BMJ Open

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

Thomas Agoritsas,[1,2,3] Arnaud Merglen,[4] Anja Fog Heen,[5] Annette Kristiansen,[5] Ignacio Neumann,[3,6] Juