.com)**

Supplementary File 2.

Characteristics of all 9451 recommendations in UpToDate: certainty effect estimates and clinical fields

	N	(%)	% of Strong Rec	% of strong rec discordant	% of any rec being discordant
Clinical Fields					
1 Primary Care & General Internal Medicine	356	(3.8)	22.5	16.3	3.7
2 Emergency Medicine	295	(3.1)	25.1	**23.0**	**5.8**
3 Critical Care	144	(1.5)	34.0	10.2	3.5
4 Cardiovascular Medicine	529	(5.6)	**42.7**	5.3	2.3
5 Infectious Diseases	870	(9.2)	**41.3**	**19.5**	**8.0**
6 Nephrology and Hypertension	475	(5.0)	**39.8**	6.3	2.5
7 Pulmonary Medicine	347	(3.7)	34.6	6.7	2.3
8 Hematology (non-oncology)	192	(2.0)	**41.1**	**26.6**	**10.9**
9 Neurology	395	(4.2)	31.1	**17.9**	**5.6**
10 Allergy and Immunology	261	(2.8)	23.4	3.3	0.8
11 Endocrinology & Diabetes	504	(5.3)	19.0	6.3	1.2
12 Gastroenterology & Hepatology	471	(5.0)	22.9	7.4	1.7
13 Rheumatology	216	(2.3)	21.3	4.3	0.9
14 Palliative Care	33	(0.3)	18.2	0.0	0.0
15 Oncology	1255	(13.3)	**36.0**	7.7	2.8
16 Hemato-oncology	263	(2.8)	27.8	**20.5**	**5.7**
17 Pediatrics	1057	(11.2)	**39.5**	**17.7**	**7.0**
18 Pediatric Emergency Medicine	126	(1.3)	28.6	5.6	1.6
19 Gynecology & Obstetrics	709	(7.5)	22.3	12.0	2.7
20 General Surgery	403	(4.3)	**32.8**	12.1	4.0
21 Anesthesiology	48	(0.5)	16.7	**37.5**	**6.3**
22 Dermatology	240	(2.5)	5.8	0.0	0.0
23 Psychiatry	262	(2.8)	16.8	9.1	1.5
TOTAL	**9451**	**(100)**	**31.2**	**12.4**	**3.9**

BMJ Open

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018593.R1
Article Type:	Research
Date Submitted by the Author:	11-Sep-2017
Complete List of Authors:	Agoritsas, Thomas; University Hospitals of Geneva, Division of General Internal Medicine & Division of Clinical Epidemiology; McMaster University Faculty of Health Sciences, Department of Health Research Methods, Evidence, and Impact Merglen, Arnaud; University Hospitals of Geneva, Division of General Pediatrics Heen, Anja; Innlandet Hospital Trust-division Gjøvik, Department of Internal Medicine Kristiansen, Annette; Inland hospital trust, Internal medicine, Gjøvik Neumann, Ignacio; Pontificia Universidad Catolica de Chile, Department of Internal Medicine; McMaster University, Department of Health Research Methods, Evidence, and Impact, Brito, Juan; Mayo Clinic Minnesota, Department of Medicine and Knowledge and Evaluation Research Unit, Brignardello-Petersen, Romina; McMaster University, Department of Health Research Methods, Evidence, and Impact Alexander, Paul; McMaster University, Department of Health Research Methods, Evidence, and Impact Rind, David; Institute for Clinical and Economic Review Vandvik, Per; Norwegian Knowledge Centre for the Health Services, Guyatt, Gordon; Mcmaster University, Department of Health Research Methods, Evidence, and Impact
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Epidemiology
Keywords:	Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Thomas Agoritsas, MD, PhD [1,2]

Arnaud Merglen, MD, MSc [3]

Anja Fog Heen, MD [4]

Annette Kristiansen, MD, PhD [4]

Ignacio Neumann, MD, PhD [2,5]

Juan P Brito, MD, MSc [6]

Romina Brignardello-Petersen, DDS, MSc, PhD [2, 7]

Paul E Alexander, MSc, PhD [2]

David M Rind, MD, MSc [8]

Per O. Vandvik, MD, PhD [4,9]

Gordon H Guyatt, MD, MSc [2]

Affiliations:

[1] Division of General Internal Medicine & Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland.

[2] Department of Health Research Methods, Evidence, and Impact, McMaster University, Faculty of Health Sciences, Hamilton, Ontario, Canada

[3] Division of General Pediatrics, University Hospitals of Geneva & Faculty of Medicine, University of Geneva, Geneva, Switzerland.

[4] Department of Internal Medicine, Innlandet Hospital Trust-division Gjøvik, Norway

[5] Department of Internal Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

[6] Division of Endocrinology, Diabetes, Metabolism and Nutrition, Department of Medicine and Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, USA

[7] Faculty of Dentistry, University of Chile, Chile.

[8] Institute for Clinical and Economic Review, Boston, MA, USA

[9] Institute of Health and Society, Faculty of Medicine, University of Oslo, Norway

*** Correspondence to**: Thomas Agoritsas, MD, PhD (thomas.agoritsas@gmail.com)

Division of General Internal Medicine & Division of Clinical Epidemiology

Department Internal Medicine, Rehabilitation and Geriatrics (DMIRG)

University Hospitals of Geneva, Rue Gabrielle-Perret-Gentil 4,

1211 Genève 14, Switzerland

Phone: +41 79 55 34 543

Keywords: Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine.

Abstract: 283 words / **Manuscript**: 3064 words

Tables: 5 / Supplementary Files: 2

ABSTRACT

Introduction

UpToDate is widely used by clinicians worldwide and includes more than 9,400 recommendations that apply the GRADE framework. GRADE guidance warns against strong recommendations when certainty of the evidence is low or very low (discordant recommendations), but has identified five paradigmatic situations in which discordant recommendations may be justified.

Objectives

Our objective was to document the strength of recommendations in UpToDate and assess the frequency and appropriateness of discordant recommendations.

Design

Analytic survey of all recommendations in UpToDate

Methods

We identified all GRADE recommendations in UpToDate, and examined their strength (strong or weak) and certainty of the evidence (high, moderate, or low certainty). We identified all discordant recommendations as of January 2015, and pairs of reviewers independently classified them either into one of the five appropriate paradigms or into one of three categories inconsistent with GRADE guidance.

Results

UpToDate included 9451 GRADE recommendations, of which 6501 (68.8%) were formulated as weak recommendations and 2950 (31.2%) as strong. Among the strong,

844 (28.6%) were based on high certainty in effect estimates, 1,740 (59.0%) on moderate certainty, and 366 (12.4%) on low certainty. Of the 349 discordant recommendations 204 (58.5%) were judged appropriate (consistent with one of the five paradigms); we classified 47 (13.5%) as good practice statements; 38 (10.9%) misclassified the evidence as low certainty when it was at least moderate; and 60 (17.2%) warranted a weak rather than a strong recommendation.

Conclusion

The proportion of discordant recommendations in UpToDate is small, and the proportion that is truly problematic (strong recommendations that would best have been weak) very small. Clinicians should nevertheless be cautious, and look for clear explanations – in UpToDate and elsewhere – when guidelines offer strong recommendations based on low certainty evidence.

Strengths and limitations of this study

- We assessed the strength of recommendations in the largest known sample of recommendations using GRADE (N=9451) addressing a wide array of clinical fields.
- We used a taxonomy to appraise discordant recommendations that has been successfully implemented in two prior assessments of clinical practice guidelines.
- We based our assessment solely on information published in UpToDate, while authors of the topics may have considered other factors in deciding to issue a discordant recommendation.
- UpToDate topics are narrative in nature and do not include formal summary of finding tables. As a result, the comparators were often not clearly stated, which may have influenced the reviewers' inferences about the discordant recommendations.

INTRODUCTION

To ensure that patients receive optimal care, consistent with their values and preferences, clinicians need trustworthy recommendations based on transparent ratings of certainty of evidence and strength of recommendations.[1] The widely adopted GRADE system (Grading of Recommendations Assessment, Development and Evaluation) offers a systematic and transparent framework to rate certainty (also referred to as quality or confidence) of evidence and to move from evidence to recommendations.[2-5]

Using GRADE, guideline-makers issue strong recommendations when they are confident that the desirable consequences clearly outweigh the undesirable consequences.[6 7] Conversely they should issue weak (also called conditional) when the balance of desirable and undesirable consequences between alternatives is close, the certainty in evidence is low, uncertainty or variability in patients' values and preferences is large, or cost-effectiveness is questionable.[6] Strong recommendations represent "just do it" recommendations applicable to almost all patients; weak recommendations are applicable to the majority of patients and include preference-sensitive decisions that require clinicians to ensure, through shared-decision making, that patients' choices are congruent with their values.[8]

GRADE views strong recommendations in the face of low certainty evidence (we will refer to such situations as *discordant recommendations*) as questionable, and often inappropriate. Some guidelines have a clear surfeit of discordant recommendations. For example, of 456 recommendations in 116 WHO guidelines, 160 (35%) proved discordant.[9 10] Similarly 121 of 357 (34%) recommendations in 17 Endocrine Society Guidelines proved discordant.[11 12]

Though discordant recommendations often represent a violation of GRADE guidance,

this is not always the case. GRADE has identified 5 seldom-occurring paradigmatic situations in which a strong recommendation is warranted despite low certainty in the evidence (Table 1).[6 13] Further, there is more than one explanation for an apparent violation of GRADE guidance (a discordant recommendation that fails to meet one of these criteria). First, the discordant recommendation may actually represent a good practice statement, in which indirect evidence justifies an inference that the recommended management option is far superior to the alternative.[14] Indirect evidence refers to evidence that does not directly address the question at hand, but nevertheless bears on the question. For instance, though there are no randomized trials of use of a parachute after jumping out of plane, there is ample indirect evidence suggesting its impact on mortality from the jump. Second, the panel may have misclassified the certainty of the evidence (it may actually be moderate or high). Third, and most concerning, the optimal management option is in fact value and preference-sensitive and the panel should have issued a weak recommendation (Table 2).[6 13]

Of the 160 discordant recommendations in the WHO guideline, 73 (46%) fell into the most concerning category of those that warranted a weak recommendation.[9 10] Of the 121 discordant recommendations in the Endocrine Society guidelines, 33 (27%) warranted a weak recommendation.[11] These results demonstrate that excessive use of strong recommendations in the face of low certainty evidence is common and concerning.

UpToDate (www.uptodate.com)[15] is an electronic medical textbook that uses GRADE and includes over 9,400 GRADE recommendations[15 16]. UpToDate has instituted intensive training in GRADE methods for their in-house deputy editors who are largely responsible for UpToDate material. Training involves regular large and small group

seminars, and individual feedback from in-house methodologists.

Because it is enormously popular and used by clinicians worldwide, the possibility that UpToDate is issuing misleading strong recommendations on the basis of low certainty evidence constitutes a matter of concern. Therefore, we set out to determine, among all GRADE recommendations in UpToDate, the distribution of strong and weak recommendations, the proportion of discordant recommendations, and to characterize discordant recommendations based on the taxonomy described above (Table 1 & 2). In doing so, we restricted ourselves to the evidence presented in UpToDate, rather than conducting our own literature review. The reason is that our interest was in evaluating UpToDate editors' ability to formulate a GRADEd recommendation from the data they present rather than their ability to find the most relevant data in the literature.

METHODS

Design and data source

We conducted an analytic survey of all GRADE recommendations included in UpToDate. We collaborated with UpToDate to identify all 9451 included in UpToDate as of June 2014, and determined their strength (strong or weak), and their certainty in evidence (high, moderate, or low – UpToDate does not use GRADE's "very low" category). We abstracted the title of each topic, as well as their corresponding clinical domains and age-group populations. From this database, we identified all discordant recommendations included in UpToDate as of January 2015.

Data abstraction on the discordant recommendations

UpToDate topics summarizing the evidence and rationale supporting the recommendations are mostly in narrative formats, and do not provide summary of finding tables or evidence profiles.[17] To assess the appropriateness of discordant recommendations according to the paradigmatic situation defined in the GRADE framework, we therefore standardized data abstraction to collect relevant information from the main text (detailed instruction Supplementary File 1).

Eight reviewers working in six pairs – all working actively as clinicians and proficient in GRADE methodology – performed data abstraction and assessed the appropriateness of discordant recommendations in duplicate. They abstracted the following information related to each discordant recommendation:

- Patient population (clinical field and age group);
- Type of intervention (drug, procedure, device, etc.) and type of comparator

(existing standard care, no intervention, alternative intervention, etc.);

- The clarity of the comparator, classified as (i) clearly and explicitly stated; (ii) not clearly and explicitly stated, but obvious; (iii) not clearly and explicitly stated or obvious, but relatively easy to infer; (iv) not at all clear - uncertain;
- Outcomes: whether there was an explicit statement on mortality as well as the balance of benefits and harms;
- Whether there was an explicit statement on the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternative courses of action;
- Whether issues of cost or resources were explicitly discussed;
- The evidence supporting the recommendation, both for systematic reviews and primary study designs (randomized trials, observational studies, etc.)
- Whether the evidence summary suggested large effects in critical outcomes, or that indirect evidence, not incorporated in the grading, seemed to drive the recommendation.

Based on this abstracted information, each reviewer independently classified each of the discordant recommendations as either consistent with one of the five previously identified optimal categories for discordant recommendations (Table 1)[6 10 13] or in one of three categories in which we judged discordant recommendations to be inconsistent with GRADE guidance (Table 2): (i) good practice statements; (ii) a misclassification of the evidence – the evidence warranted moderate or high certainty rather than low; or (iii) uncertainty in the estimates of effect would best lead to a weak recommendation. We assessed agreement for whether recommendations were appropriate (vs. inappropriate) according to GRADE guidance using the chance-corrected kappa statistic.

The reviewers resolved all disagreements by discussion or through referral to an additional reviewer.

Data analysis and reporting

We abstracted data in an MS Excel database (v. 14.4) with pre-specified response categories whenever possible, and exported in SPSS (v. 22.0) for analysis. We analyzed the recommendation and sample characteristics as natural frequencies and proportions.

RESULTS

The 2971 topics in UpToDate that included GRADE recommendations covered a broad spectrum of clinical fields and health care, including 16.1% in oncology, 49.2% topics in other internal medicine specialties or primary care, and 12.5% in pediatrics. These topics included 9451 GRADE recommendations, of which 6501 (68.8%) were formulated as weak recommendations and 2950 (31.2%) as strong recommendations (Table 3). The proportion of strong recommendations varied greatly across clinical fields, ranging from 5.8% (in dermatology) to 42.7% (in cardiovascular medicine) (Supplementary File 2).

Of the 2950 strong recommendations, 844 (28.6%) were based on high certainty evidence, 1740 (59.0%) on moderate certainty, and 366 (12.4%) were discordant strong recommendations based on low certainty evidence (Table 3). Because UpToDate is continuously updated, 17 recommendations were modified in strength and/or certainty between the time all 9451 recommendations were retrieved, and the time all topics were downloaded for abstraction, as of January 2015.[15] The final study cohort therefore comprised a total of 349 discordant recommendations.

The 349 discordant recommendations were issued across 274 individual topics in UpToDate (each including a range of one to five recommendations), and the topics addressed covered a broad spectrum of health care issues within each clinical field, (Supplementary File 2). Interventions included drugs (56.4% of recommendations), surgery (19.8%), medical devices (6.9%), diagnostic or screening tests (20.9%), and other behavioral or multi-disciplinary interventions (10.0%). These interventions were most often compared to another intervention or to standard of care (56.7%) and less often to no intervention or placebo (36.1%).

The 349 discordant recommendations represent 3.7% of all 9451 recommendations. The proportion of discordant recommendations varied from 0% (e.g. in palliative care, dermatology or for recommendations applying specifically to the elderly population), to 7.0% in pediatrics, 8.0% in infectious disease, and 10.9% in hematology (Supplementary File 2).

Evidence supporting the discordant recommendations

The comparator was clearly and explicitly stated in 73 (20.9%) of the 349 recommendations, not clearly but either obvious or relatively easy to infer in 230 (65.9%) and uncertain in 46 (13.2%). The direction of the recommendation was most often framed in favor of the intervention (78.5%) rather than against it (Table 4).

The full-text of the UpToDate topic often provided a rationale supporting the recommendation. An explicit statement on the balance of benefits and harms was present in 92 (26.4%), and an implicit statement in 157 (45.0%), and no statement in 100 (28.7%). Explicit statements addressing the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternatives were present in 10 (2.9%) of the recommendations; they could be inferred in 171 (49.0%), but not in the remaining 168 (48.1%) of discordant recommendations. Cost or resources considerations were mentioned in 15 (4.3%). The evidence cited to support each discordant recommendation varied substantially, with a median of 4 references cited, range from 0 to 33, with 45 (12.9%) of recommendations without any citation. Observational studies dominated (203, 58.2%); 49 (14.0%) were supported by a systematic review (Table 4).

Appropriateness of the discordant recommendations

Kappa for the initial taxonomic judgment regarding whether the recommendation was appropriate or inappropriate according to GRADE guidance was 0.46 (moderate agreement). The two reviewers required consensus discussions for 43% of the discordant recommendations. Third party adjudication to determine the appropriate classification was required in 12 of the discordant recommendations (3.4%).

Reviewers judged 204 (58.5%) of the 349 discordant recommendations to be consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations (Table 5). The most common paradigm was a "life-threatening or potentially catastrophical situation", followed by "potential similar benefits, one clearly less risky or costly", "potential catastrophic harm", "uncertain benefits, certain harm", and "established similar benefits, one potentially more risky or costly" (Table 5).

Reviewers judged 47 (13.5%) of the 349 discordant recommendations as "good practice statements"; 38 (10.9%) as a "misclassification of certainty (evidence warranted moderate or high certainty)"; and 60 (17.2%) as warranting a weak recommendation (see Table 5).

DISCUSSION

Among 9451 GRADE recommendations in UpToDate, about two thirds were formulated as weak recommendations and the remainder as strong recommendations. Of all recommendations, only 3.7% (n=349) were strong recommendations based on low certainty in effect estimates (Table 3). Of these discordant recommendations, over half were consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations; approximately 14% represented "good practice statements"; approximately 11% were based on a misclassification of certainty (evidence warranted moderate or high certainty), and approximately 17% were judged to warrant a weak recommendation (Table 5). The proportion of appropriate discordant recommendations varied across intervention types or clinical fields (Supplementary File 2). Although most topics in UpToDate provided a rationale to support the discordant recommendation, 29% lacked statements about benefits and harms and 13% did not provide citations, which points at potential areas of improvement for UpToDate related to standards for trustworthy guidelines.[1]

Strengths and limitations

This study assessed the strength of recommendations in the largest known sample of recommendations developed using GRADE. Indeed, even large guidelines include a few hundred recommendations[18], whereas UpToDate topics have one of the largest known coverage in clinical fields and included 9451 recommendations at the time of this assessment.

The taxonomy that we used has been successfully implemented in two prior studies of clinical guidelines[10 11] (see below: relation to prior work). Our reviewers could all be

characterized as expert GRADE methodologists: they were clinical epidemiologists with an in-depth understanding of GRADE methodology acquired through use of GRADE in a large number of assessments over a period of years and were therefore well equipped to assess judgments on evidence and recommendations. This differs markedly from UpToDate authors (some with little understanding of GRADE) and UpToDate editors (all of whom have received basic GRADE training, but some little more than that). Despite the advanced skills of our reviewers, chance corrected kappa agreement on the appropriateness of recommendations was moderate (0.48).[19] Consensus discussions were needed for 43% of discordant recommendations, although formal adjudication by third parties was required for only 12 discordant recommendations (3.4%).

The necessity for frequent consensus discussions reflects the substantial judgment required in categorizing recommendations. This is in part due to the narrative nature of UpToDate topics, which does not include formal summary of finding tables or evidence profiles[17], often discussing the evidence and rationale for several recommendations in a free-text cross-referenced structure that sometimes omits statements regarding benefits and harms, and lacks citations. The one previous study using this taxonomy that addressed chance-corrected agreement reported a kappa of 0.68. The higher kappa may well be a result of more explicit reporting with use of summary of findings tables in the WHO guidelines that were the subject of investigation. The concern regarding the need for consensus discussions is perhaps increased because a single team using a single system of categorization undertook the study. A further limitation of our study is that decisions were based solely on information published in UpToDate, while authors of the topics may have considered other factors.[20]

Another element contributing to the challenges in making categorizations is the clarity of the comparison on which the recommendation applies. As in previous assessment in

guidelines[9], the comparator was clearly and explicitly stated in only 73 (20.9%) of discordant recommendations and was uncertain in 46 (13.2%). When comparators were not clear and explicit, reviewers' inferences may not always have been correct.[20]

Relation to previous work

Two prior studies provided a formal structured exploration of discordant recommendations using the GRADE approach. An assessment of 357 recommendations in 17 Endocrine Society Guidelines found that only 29% of discordant recommendations were consistent with one of the 5 paradigmatic situations.[11] A second study of 456 recommendations in 116 WHO guidelines using GRADE found that of 160 discordant recommendations, only 15.6% were judged consistent with GRADE guidance.[9 10]

Our results contrast with these previous two studies. First, the proportion of weak recommendations was approximately 30% higher in UpToDate than in WHO and Endocrine Society guidelines. This proportion was, however, similar to the 9th edition ACCP guideline on Antithrombotic Therapy and Prevention of Thrombosis, after it implemented GRADE.[18 21] Second, the proportion of inappropriate, discordant recommendation was considerably lower. Of the discordant recommendations, the proportion that should have been weak was about 17%, rather than 27% (Endocrine Society)[11] or 46% (WHO guidelines).[9]

A subsequent interview of panel members involved in the WHO guidelines highlighted reasons contributing to discordant recommendations. These included political considerations around long-established practices, the need for funding and policy formulation, or the fear of pushback from media.[20] Panel members also expressed skepticism regarding the value of making weak recommendations, or concerns they may

be ignored[20], although another study reported that WHO weak recommendations are frequently adopted in national policies (uptake of 61% for weak recommendations versus 82% for strong recommendations).[22] Finally, the authors identified both financial and intellectual conflicts of interest among panel members as an explanation for discordant recommendations.[20 23] Any or all of these factors may have contributed to UpToDate discordant recommendations.

Implications and conclusion

For users of UpToDate, our results are generally, though not absolutely, reassuring. The proportion of discordant recommendations is very small – only 3.7% of all recommendations. Furthermore, of the three categories inconsistent with GRADE guidance – good practice statement, misclassification of the certainty, and evidence warranting a weak recommendation (Table 2) – the third is by far the most problematic.[9] Good practice statements are appropriate when indirect evidence that is difficult to collect and summarize warrants high certainty in the impact of a given intervention and when the balance benefits and harms is large.[14] Thus, in terms of implications for clinical practice, good practice statements have the same force as strong recommendations. Similarly with misclassification of certainty: since the certainty is actually moderate or high, a strong recommendation is appropriate. Recommendations that should have been weak instead of strong provide inappropriate "just do it" guidance for clinical practice, although they are actually preference-sensitive and should thus warrant shared-decision making.[8] Of the 349 discordant recommendations in UpToDate, only 60 fall in the category of inappropriate strong recommendations.

Thus, clinicians using UpToDate can anticipate that they will be misleadingly instructed

to take a "just do it" rather than an "it depends" approach to clinical decision making in 0.6% (6 of 1,000) UpToDate recommendations.[15] This seems close to a threshold in which one might ignore the problem. Nevertheless, we would still encourage clinicians to be alert to the possibility of an inappropriate strong recommendation – in UpToDate or elsewhere – whenever the recommendation is based on low certainty evidence and authors fail to provide an explicit rationale corresponding to one of the categories in Table 1.

A likely explanation for UpToDate's success in avoiding inappropriate discordant recommendations is the training and feedback that their deputy editors receive. For organizations using GRADE, our results suggest the desirability of such training for those involved in formulating recommendations to optimize use of GRADE.

Finally our results highlight the need for authors of trustworthy recommendations or guidelines[1] to provide clear and explicit comparators, as well as transparent and systematic reports of the key ingredients of their rationale when moving from evidence to recommendation.[17 24 25] Future avenues for research should also look at optimal presentation formats of EBM textbooks and guidelines, to ensure clinicians actually understand both the rationale and potential implications of all recommendations for clinical practice.[8 26-29]

LIST OF ABBREVIATIONS

GRADE: Grading of Recommendations Assessment, Development and Evaluation

WHO: World Health Organization

ACCP: American College of Chest Physicians

COMPETING INTERESTS

TA, AK, IN, RBP, PEA, DR, POV, and GHG are active members of the GRADE working group.

DR, at the time the data on graded recommendations was extracted from UpToDate and until 2016, was an employee of UpToDate – he reports personal fees from UpToDate, outside the submitted work.

GHG contributes to the training in GRADE methods for UpToDate in-house deputy editors, for which he reports personal fees from UpToDate, outside the submitted work.

CONTRIBUTORS

Thomas Agoritsas (TA) and Gordon H. Guyatt (GHG) designed the study. David Rind (DR) provided the list of all recommendations and grading from UpToDate. Paul E. Alexander (PEA) helped structuring data abstraction. Thomas Agoritsas (TA), Arnaud Merglen (AM), Anja F. Heen (AFH), Annette Kristiansen (AK), Ignacio Neumann (IN), Juan P. Brito (JPB), Romina Brignardello-Petersen (RBP), and Per O. Vandvik (POV) reviewed the recommendations in duplicate and classified them according to GRADE taxonomy. Thomas Agoritsas (TA) and Gordon Guyatt (GG) wrote the first draft of the manuscript. All authors have read the manuscripts and made improvements of the content and wording.

FUNDING / ACKNOWLEDGMENTS

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

DATA SHARING STATEMENT

There were no additional unpublished data from this study.

REFERENCES

1. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. . In: Graham R, Mancher M, Miller Wolman D, et al., eds. Clinical Practice Guidelines We Can Trust. Washington (DC), 2011.

2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;**336**(7650):924-6.

3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;**64**(4):383-94.

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. Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013.

7. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;**64**(4):401-6.

8. Agoritsas T, Heen AF, Brandt L, et al. Decision aids that really promote shared decision making: the pace quickens. BMJ 2015;**350**:g7624.

9. Alexander PE, Brito JP, Neumann I, et al. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. J Clin Epidemiol 2016;**72**:98-106.

10. Alexander PE, Bero L, Montori VM, et al. World Health Organization recommendations are often strong based on low confidence in effect estimates. J Clin Epidemiol 2014;**67**(6):629-34.

11. Brito JP, Domecq JP, Murad MH, et al. The Endocrine Society guidelines: when the confidence cart goes before the evidence horse. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3246-52.

12. Vigersky RA, Bhasin S, Martin KA. The Endocrine Society Clinical Practice Guidelines: a self-assessment. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3174-7.

13. Neumann I, Santesso N, Akl EA, et al. A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach. J Clin Epidemiol 2016;**72**:45-55.

14. Guyatt GH, Schunemann HJ, Djulbegovic B, et al. Guideline panels should not GRADE good practice statements. J Clin Epidemiol 2014.

15. UpToDate, Waltham, MA. http://www.uptodate.com (Accessed on July 7th, 2017).

16. Agoritsas T, T Vandvik PO, Neumann I, Rochwerg B, Jaeschke R, Hayward R, Guyatt GH, McKibbon A. Chapter 5. Finding Current Best Evidence, in JAMA Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3nd Edition, McGraw-Hill Medical, 2015.

17. Guyatt G, Oxman AD, Akl E, et al. GRADE guidelines 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2010.

18. Agoritsas T, Neumann I, Mendoza C, et al. Guideline conflict of interest management and methodology heavily impacts on the strength of recommendations: comparison between two iterations of the American College of Chest Physicians Antithrombotic Guidelines. J Clin Epidemiol 2017;**81**:141-43.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;**33**(1):159-74.

20. Alexander PE, Gionfriddo MR, Li SA, et al. A number of factors explain why WHO guideline developers make strong recommendations inconsistent with GRADE guidance. J Clin Epidemiol 2016;**70**:111-22.

21. Guyatt GH, Norris SL, Schulman S, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;**141**(2 Suppl):53S-70S.

22. Nasser SM, Cooke G, Kranzer K, et al. Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. J Clin Epidemiol 2015;**68**(6):703-7.

23. Alexander PE, Li SA, Gionfriddo MR, et al. Senior GRADE methodologists encounter challenges as part of WHO guideline development panels: an inductive content analysis. J Clin Epidemiol 2016;**70**:123-8.

24. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;**64**(4):395-400.

25. Guyatt GH, Oxman AD, Kunz R, et al. Going from evidence to recommendations. BMJ 2008;**336**(7652):1049-51.

26. Vandvik PO, Brandt L, Alonso-Coello P, et al. Creating clinical practice guidelines we can trust, use, and share: a new era is imminent. Chest 2013;**144**(2):381-9.

27. Kristiansen A, Brandt L, Alonso-Coello P, et al. Development of a novel multilayered

presentation format for clinical practice guidelines. Chest 2014.

28. Treweek S, Oxman AD, Alderson P, et al. Developing and evaluating communication strategies to support informed decisions and practice based on evidence (DECIDE): protocol and preliminary results. Implement Sci 2013;**8**:6.

29. Siemieniuk RA, Agoritsas T, Macdonald H, et al. Introduction to BMJ Rapid Recommendations. BMJ 2016;**354**:i5191.

Table 1. Paradigmatic situations in which a strong recommendation may be warranted despite low or very low certainty in effect estimates *(appropriate strength, consistent with GRADE)*

Situation	Certainty in Estimates (Quality of Evidence)		Balance of Benefits and Harms	Values and Preferences	Resource Considerations	Recommendation	Example
	Benefits	Harms					
1. Life-threatening (or catastrophical) situation	Low or very low	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation; adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits justify the intervention	Strong recommendation in favor of the intervention	Indirect evidence from seasonal influenza suggests that patients with avian influenza may benefit from the use of oseltamivir (low certainty in effect estimates). Given the high mortality of the disease and the absence of effective alternatives, the WHO made a strong recommendation in favor of the use of oseltamivir rather than no treatment in patients with avian influenza.
2. Uncertain benefit, certain harm	Low or very low	High or moderate	Possible but uncertain benefit; substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	High incremental cost (or resource use) relative to the benefits may not justify the intervention	Strong recommendation against the intervention	In patients with idiopathic pulmonary fibrosis, treatment with azathioprine plus prednisone offers a possible but uncertain benefit in comparison with no treatment. The intervention, however, is associated with a substantial established harm. An international guideline made a recommendation against the combination of corticosteroids plus azathioprine in patients with idiopathic pulmonary fibrosis.
3. Potential equivalence, one option clearly less risky or costly	Low or very low	High or moderate	Magnitude of benefit apparently similar—though uncertain—for alternatives; we are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation for less harmful/less expensive	Low-quality evidence suggests that initial Helicobacter pylori eradication in patients with early stage extranodal marginal zone (MALT) B-cell lymphoma results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy, but with high certainty of less harm, morbidity, and cost. Consequently, UpToDate made a strong recommendation in favor of H pylori eradication rather than radiotherapy in patients with MALT lymphoma.
4. High certainty in similar benefits, one option potentially more risky or costly	High or moderate	Low or very low	Established that magnitude of benefit is similar for alternative management strategies; best (though uncertain) estimate is that one alternative has appreciably greater harm	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation against the intervention with possible greater harm	In women requiring anticoagulation and planning conception or in pregnancy, high certainty estimates suggest similar effects of different anticoagulants. However, indirect evidence (low certainty in effect estimates) suggests potential harm to the unborn infant with oral direct thrombin (eg, dabigatran) and factor Xa inhibitors (eg, rivaroxaban, apixaban). The AT9 guidelines recommended against the use of such anticoagulants in women planning conception or in pregnancy.
5. Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) relative to the benefits, may not justify the intervention	Strong recommendation against the intervention	In males with androgen deficiency, testosterone supplementation likely improves quality of life. Low- certainty evidence suggests that testosterone increases cancer spread in patients with prostate cancer. The US Endocrine Society made a recommendation against testosterone supplementation in patients with prostate cancer.

Reproduced and adapted from Neumann & al.[13]

Table 2. Reasons for issuing strong recommendation based on low certainty in effect estimates inconsistent with GRADE guidance

Situation	Example
Best practice recommendation (for which sensible alternatives do not exist)	"For patients with congenital adrenal hyperplasia, we recommend monitoring patients for signs of glucocorticoid excess, as well as for signs of inadequate androgen suppression." This statement should not have been GRADEed as sensible alternatives do not exist
The strong recommendation was warranted because the certainty of the evidence was actually moderate rather then low	"We recommend intensive lifestyle modification to the entire family and to the patient, and as the prerequisite for all overweight and obesity treatments for children and adolescents." The authors classified this as low quality evidence; our judgment is that the correct classification is moderate quality.
Lack of compelling explanation (the recommendation should have been weak)	"If a patient is unable or unwilling to undergo surgery, we recommend medical treatment with mineralocorticoids" Lack of evidence of mineralocorticoids being superior to other medical treatment (eg, anti-hypertensive medications)

Elements adapted from Brito et al.[11]

Table 3. Distribution of the strength of the recommendations in UpToDate according to the certainty in evidence

	Weak Recomendations	Strong Recommendations	All Recommendations
	N (%)	N (%)	N (%)
Low certainty	4335 (66.7%)	366 (12.4%)	4701 (49.7%)
Moderate certainty	2019 (31.1%)	1740 (59.0%)	3759 (39.8%)
High certainty	147 (2.3%)	844 (28.6%)	991 (10.5%)
Total	**6501** (68.8% of all rec)	**2950** (31.2% of all rec)	**9451** (100%)

Table 4. Characteristics of all 349 discordant recommendations in UpToDate, and proportion of appropriate discordant recommendations

	N	(%)	% of appropriate discordant *(p-value)*
Clinical Specialtes			*(p = 0.160)* *
Primary Care and General Internal Medicine	15	(4.3)	53.3
Emergency Medicine	16	(4.6)	81.3
Critical Care	5	(1.4)	80.0
Internal Medicine specialties	158	(45.3)	57.6
Oncology (including hemato-oncology)	43	(12.3)	55.8
Pediatrics	73	(20.9)	47.9
Obstetrics, Gynecology and Women Health	19	(5.4)	73.7
General Surgery	13	(3.7)	69.2
Anesthesiology	3	(0.9)	100.0
Psychiatry	4	(1.1)	75.0
Intervention type			*(p = 0.010)*
Drug intervention	197	(56.4)	61.4
Surgical interventions	69	(19.8)	59.4
Medical device	24	(6.9)	62.5
Behavioural or multi-disciplinary intervention	35	(10.0)	57.1
Diagnostic test, screening programms	24	(6.9)	29.2
Clarity of the comparator			*(p <0.001)*
Comparator not at all clear – uncertain	46	(13.2)	37.0
Comparator not clearly and explicitly stated or obvious, but relatively easy to infer	120	(34.4)	48.3
Comparator not clearly and explicitly stated, but obvious	110	(31.5)	68.2
Comparator clearly and explicitly stated	73	(20.9)	74.0
Type of comparator			*(p = 0.083)*
Too unclear	25	(7.2)	44.0
No intervention (or placebo)	126	(36.1)	54.0
Other intervention(s) (standard of care or alternative(s))	198	(56.7)	63.1
Direction of the recommendation			*(p <0.001)*
For the intervention (i.e. against the comparator)	274	(78.5)	51.1
Against the intervention (i.e. for the comparator)	75	(21.5)	85.3
Mortality			*(p <0.001)*
No statement about mortality	189	(54.2)	47.1
Implicit statement about mortality	47	(13.5)	68.1
Explicit statement about mortality	113	(32.4)	73.5
Balance of benefits and harms			*(p <0.001)*
No statement about the balance of outcomes	100	(28.7)	28.0
Implicit statement about the balance of outcomes	157	(45.0)	66.9
Explicit statement about the balance of outcomes	92	(26.4)	77.2

Relative Importance of outcomes - Values & Preferences			*(p <0.001)*
No statement about the relative importance of outcomes	168	(48.1)	42.9
Implicit statement about the relative importance of outcomes	171	(49.0)	73.1
Explicit statement about the relative importance of outcomes	10	(2.9)	70.0
Cost of resources			*(p = 0.023)*
No statement about cost or resources	334	(95.7)	57.2
Cost or resources clearly and explicitly stated	15	(4.3)	86.7
Supporting systematic review (SR)			*(p = 0.175)*
No SR is cited	300	(86.0)	56.3
SR of Observational studies	22	(6.3)	63.6
SR of both RCT and Observational studies	13	(3.7)	76.9
SR of Randomized Trials (RCT)	14	(4.0)	78.6
Design of primary studies			*(p = 0.002)*
No reference cited	45	(12.9)	35.6
Other type (eg narrative review, book chapter)	48	(13.8)	54.2
Observational studies	203	(58.2)	61.1
Randomized Trials (RCT)	53	(15.2)	71.7
Total	**349**	**(100)**	**58.5**

* The null hypothesis for the p-value is that the proportions do not differ across categories.

Table 5. Summary judgments on the appropriateness of 349 discordant strong recommendation based on low certainty in effect in UpToDate

	N	(%)
Appropriate discordant recommendations (consistent with GRADE)		
1. Life-threatening (or catastrophical) situation	70	(20.1)
2. Uncertain Benefit, Certain Harm	28	(8.0)
3. Potential similar benefits, One clearly less risky (or costly)	56	(16.0)
4. Established similar benefits, One potentially more risky (or costly)	18	(5.2)
5. Potential catastrophic harm	32	(9.2)
Total	**204**	**(58.5)**
Inappropriate discordant recommendations (inconsistent with GRADE)		
6. Good Practice Statement	47	(13.5)
7. Misclassification of certainty (judged moderate or high)	38	(10.9)
8. Lack of explanation, should have been weak recommendation (GRADE 2C)	60	(17.2)
Total	**145**	**(41.5)**

Supplementary File 1. INSTRUCTION FOR ABSTRACTION

Please read carefully before starting abstraction.
For any questions please contact me ASAP (thomas.agoritsas@gmail.com)

Background

- In a first phase of the project, we characterized the strength and confidence of the 9451 GRADE recommendations in UTD.
- In this last phase, we are focusing on
  - The **349 strong recommendations based on low confidence (GRADE 1C)**
  - Which are included in a total of **274 topics** (=chapters in UpToDate).
- The main objective is to categorize them based according the following taxonomy
  - Appropriate grading: recommendation consistent with one of the five paradigmatic situations defined the GRADE framework (see examples in Table 1 below):
    - *[App#1] Life-threatening situation*
    - *[App#2] Uncertain Benefit, Certain Harm*
    - *[App#3] Potential similar benefits, one clearly less risky (or costly)*
    - *[App#4] Established similar benefits, one potentially more risky (or costly)*
    - *[App#5] Potential catastrophic harm of one option*
  - Inappropriate grading: recommendation inconsistent with GRADE (see examples in Table 2 below), in short:
    - *[Inapp#1] Good Practice Statement*
    - *[Inapp#2] Misclassification of confidence (should have been GRADE 1B or 1A)*
    - *[Inapp#3] Lack of explanation, should have been a weak rec (GRADE 2C)*

→ **Before starting, please read the examples and Appendix Tables 1 & 2, they are also embedded in separate tabs in the abstraction excel file. Do not focus on memorizing them, as data abstraction will guide you in your judgment.**

Abstraction Excel File & Variables

- We will conduct the whole abstraction process in the attached standardized excel file
- We have kept the variables to abstract to the minimum necessary, most with **pre-defined response categories in drop-down menus** (and infrequently as free-text for copy-pasting).
- **Each recommendation has a separate row in the file**. There are sometimes more than GRADE 1C per UTD topic (the number are indicated).
- **As a guiding principle, keep in mind that the main objective is to assess the most appropriate taxonomy (first as appropriate or inappropriate, then subcategory).**
  These are the last variables in the file.
- This requires judgment based on what is reported in the UpToDate topic, mostly in narrative form. Indeed, there are typically no "summary of findings tables" of "evidence profiles" to explicit GRADE assessment. Absolute certainty in taxonomy is sometimes hard to achieve, but try and assign the best fit you can.
- The abstraction form is organized as follows to guide your final judgment:

UpToDate TOPIC & RECOMMENDATION	Pre-entered data to help you identify the relevant topic in the UpToDate topic in the dropbox
POPULATION	Check pre-entered clinical field, document age-group
INTERVENTION	Intervention type
COMPARATOR	Clarity and type of the comparator, direction of rec
OUTCOME – Benefits & Harms	Statement about: mortality, balance of benefits & harms, their relative importance (ie. values/preferences), cost
EVIDENCE supporting the rec	Date of literature review & updates number and type of supporting evidence Indication regarding the potential role of indirect evidence, Presence of large effects.
Conflict of interest (COI)	Copy paste statement, presence of financial COI
TAXONOMY – APPROPRIATE	**Separate judgement on each of the 5 paradigmatic situation defined by GRADE** (clear, possible, no)
TAXONOMY – INAPPROPRIATE	**Separate judgmenet on each of the 3 inappropriate situations**
CONFLICT RESOLUTION	**TAXONOMY DECISION** (for kappa), confidence in the decision (to document), assessing agreement and **RECONCILED TAXONOMY** within each pair of abstractor.
ADJUDICATION	Recording adjudication third reviewer if this was necessary

- **Specific guidance for each variable is found in the GREY BOXES on top of each column.**
- **Please read and select best option from DROP-DOWN menus within each cell**
- A few cells are for free-text to copy paste from the topic. Be sure to double click in the cell before pasting content, to keep the format intact.
- Most variables are followed by a column labelled **"additional comments"** or "rationale". These are for your personal notes to guide conflict resolution.
- A few examples already abstracted are shown in the first rows as an indication.

Abstraction: STEP-BY-STEP

Start abstracting a few first recommendations to get familiar with the process and contact me (thomas.agoritsas@gmail.com) for any question. I'm happy to have a quick skype if and as often as necessary.

- Go to the recommendation in the next row in the excel file.
- The [topic_#] and [topic_title] correspond to the name of the **PDF file** for the corresponding UTD topic → Open it.
- **Search automatically (ctrl-R or command-F) for "1C" → This will directly lead you to the GRADE 1C recommendation(s) at the end of the topic under the "Summary and Recommendations" section.**
- Read it carefully, and take a few seconds to try and get some first rough impression re: potential taxonomy
- **Then, find the corresponding paragraph in the topic that supports the recommendation** (it is often indicated soon after the recommendation (e.g. "See treatment…"). If not, try and find which paragraph(s) discuss(es) the recommendation.
- Abstract all variables in the order of the file as this will guide your formal judgment.
- Then judge each of the 5 appropriate and 3 inappropriate categories in the taxonomy.
- Then decide which one fits best and document the confidence you have in your assessment.
- Go to the next recommendation in the next row. (if this is in the same topic, you'll be able to copy several or your answers.

Conflict resolution and adjudication

- **You've been assigned with a paired reviewer. Schedule a first conflict resolution in the following days to ensure you are on the same page.**
- Record the agreed taxonomy in the specific column ("RECONCILED TAXONOMY"). Do NOT modify your initial judgement ("TAXONOMY DECISION") - as this will use to calculate kappa.
- **If you cannot resolve conflict, send us your questions for adjudication by a third reviewer.**
- Record adjudication in the final column.

➢ **Thanks again for your help. Do not hesitate to contact me for any questions (thomas.agoritsas@gmail.com)**

Supplementary File 2.

Characteristics of all 9451 recommendations in UpToDate: certainty effect estimates and clinical fields

	N	(%)	% of Strong Rec	% of strong rec discordant	% of any rec being discordant
Clinical Fields					
1 Primary Care & General Internal Medicine	356	(3.8)	22.5	16.3	3.7
2 Emergency Medicine	295	(3.1)	25.1	**23.0**	**5.8**
3 Critical Care	144	(1.5)	34.0	10.2	3.5
4 Cardiovascular Medicine	529	(5.6)	**42.7**	5.3	2.3
5 Infectious Diseases	870	(9.2)	**41.3**	**19.5**	**8.0**
6 Nephrology and Hypertension	475	(5.0)	**39.8**	6.3	2.5
7 Pulmonary Medicine	347	(3.7)	34.6	6.7	2.3
8 Hematology (non-oncology)	192	(2.0)	**41.1**	**26.6**	**10.9**
9 Neurology	395	(4.2)	31.1	**17.9**	**5.6**
10 Allergy and Immunology	261	(2.8)	23.4	3.3	0.8
11 Endocrinology & Diabetes	504	(5.3)	19.0	6.3	1.2
12 Gastroenterology & Hepatology	471	(5.0)	22.9	7.4	1.7
13 Rheumatology	216	(2.3)	21.3	4.3	0.9
14 Palliative Care	33	(0.3)	18.2	0.0	0.0
15 Oncology	1255	(13.3)	**36.0**	7.7	2.8
16 Hemato-oncology	263	(2.8)	27.8	**20.5**	**5.7**
17 Pediatrics	1057	(11.2)	**39.5**	**17.7**	**7.0**
18 Pediatric Emergency Medicine	126	(1.3)	28.6	5.6	1.6
19 Gynecology & Obstetrics	709	(7.5)	22.3	12.0	2.7
20 General Surgery	403	(4.3)	**32.8**	12.1	4.0
21 Anesthesiology	48	(0.5)	16.7	**37.5**	**6.3**
22 Dermatology	240	(2.5)	5.8	0.0	0.0
23 Psychiatry	262	(2.8)	16.8	9.1	1.5
TOTAL	**9451**	**(100)**	**31.2**	**12.4**	**3.9**

BMJ Open

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2017-018593.R2
Article Type:	Research
Date Submitted by the Author:	24-Sep-2017
Complete List of Authors:	Agoritsas, Thomas; University Hospitals of Geneva, Division of General Internal Medicine & Division of Clinical Epidemiology; McMaster University Faculty of Health Sciences, Department of Health Research Methods, Evidence, and Impact Merglen, Arnaud; University Hospitals of Geneva, Division of General Pediatrics Heen, Anja; Innlandet Hospital Trust-division Gjøvik, Department of Internal Medicine Kristiansen, Annette; Inland hospital trust, Internal medicine, Gjøvik Neumann, Ignacio; Pontificia Universidad Catolica de Chile, Department of Internal Medicine; McMaster University, Department of Health Research Methods, Evidence, and Impact, Brito, Juan; Mayo Clinic Minnesota, Department of Medicine and Knowledge and Evaluation Research Unit, Brignardello-Petersen, Romina; McMaster University, Department of Health Research Methods, Evidence, and Impact Alexander, Paul; McMaster University, Department of Health Research Methods, Evidence, and Impact Rind, David; Institute for Clinical and Economic Review Vandvik, Per; Norwegian Knowledge Centre for the Health Services, Guyatt, Gordon; Mcmaster University, Department of Health Research Methods, Evidence, and Impact
Primary Subject Heading:	Evidence based practice
Secondary Subject Heading:	Epidemiology
Keywords:	Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine

UpToDate adherence to GRADE criteria for strong recommendations: an analytic survey

Thomas Agoritsas, MD, PhD [1,2]

Arnaud Merglen, MD, MSc [3]

Anja Fog Heen, MD [4]

Annette Kristiansen, MD, PhD [4]

Ignacio Neumann, MD, PhD [2,5]

Juan P Brito, MD, MSc [6]

Romina Brignardello-Petersen, DDS, MSc, PhD [2, 7]

Paul E Alexander, MSc, PhD [2]

David M Rind, MD, MSc [8]

Per O. Vandvik, MD, PhD [4,9]

Gordon H Guyatt, MD, MSc [2]

Affiliations:

[1] Division of General Internal Medicine & Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland.

[2] Department of Health Research Methods, Evidence, and Impact, McMaster University, Faculty of Health Sciences, Hamilton, Ontario, Canada

[3] Division of General Pediatrics, University Hospitals of Geneva & Faculty of Medicine, University of Geneva, Geneva, Switzerland.

[4] Department of Internal Medicine, Innlandet Hospital Trust-division Gjøvik, Norway

[5] Department of Internal Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile

[6] Division of Endocrinology, Diabetes, Metabolism and Nutrition, Department of Medicine and Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, USA

[7] Faculty of Dentistry, University of Chile, Chile.

[8] Institute for Clinical and Economic Review, Boston, MA, USA

[9] Institute of Health and Society, Faculty of Medicine, University of Oslo, Norway

*** Correspondence to**: Thomas Agoritsas, MD, PhD (thomas.agoritsas@gmail.com)

Division of General Internal Medicine & Division of Clinical Epidemiology

Department Internal Medicine, Rehabilitation and Geriatrics (DMIRG)

University Hospitals of Geneva, Rue Gabrielle-Perret-Gentil 4,

1211 Genève 14, Switzerland

Phone: +41 79 55 34 543

Keywords: Clinical Practice Guidelines, Strength of Recommendations, Quality of the Evidence, Clinical Decision Making, Evidence-Based Medicine.

Abstract: 295 words / **Manuscript**: 3064 words

Tables: 5 / Supplementary Files: 2

ABSTRACT

Introduction

UpToDate is widely used by clinicians worldwide and includes more than 9,400 recommendations that apply the GRADE framework. GRADE guidance warns against strong recommendations when certainty of the evidence is low or very low (discordant recommendations), but has identified five paradigmatic situations in which discordant recommendations may be justified.

Objectives

Our objective was to document the strength of recommendations in UpToDate and assess the frequency and appropriateness of discordant recommendations.

Design

Analytic survey of all recommendations in UpToDate

Methods

We identified all GRADE recommendations in UpToDate, and examined their strength (strong or weak) and certainty of the evidence (high, moderate, or low certainty). We identified all discordant recommendations as of January 2015, and pairs of reviewers independently classified them either into one of the five appropriate paradigms or into one of three categories inconsistent with GRADE guidance, based on the evidence presented in UpToDate.

Results

UpToDate included 9451 GRADE recommendations, of which 6501 (68.8%) were

formulated as weak recommendations and 2950 (31.2%) as strong. Among the strong, 844 (28.6%) were based on high certainty in effect estimates, 1,740 (59.0%) on moderate certainty, and 366 (12.4%) on low certainty. Of the 349 discordant recommendations 204 (58.5%) were judged appropriate (consistent with one of the five paradigms); we classified 47 (13.5%) as good practice statements; 38 (10.9%) misclassified the evidence as low certainty when it was at least moderate; and 60 (17.2%) warranted a weak rather than a strong recommendation.

Conclusion

The proportion of discordant recommendations in UpToDate is small (3.7% of all recommendations), and the proportion that is truly problematic (strong recommendations that would best have been weak) very small (0.6%). Clinicians should nevertheless be cautious, and look for clear explanations – in UpToDate and elsewhere – when guidelines offer strong recommendations based on low certainty evidence.

Strengths and limitations of this study

- We assessed the strength of recommendations in the largest known sample of recommendations using GRADE (N=9451) addressing a wide array of clinical fields.
- We used a taxonomy to appraise discordant recommendations that has been successfully implemented in two prior assessments of clinical practice guidelines.
- We based our assessment solely on information published in UpToDate, while authors of the topics may have considered other factors in deciding to issue a discordant recommendation.
- UpToDate topics are narrative in nature and do not include formal summary of finding tables. As a result, the comparators were often not clearly stated, which may have influenced the reviewers' inferences about the discordant recommendations.

INTRODUCTION

To ensure that patients receive optimal care, consistent with their values and preferences, clinicians need trustworthy recommendations based on transparent ratings of certainty of evidence and strength of recommendations.[1] The widely adopted GRADE system (Grading of Recommendations Assessment, Development and Evaluation) offers a systematic and transparent framework to rate certainty (also referred to as quality or confidence) of evidence and to move from evidence to recommendations.[2-5]

Using GRADE, guideline-makers issue strong recommendations when they are confident that the desirable consequences clearly outweigh the undesirable consequences.[6 7] Conversely they should issue weak (also called conditional) when the balance of desirable and undesirable consequences between alternatives is close, the certainty in evidence is low, uncertainty or variability in patients' values and preferences is large, or cost-effectiveness is questionable.[6] Strong recommendations represent "just do it" recommendations applicable to almost all patients; weak recommendations are applicable to the majority of patients and include preference-sensitive decisions that require clinicians to ensure, through shared-decision making, that patients' choices are congruent with their values.[8]

GRADE views strong recommendations in the face of low certainty evidence (we will refer to such situations as *discordant recommendations*) as questionable, and often inappropriate. Some guidelines have a clear surfeit of discordant recommendations. For example, of 456 recommendations in 116 WHO guidelines, 160 (35%) proved discordant.[9 10] Similarly 121 of 357 (34%) recommendations in 17 Endocrine Society Guidelines proved discordant.[11 12]

Though discordant recommendations often represent a violation of GRADE guidance,

this is not always the case. GRADE has identified 5 seldom-occurring paradigmatic situations in which a strong recommendation is warranted despite low certainty in the evidence (Table 1).[6 13] Further, there is more than one explanation for an apparent violation of GRADE guidance (a discordant recommendation that fails to meet one of these criteria). First, the discordant recommendation may actually represent a good practice statement, in which indirect evidence justifies an inference that the recommended management option is far superior to the alternative.[14] Indirect evidence refers to evidence that does not directly address the question at hand, but nevertheless bears on the question. For instance, though there are no randomized trials of use of a parachute after jumping out of plane, there is ample indirect evidence suggesting its impact on mortality from the jump. Second, the panel may have misclassified the certainty of the evidence (it may actually be moderate or high). Third, and most concerning, the optimal management option is in fact value and preference-sensitive and the panel should have issued a weak recommendation (Table 2).[6 13]

Of the 160 discordant recommendations in the WHO guideline, 73 (46%) fell into the most concerning category of those that warranted a weak recommendation.[9 10] Of the 121 discordant recommendations in the Endocrine Society guidelines, 33 (27%) warranted a weak recommendation.[11] These results demonstrate that excessive use of strong recommendations in the face of low certainty evidence is common and concerning.

UpToDate (www.uptodate.com)[15] is an electronic medical textbook that uses GRADE and includes over 9,400 GRADE recommendations[15 16]. UpToDate has instituted intensive training in GRADE methods for their in-house deputy editors who are largely responsible for UpToDate material. Training involves regular large and small group

seminars, and individual feedback from in-house methodologists.

Because it is enormously popular and used by clinicians worldwide, the possibility that UpToDate is issuing misleading strong recommendations on the basis of low certainty evidence constitutes a matter of concern. Therefore, we set out to determine, among all GRADE recommendations in UpToDate, the distribution of strong and weak recommendations, the proportion of discordant recommendations, and to characterize discordant recommendations based on the taxonomy described above (Table 1 & 2). In doing so, we restricted ourselves to the evidence presented in UpToDate, rather than conducting our own literature review. The reason is that our interest was in evaluating UpToDate editors' ability to formulate a GRADEd recommendation from the data they present rather than their ability to find the most relevant data in the literature.

METHODS

Design and data source

We conducted an analytic survey of all GRADE recommendations included in UpToDate. We collaborated with UpToDate to identify all 9451 included in UpToDate as of June 2014, and determined their strength (strong or weak), and their certainty in evidence (high, moderate, or low – UpToDate does not use GRADE's "very low" category). We abstracted the title of each topic, as well as their corresponding clinical domains and age-group populations. From this database, we identified all discordant recommendations included in UpToDate as of January 2015.

Data abstraction on the discordant recommendations

UpToDate topics summarizing the evidence and rationale supporting the recommendations are mostly in narrative formats, and do not provide summary of finding tables or evidence profiles.[17] To assess the appropriateness of discordant recommendations according to the paradigmatic situation defined in the GRADE framework, we therefore standardized data abstraction to collect relevant information from the main text (detailed instruction Supplementary File 1).

Eight reviewers working in six pairs – all working actively as clinicians and proficient in GRADE methodology – performed data abstraction and assessed the appropriateness of discordant recommendations in duplicate. They abstracted the following information related to each discordant recommendation:

- Patient population (clinical field and age group);
- Type of intervention (drug, procedure, device, etc.) and type of comparator

(existing standard care, no intervention, alternative intervention, etc.);

- The clarity of the comparator, classified as (i) clearly and explicitly stated; (ii) not clearly and explicitly stated, but obvious; (iii) not clearly and explicitly stated or obvious, but relatively easy to infer; (iv) not at all clear - uncertain;
- Outcomes: whether there was an explicit statement on mortality as well as the balance of benefits and harms;
- Whether there was an explicit statement on the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternative courses of action;
- Whether issues of cost or resources were explicitly discussed;
- The evidence supporting the recommendation, both for systematic reviews and primary study designs (randomized trials, observational studies, etc.)
- Whether the evidence summary suggested large effects in critical outcomes, or that indirect evidence, not incorporated in the grading, seemed to drive the recommendation.

Based on this abstracted information, each reviewer independently classified each of the discordant recommendations as either consistent with one of the five previously identified optimal categories for discordant recommendations (Table 1)[6 10 13] or in one of three categories in which we judged discordant recommendations to be inconsistent with GRADE guidance (Table 2): (i) good practice statements; (ii) a misclassification of the evidence – the evidence warranted moderate or high certainty rather than low; or (iii) uncertainty in the estimates of effect would best lead to a weak recommendation. We assessed agreement for whether recommendations were appropriate (vs. inappropriate) according to GRADE guidance using the chance-corrected kappa statistic.

The reviewers resolved all disagreements by discussion or through referral to an additional reviewer.

Data analysis and reporting

We abstracted data in an MS Excel database (v. 14.4) with pre-specified response categories whenever possible, and exported in SPSS (v. 22.0) for analysis. We analyzed the recommendation and sample characteristics as natural frequencies and proportions.

RESULTS

The 2971 topics in UpToDate that included GRADE recommendations covered a broad spectrum of clinical fields and health care, including 16.1% in oncology, 49.2% topics in other internal medicine specialties or primary care, and 12.5% in pediatrics. These topics included 9451 GRADE recommendations, of which 6501 (68.8%) were formulated as weak recommendations and 2950 (31.2%) as strong recommendations (Table 3). The proportion of strong recommendations varied greatly across clinical fields, ranging from 5.8% (in dermatology) to 42.7% (in cardiovascular medicine) (Supplementary File 2).

Of the 2950 strong recommendations, 844 (28.6%) were based on high certainty evidence, 1740 (59.0%) on moderate certainty, and 366 (12.4%) were discordant strong recommendations based on low certainty evidence (Table 3). Because UpToDate is continuously updated, 17 recommendations were modified in strength and/or certainty between the time all 9451 recommendations were retrieved, and the time all topics were downloaded for abstraction, as of January 2015.[15] The final study cohort therefore comprised a total of 349 discordant recommendations.

The 349 discordant recommendations were issued across 274 individual topics in UpToDate (each including a range of one to five recommendations), and the topics addressed covered a broad spectrum of health care issues within each clinical field, (Supplementary File 2). Interventions included drugs (56.4% of recommendations), surgery (19.8%), medical devices (6.9%), diagnostic or screening tests (20.9%), and other behavioral or multi-disciplinary interventions (10.0%). These interventions were most often compared to another intervention or to standard of care (56.7%) and less often to no intervention or placebo (36.1%).

The 349 discordant recommendations represent 3.7% of all 9451 recommendations. The proportion of discordant recommendations varied from 0% (e.g. in palliative care, dermatology or for recommendations applying specifically to the elderly population), to 7.0% in pediatrics, 8.0% in infectious disease, and 10.9% in hematology (Supplementary File 2).

Evidence supporting the discordant recommendations

The comparator was clearly and explicitly stated in 73 (20.9%) of the 349 recommendations, not clearly but either obvious or relatively easy to infer in 230 (65.9%) and uncertain in 46 (13.2%). The direction of the recommendation was most often framed in favor of the intervention (78.5%) rather than against it (Table 4).

The full-text of the UpToDate topic often provided a rationale supporting the recommendation. An explicit statement on the balance of benefits and harms was present in 92 (26.4%), and an implicit statement in 157 (45.0%), and no statement in 100 (28.7%). Explicit statements addressing the relative importance of outcomes and/or on patients' values and preferences in making the trade-offs between alternatives were present in 10 (2.9%) of the recommendations; they could be inferred in 171 (49.0%), but not in the remaining 168 (48.1%) of discordant recommendations. Cost or resources considerations were mentioned in 15 (4.3%). The evidence cited to support each discordant recommendation varied substantially, with a median of 4 references cited, range from 0 to 33, with 45 (12.9%) of recommendations without any citation. Observational studies dominated (203, 58.2%); 49 (14.0%) were supported by a systematic review (Table 4).

Appropriateness of the discordant recommendations

Kappa for the initial taxonomic judgment regarding whether the recommendation was appropriate or inappropriate according to GRADE guidance was 0.46 (moderate agreement). The two reviewers required consensus discussions for 43% of the discordant recommendations. Third party adjudication to determine the appropriate classification was required in 12 of the discordant recommendations (3.4%).

Reviewers judged 204 (58.5%) of the 349 discordant recommendations to be consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations (Table 5). The most common paradigm was a "life-threatening or potentially catastrophical situation", followed by "potential similar benefits, one clearly less risky or costly", "potential catastrophic harm", "uncertain benefits, certain harm", and "established similar benefits, one potentially more risky or costly" (Table 5).

Reviewers judged 47 (13.5%) of the 349 discordant recommendations as "good practice statements"; 38 (10.9%) as a "misclassification of certainty (evidence warranted moderate or high certainty)"; and 60 (17.2%) as warranting a weak recommendation (see Table 5).

DISCUSSION

Among 9451 GRADE recommendations in UpToDate, about two thirds were formulated as weak recommendations and the remainder as strong recommendations. Of all recommendations, only 3.7% (n=349) were strong recommendations based on low certainty in effect estimates (Table 3). Of these discordant recommendations, over half were consistent with one of the five paradigmatic situations in which it is appropriate to offer discordant recommendations; approximately 14% represented "good practice statements"; approximately 11% were based on a misclassification of certainty (evidence warranted moderate or high certainty), and approximately 17% were judged to warrant a weak recommendation (Table 5). The proportion of appropriate discordant recommendations varied across intervention types or clinical fields (Supplementary File 2). Although most topics in UpToDate provided a rationale to support the discordant recommendation, 29% lacked statements about benefits and harms and 13% did not provide citations, which points at potential areas of improvement for UpToDate related to standards for trustworthy guidelines.[1]

Strengths and limitations

This study assessed the strength of recommendations in the largest known sample of recommendations developed using GRADE. Indeed, even large guidelines include a few hundred recommendations[18], whereas UpToDate topics have one of the largest known coverage in clinical fields and included 9451 recommendations at the time of this assessment.

The taxonomy that we used has been successfully implemented in two prior studies of clinical guidelines[10 11] (see below: relation to prior work). Our reviewers could all be

characterized as expert GRADE methodologists: they were clinical epidemiologists with an in-depth understanding of GRADE methodology acquired through use of GRADE in a large number of assessments over a period of years and were therefore well equipped to assess judgments on evidence and recommendations. This differs markedly from UpToDate authors (some with little understanding of GRADE) and UpToDate editors (all of whom have received basic GRADE training, but some little more than that). Despite the advanced skills of our reviewers, chance corrected kappa agreement on the appropriateness of recommendations was moderate (0.48).[19] Consensus discussions were needed for 43% of discordant recommendations, although formal adjudication by third parties was required for only 12 discordant recommendations (3.4%).

The necessity for frequent consensus discussions reflects the substantial judgment required in categorizing recommendations. This is in part due to the narrative nature of UpToDate topics, which does not include formal summary of finding tables or evidence profiles[17], often discussing the evidence and rationale for several recommendations in a free-text cross-referenced structure that sometimes omits statements regarding benefits and harms, and lacks citations. The one previous study using this taxonomy that addressed chance-corrected agreement reported a kappa of 0.68. The higher kappa may well be a result of more explicit reporting with use of summary of findings tables in the WHO guidelines that were the subject of investigation. The concern regarding the need for consensus discussions is perhaps increased because a single team using a single system of categorization undertook the study. A further limitation of our study is that decisions were based solely on information published in UpToDate, while authors of the topics may have considered other factors.[20]

Another element contributing to the challenges in making categorizations is the clarity of the comparison on which the recommendation applies. As in previous assessment in

guidelines[9], the comparator was clearly and explicitly stated in only 73 (20.9%) of discordant recommendations and was uncertain in 46 (13.2%). When comparators were not clear and explicit, reviewers' inferences may not always have been correct.[20]

Relation to previous work

Two prior studies provided a formal structured exploration of discordant recommendations using the GRADE approach. An assessment of 357 recommendations in 17 Endocrine Society Guidelines found that only 29% of discordant recommendations were consistent with one of the 5 paradigmatic situations.[11] A second study of 456 recommendations in 116 WHO guidelines using GRADE found that of 160 discordant recommendations, only 15.6% were judged consistent with GRADE guidance.[9 10]

Our results contrast with these previous two studies. First, the proportion of weak recommendations was approximately 30% higher in UpToDate than in WHO and Endocrine Society guidelines. This proportion was, however, similar to the 9th edition ACCP guideline on Antithrombotic Therapy and Prevention of Thrombosis, after it implemented GRADE.[18 21] Second, the proportion of inappropriate, discordant recommendation was considerably lower. Of the discordant recommendations, the proportion that should have been weak was about 17%, rather than 27% (Endocrine Society)[11] or 46% (WHO guidelines).[9]

A subsequent interview of panel members involved in the WHO guidelines highlighted reasons contributing to discordant recommendations. These included political considerations around long-established practices, the need for funding and policy formulation, or the fear of pushback from media.[20] Panel members also expressed skepticism regarding the value of making weak recommendations, or concerns they may

be ignored[20], although another study reported that WHO weak recommendations are frequently adopted in national policies (uptake of 61% for weak recommendations versus 82% for strong recommendations).[22] Finally, the authors identified both financial and intellectual conflicts of interest among panel members as an explanation for discordant recommendations.[20 23] Any or all of these factors may have contributed to UpToDate discordant recommendations.

Implications and conclusion

For users of UpToDate, our results are generally, though not absolutely, reassuring. The proportion of discordant recommendations is very small – only 3.7% of all recommendations. Furthermore, of the three categories inconsistent with GRADE guidance – good practice statement, misclassification of the certainty, and evidence warranting a weak recommendation (Table 2) – the third is by far the most problematic.[9] Good practice statements are appropriate when indirect evidence that is difficult to collect and summarize warrants high certainty in the impact of a given intervention and when the balance benefits and harms is large.[14] Thus, in terms of implications for clinical practice, good practice statements have the same force as strong recommendations. Similarly with misclassification of certainty: since the certainty is actually moderate or high, a strong recommendation is appropriate. Recommendations that should have been weak instead of strong provide inappropriate "just do it" guidance for clinical practice, although they are actually preference-sensitive and should thus warrant shared-decision making.[8] Of the 349 discordant recommendations in UpToDate, only 60 fall in the category of inappropriate strong recommendations.

Thus, clinicians using UpToDate can anticipate that they will be misleadingly instructed

to take a "just do it" rather than an "it depends" approach to clinical decision making in 0.6% (6 of 1,000) UpToDate recommendations.[15] This seems close to a threshold in which one might ignore the problem. Nevertheless, we would still encourage clinicians to be alert to the possibility of an inappropriate strong recommendation – in UpToDate or elsewhere – whenever the recommendation is based on low certainty evidence and authors fail to provide an explicit rationale corresponding to one of the categories in Table 1.

A likely explanation for UpToDate's success in avoiding inappropriate discordant recommendations is the training and feedback that their deputy editors receive. For organizations using GRADE, our results suggest the desirability of such training for those involved in formulating recommendations to optimize use of GRADE.

Finally our results highlight the need for authors of trustworthy recommendations or guidelines[1] to provide clear and explicit comparators, as well as transparent and systematic reports of the key ingredients of their rationale when moving from evidence to recommendation.[17 24 25] Future avenues for research should also look at optimal presentation formats of EBM textbooks and guidelines, to ensure clinicians actually understand both the rationale and potential implications of all recommendations for clinical practice.[8 26-29]

LIST OF ABBREVIATIONS

GRADE: Grading of Recommendations Assessment, Development and Evaluation

WHO: World Health Organization

ACCP: American College of Chest Physicians

COMPETING INTERESTS

TA, AK, IN, RBP, PEA, DR, POV, and GHG are active members of the GRADE working group.

DR, at the time the data on graded recommendations was extracted from UpToDate and until 2016, was an employee of UpToDate – he reports personal fees from UpToDate, outside the submitted work.

GHG contributes to the training in GRADE methods for UpToDate in-house deputy editors, for which he reports personal fees from UpToDate, outside the submitted work.

CONTRIBUTORS

Thomas Agoritsas (TA) and Gordon H. Guyatt (GHG) designed the study. David Rind (DR) provided the list of all recommendations and grading from UpToDate. Paul E. Alexander (PEA) helped structuring data abstraction. Thomas Agoritsas (TA), Arnaud Merglen (AM), Anja F. Heen (AFH), Annette Kristiansen (AK), Ignacio Neumann (IN), Juan P. Brito (JPB), Romina Brignardello-Petersen (RBP), and Per O. Vandvik (POV) reviewed the recommendations in duplicate and classified them according to GRADE taxonomy. Thomas Agoritsas (TA) and Gordon Guyatt (GG) wrote the first draft of the manuscript. All authors have read the manuscripts and made improvements of the content and wording.

FUNDING / ACKNOWLEDGMENTS

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

DATA SHARING STATEMENT

There were no additional unpublished data from this study.

REFERENCES

1. Institute of Medicine (US) Committee on Standards for Developing Trustworthy Clinical Practice Guidelines. . In: Graham R, Mancher M, Miller Wolman D, et al., eds. Clinical Practice Guidelines We Can Trust. Washington (DC), 2011.

2. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ 2008;**336**(7650):924-6.

3. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2011;**64**(4):383-94.

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. Andrews J, Guyatt G, Oxman AD, et al. GRADE guidelines: 15. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol 2013.

7. Balshem H, Helfand M, Schunemann HJ, et al. GRADE guidelines: 3. Rating the quality of evidence. J Clin Epidemiol 2011;**64**(4):401-6.

8. Agoritsas T, Heen AF, Brandt L, et al. Decision aids that really promote shared decision making: the pace quickens. BMJ 2015;**350**:g7624.

9. Alexander PE, Brito JP, Neumann I, et al. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. J Clin Epidemiol 2016;**72**:98-106.

10. Alexander PE, Bero L, Montori VM, et al. World Health Organization recommendations are often strong based on low confidence in effect estimates. J Clin Epidemiol 2014;**67**(6):629-34.

11. Brito JP, Domecq JP, Murad MH, et al. The Endocrine Society guidelines: when the confidence cart goes before the evidence horse. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3246-52.

12. Vigersky RA, Bhasin S, Martin KA. The Endocrine Society Clinical Practice Guidelines: a self-assessment. The Journal of clinical endocrinology and metabolism 2013;**98**(8):3174-7.

13. Neumann I, Santesso N, Akl EA, et al. A guide for health professionals to interpret and use recommendations in guidelines developed with the GRADE approach. J Clin Epidemiol 2016;**72**:45-55.

14. Guyatt GH, Schunemann HJ, Djulbegovic B, et al. Guideline panels should not GRADE good practice statements. J Clin Epidemiol 2014.

15. UpToDate, Waltham, MA. http://www.uptodate.com (Accessed on July 7th, 2017).

16. Agoritsas T, T Vandvik PO, Neumann I, Rochwerg B, Jaeschke R, Hayward R, Guyatt GH, McKibbon A. Chapter 5. Finding Current Best Evidence, in JAMA Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 3nd Edition, McGraw-Hill Medical, 2015.

17. Guyatt G, Oxman AD, Akl E, et al. GRADE guidelines 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol 2010.

18. Agoritsas T, Neumann I, Mendoza C, et al. Guideline conflict of interest management and methodology heavily impacts on the strength of recommendations: comparison between two iterations of the American College of Chest Physicians Antithrombotic Guidelines. J Clin Epidemiol 2017;**81**:141-43.

19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;**33**(1):159-74.

20. Alexander PE, Gionfriddo MR, Li SA, et al. A number of factors explain why WHO guideline developers make strong recommendations inconsistent with GRADE guidance. J Clin Epidemiol 2016;**70**:111-22.

21. Guyatt GH, Norris SL, Schulman S, et al. Methodology for the development of antithrombotic therapy and prevention of thrombosis guidelines: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest 2012;**141**(2 Suppl):53S-70S.

22. Nasser SM, Cooke G, Kranzer K, et al. Strength of recommendations in WHO guidelines using GRADE was associated with uptake in national policy. J Clin Epidemiol 2015;**68**(6):703-7.

23. Alexander PE, Li SA, Gionfriddo MR, et al. Senior GRADE methodologists encounter challenges as part of WHO guideline development panels: an inductive content analysis. J Clin Epidemiol 2016;**70**:123-8.

24. Guyatt GH, Oxman AD, Kunz R, et al. GRADE guidelines: 2. Framing the question and deciding on important outcomes. J Clin Epidemiol 2011;**64**(4):395-400.

25. Guyatt GH, Oxman AD, Kunz R, et al. Going from evidence to recommendations. BMJ 2008;**336**(7652):1049-51.

26. Vandvik PO, Brandt L, Alonso-Coello P, et al. Creating clinical practice guidelines we can trust, use, and share: a new era is imminent. Chest 2013;**144**(2):381-9.

27. Kristiansen A, Brandt L, Alonso-Coello P, et al. Development of a novel multilayered

presentation format for clinical practice guidelines. Chest 2014.

28. Treweek S, Oxman AD, Alderson P, et al. Developing and evaluating communication strategies to support informed decisions and practice based on evidence (DECIDE): protocol and preliminary results. Implement Sci 2013;**8**:6.

29. Siemieniuk RA, Agoritsas T, Macdonald H, et al. Introduction to BMJ Rapid Recommendations. BMJ 2016;**354**:i5191.

Table 1. Paradigmatic situations in which a strong recommendation may be warranted despite low or very low certainty in effect estimates *(appropriate strength, consistent with GRADE)*

Situation	Certainty in Estimates (Quality of Evidence)		Balance of Benefits and Harms	Values and Preferences	Resource Considerations	Recommendation	Example
	Benefits	Harms					
1. Life-threatening (or catastrophical) situation	Low or very low	Immaterial (very low to high)	Intervention may reduce mortality in a life-threatening situation; adverse events not prohibitive	A very high value is placed on an uncertain but potentially life-preserving benefit	Small incremental cost (or resource use) relative to the benefits justify the intervention	Strong recommendation in favor of the intervention	Indirect evidence from seasonal influenza suggests that patients with avian influenza may benefit from the use of oseltamivir (low certainty in effect estimates). Given the high mortality of the disease and the absence of effective alternatives, the WHO made a strong recommendation in favor of the use of oseltamivir rather than no treatment in patients with avian influenza.
2. Uncertain benefit, certain harm	Low or very low	High or moderate	Possible but uncertain benefit; substantial established harm	A much higher value is placed on the adverse events in which we are confident than in the benefit, which is uncertain	High incremental cost (or resource use) relative to the benefits may not justify the intervention	Strong recommendation against the intervention	In patients with idiopathic pulmonary fibrosis, treatment with azathioprine plus prednisone offers a possible but uncertain benefit in comparison with no treatment. The intervention, however, is associated with a substantial established harm. An international guideline made a recommendation against the combination of corticosteroids plus azathioprine in patients with idiopathic pulmonary fibrosis.
3. Potential equivalence, one option clearly less risky or costly	Low or very low	High or moderate	Magnitude of benefit apparently similar—though uncertain—for alternatives; we are confident less harm or cost for one of the competing alternatives	A high value is placed on the reduction in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation for less harmful/less expensive	Low-quality evidence suggests that initial Helicobacter pylori eradication in patients with early stage extranodal marginal zone (MALT) B-cell lymphoma results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy, but with high certainty of less harm, morbidity, and cost. Consequently, UpToDate made a strong recommendation in favor of H pylori eradication rather than radiotherapy in patients with MALT lymphoma.
4. High certainty in similar benefits, one option potentially more risky or costly	High or moderate	Low or very low	Established that magnitude of benefit is similar for alternative management strategies; best (though uncertain) estimate is that one alternative has appreciably greater harm	A high value is placed on avoiding the potential increase in harm	High incremental cost (or resource use) relative to the benefits may not justify one of the alternatives	Strong recommendation against the intervention with possible greater harm	In women requiring anticoagulation and planning conception or in pregnancy, high certainty estimates suggest similar effects of different anticoagulants. However, indirect evidence (low certainty in effect estimates) suggests potential harm to the unborn infant with oral direct thrombin (eg, dabigatran) and factor Xa inhibitors (eg, rivaroxaban, apixaban). The AT9 guidelines recommended against the use of such anticoagulants in women planning conception or in pregnancy.
5. Potential catastrophic harm	Immaterial (very low to high)	Low or very low	Potential important harm of the intervention, magnitude of benefit is variable	A high value is placed on avoiding potential increase in harm	High incremental cost (or resource use) relative to the benefits, may not justify the intervention	Strong recommendation against the intervention	In males with androgen deficiency, testosterone supplementation likely improves quality of life. Low- certainty evidence suggests that testosterone increases cancer spread in patients with prostate cancer. The US Endocrine Society made a recommendation against testosterone supplementation in patients with prostate cancer.

Reproduced and adapted from Neumann & al.[13]

Table 2. Reasons for issuing strong recommendation based on low certainty in effect estimates inconsistent with GRADE guidance

Situation	Example
Best practice recommendation (for which sensible alternatives do not exist)	"For patients with congenital adrenal hyperplasia, we recommend monitoring patients for signs of glucocorticoid excess, as well as for signs of inadequate androgen suppression." This statement should not have been GRADEed as sensible alternatives do not exist
The strong recommendation was warranted because the certainty of the evidence was actually moderate rather then low	"We recommend intensive lifestyle modification to the entire family and to the patient, and as the prerequisite for all overweight and obesity treatments for children and adolescents." The authors classified this as low quality evidence; our judgment is that the correct classification is moderate quality.
Lack of compelling explanation (the recommendation should have been weak)	"If a patient is unable or unwilling to undergo surgery, we recommend medical treatment with mineralocorticoids" Lack of evidence of mineralocorticoids being superior to other medical treatment (eg, anti-hypertensive medications)

Elements adapted from Brito et al.[11]

Table 3. Distribution of the strength of the recommendations in UpToDate according to the certainty in evidence

	Weak Recomendations	Strong Recommendations	All Recommendations
	N (%)	N (%)	N (%)
Low certainty	4335 (66.7%)	366 (12.4%)	4701 (49.7%)
Moderate certainty	2019 (31.1%)	1740 (59.0%)	3759 (39.8%)
High certainty	147 (2.3%)	844 (28.6%)	991 (10.5%)
Total	**6501** (68.8% of all rec)	**2950** (31.2% of all rec)	**9451** (100%)

Table 4. Characteristics of all 349 discordant recommendations in UpToDate, and proportion of appropriate discordant recommendations

	N	(%)	% of appropriate discordant *(p-value)*
Clinical Specialtes			*(p = 0.160)* *
Primary Care and General Internal Medicine	15	(4.3)	53.3
Emergency Medicine	16	(4.6)	81.3
Critical Care	5	(1.4)	80.0
Internal Medicine specialties	158	(45.3)	57.6
Oncology (including hemato-oncology)	43	(12.3)	55.8
Pediatrics	73	(20.9)	47.9
Obstetrics, Gynecology and Women Health	19	(5.4)	73.7
General Surgery	13	(3.7)	69.2
Anesthesiology	3	(0.9)	100.0
Psychiatry	4	(1.1)	75.0
Intervention type			*(p = 0.010)*
Drug intervention	197	(56.4)	61.4
Surgical interventions	69	(19.8)	59.4
Medical device	24	(6.9)	62.5
Behavioural or multi-disciplinary intervention	35	(10.0)	57.1
Diagnostic test, screening programms	24	(6.9)	29.2
Clarity of the comparator			*(p <0.001)*
Comparator not at all clear – uncertain	46	(13.2)	37.0
Comparator not clearly and explicitly stated or obvious, but relatively easy to infer	120	(34.4)	48.3
Comparator not clearly and explicitly stated, but obvious	110	(31.5)	68.2
Comparator clearly and explicitly stated	73	(20.9)	74.0
Type of comparator			*(p = 0.083)*
Too unclear	25	(7.2)	44.0
No intervention (or placebo)	126	(36.1)	54.0
Other intervention(s) (standard of care or alternative(s))	198	(56.7)	63.1
Direction of the recommendation			*(p <0.001)*
For the intervention (i.e. against the comparator)	274	(78.5)	51.1
Against the intervention (i.e. for the comparator)	75	(21.5)	85.3
Mortality			*(p <0.001)*
No statement about mortality	189	(54.2)	47.1
Implicit statement about mortality	47	(13.5)	68.1
Explicit statement about mortality	113	(32.4)	73.5
Balance of benefits and harms			*(p <0.001)*
No statement about the balance of outcomes	100	(28.7)	28.0
Implicit statement about the balance of outcomes	157	(45.0)	66.9
Explicit statement about the balance of outcomes	92	(26.4)	77.2

Relative Importance of outcomes - Values & Preferences			*(p <0.001)*
No statement about the relative importance of outcomes	168	(48.1)	42.9
Implicit statement about the relative importance of outcomes	171	(49.0)	73.1
Explicit statement about the relative importance of outcomes	10	(2.9)	70.0
Cost of resources			*(p = 0.023)*
No statement about cost or resources	334	(95.7)	57.2
Cost or resources clearly and explicitly stated	15	(4.3)	86.7
Supporting systematic review (SR)			*(p = 0.175)*
No SR is cited	300	(86.0)	56.3
SR of Observational studies	22	(6.3)	63.6
SR of both RCT and Observational studies	13	(3.7)	76.9
SR of Randomized Trials (RCT)	14	(4.0)	78.6
Design of primary studies			*(p = 0.002)*
No reference cited	45	(12.9)	35.6
Other type (eg narrative review, book chapter)	48	(13.8)	54.2
Observational studies	203	(58.2)	61.1
Randomized Trials (RCT)	53	(15.2)	71.7
Total	**349**	**(100)**	**58.5**

* The null hypothesis for the p-value is that the proportions do not differ across categories.

Table 5. Summary judgments on the appropriateness of 349 discordant strong recommendation based on low certainty in effect in UpToDate

	N	(%)
Appropriate discordant recommendations (consistent with GRADE)		
1. Life-threatening (or catastrophical) situation	70	(20.1)
2. Uncertain Benefit, Certain Harm	28	(8.0)
3. Potential similar benefits, One clearly less risky (or costly)	56	(16.0)
4. Established similar benefits, One potentially more risky (or costly)	18	(5.2)
5. Potential catastrophic harm	32	(9.2)
Total	**204**	**(58.5)**
Inappropriate discordant recommendations (inconsistent with GRADE)		
6. Good Practice Statement	47	(13.5)
7. Misclassification of certainty (judged moderate or high)	38	(10.9)
8. Lack of explanation, should have been weak recommendation (GRADE 2C)	60	(17.2)
Total	**145**	**(41.5)**

Supplementary File 1. INSTRUCTION FOR ABSTRACTION

Please read carefully before starting abstraction.
For any questions please contact me ASAP (thomas.agoritsas@gmail.com)

Background

- In a first phase of the project, we characterized the strength and confidence of the 9451 GRADE recommendations in UTD.
- In this last phase, we are focusing on
  - The **349 strong recommendations based on low confidence (GRADE 1C)**
  - Which are included in a total of **274 topics** (=chapters in UpToDate).
- The main objective is to categorize them based according the following taxonomy
  - Appropriate grading: recommendation consistent with one of the five paradigmatic situations defined the GRADE framework (see examples in Table 1 below):
    - *[App#1] Life-threatening situation*
    - *[App#2] Uncertain Benefit, Certain Harm*
    - *[App#3] Potential similar benefits, one clearly less risky (or costly)*
    - *[App#4] Established similar benefits, one potentially more risky (or costly)*
    - *[App#5] Potential catastrophic harm of one option*
  - Inappropriate grading: recommendation inconsistent with GRADE (see examples in Table 2 below), in short:
    - *[Inapp#1] Good Practice Statement*
    - *[Inapp#2] Misclassification of confidence (should have been GRADE 1B or 1A)*
    - *[Inapp#3] Lack of explanation, should have been a weak rec (GRADE 2C)*

→ **Before starting, please read the examples and Appendix Tables 1 & 2, they are also embedded in separate tabs in the abstraction excel file. Do not focus on memorizing them, as data abstraction will guide you in your judgment.**

Abstraction Excel File & Variables

- We will conduct the whole abstraction process in the attached standardized excel file
- We have kept the variables to abstract to the minimum necessary, most with **pre-defined response categories in drop-down menus** (and infrequently as free-text for copy-pasting).
- **Each recommendation has a separate row in the file**. There are sometimes more than GRADE 1C per UTD topic (the number are indicated).
- **As a guiding principle, keep in mind that the main objective is to assess the most appropriate taxonomy (first as appropriate or inappropriate, then subcategory).**
  These are the last variables in the file.
- This requires judgment based on what is reported in the UpToDate topic, mostly in narrative form. Indeed, there are typically no "summary of findings tables" of "evidence profiles" to explicit GRADE assessment. Absolute certainty in taxonomy is sometimes hard to achieve, but try and assign the best fit you can.
- The abstraction form is organized as follows to guide your final judgment:

UpToDate TOPIC & RECOMMENDATION	Pre-entered data to help you identify the relevant topic in the UpToDate topic in the dropbox
POPULATION	Check pre-entered clinical field, document age-group
INTERVENTION	Intervention type
COMPARATOR	Clarity and type of the comparator, direction of rec
OUTCOME – Benefits & Harms	Statement about: mortality, balance of benefits & harms, their relative importance (ie. values/preferences), cost
EVIDENCE supporting the rec	Date of literature review & updates number and type of supporting evidence Indication regarding the potential role of indirect evidence, Presence of large effects.
Conflict of interest (COI)	Copy paste statement, presence of financial COI
TAXONOMY – APPROPRIATE	**Separate judgement on each of the 5 paradigmatic situation defined by GRADE** (clear, possible, no)
TAXONOMY – INAPPROPRIATE	**Separate judgmenet on each of the 3 inappropriate situations**
CONFLICT RESOLUTION	**TAXONOMY DECISION** (for kappa), confidence in the decision (to document), assessing agreement and **RECONCILED TAXONOMY** within each pair of abstractor.
ADJUDICATION	Recording adjudication third reviewer if this was necessary

- **Specific guidance for each variable is found in the GREY BOXES on top of each column.**
- **Please read and select best option from DROP-DOWN menus within each cell**
- A few cells are for free-text to copy paste from the topic. Be sure to double click in the cell before pasting content, to keep the format intact.
- Most variables are followed by a column labelled **"additional comments"** or "rationale". These are for your personal notes to guide conflict resolution.
- A few examples already abstracted are shown in the first rows as an indication.

Abstraction: STEP-BY-STEP

Start abstracting a few first recommendations to get familiar with the process and contact me (thomas.agoritsas@gmail.com) for any question. I'm happy to have a quick skype if and as often as necessary.

- Go to the recommendation in the next row in the excel file.
- The [topic_#] and [topic_title] correspond to the name of the **PDF file** for the corresponding UTD topic → Open it.
- **Search automatically (ctrl-R or command-F) for "1C" → This will directly lead you to the GRADE 1C recommendation(s) at the end of the topic under the "Summary and Recommendations" section.**
- Read it carefully, and take a few seconds to try and get some first rough impression re: potential taxonomy
- **Then, find the corresponding paragraph in the topic that supports the recommendation** (it is often indicated soon after the recommendation (e.g. "See treatment…"). If not, try and find which paragraph(s) discuss(es) the recommendation.
- Abstract all variables in the order of the file as this will guide your formal judgment.
- Then judge each of the 5 appropriate and 3 inappropriate categories in the taxonomy.
- Then decide which one fits best and document the confidence you have in your assessment.
- Go to the next recommendation in the next row. (if this is in the same topic, you'll be able to copy several or your answers.

Conflict resolution and adjudication

- **You've been assigned with a paired reviewer. Schedule a first conflict resolution in the following days to ensure you are on the same page.**
- Record the agreed taxonomy in the specific column ("RECONCILED TAXONOMY"). Do NOT modify your initial judgement ("TAXONOMY DECISION") - as this will use to calculate kappa.
- **If you cannot resolve conflict, send us your questions for adjudication by a third reviewer.**
- Record adjudication in the final column.

➢ **Thanks again for your help. Do not hesitate to contact me for any questions (thomas.agoritsas@gmail.com)**

Supplementary File 2.

Characteristics of all 9451 recommendations in UpToDate: certainty effect estimates and clinical fields

	N	(%)	% of Strong Rec	% of strong rec discordant	% of any rec being discordant
Clinical Fields					
1 Primary Care & General Internal Medicine	356	(3.8)	22.5	16.3	3.7
2 Emergency Medicine	295	(3.1)	25.1	**23.0**	**5.8**
3 Critical Care	144	(1.5)	34.0	10.2	3.5
4 Cardiovascular Medicine	529	(5.6)	**42.7**	5.3	2.3
5 Infectious Diseases	870	(9.2)	**41.3**	**19.5**	**8.0**
6 Nephrology and Hypertension	475	(5.0)	**39.8**	6.3	2.5
7 Pulmonary Medicine	347	(3.7)	34.6	6.7	2.3
8 Hematology (non-oncology)	192	(2.0)	**41.1**	**26.6**	**10.9**
9 Neurology	395	(4.2)	31.1	**17.9**	**5.6**
10 Allergy and Immunology	261	(2.8)	23.4	3.3	0.8
11 Endocrinology & Diabetes	504	(5.3)	19.0	6.3	1.2
12 Gastroenterology & Hepatology	471	(5.0)	22.9	7.4	1.7
13 Rheumatology	216	(2.3)	21.3	4.3	0.9
14 Palliative Care	33	(0.3)	18.2	0.0	0.0
15 Oncology	1255	(13.3)	**36.0**	7.7	2.8
16 Hemato-oncology	263	(2.8)	27.8	**20.5**	**5.7**
17 Pediatrics	1057	(11.2)	**39.5**	**17.7**	**7.0**
18 Pediatric Emergency Medicine	126	(1.3)	28.6	5.6	1.6
19 Gynecology & Obstetrics	709	(7.5)	22.3	12.0	2.7
20 General Surgery	403	(4.3)	**32.8**	12.1	4.0
21 Anesthesiology	48	(0.5)	16.7	**37.5**	**6.3**
22 Dermatology	240	(2.5)	5.8	0.0	0.0
23 Psychiatry	262	(2.8)	16.8	9.1	1.5
TOTAL	**9451**	**(100)**	**31.2**	**12.4**	**3.9**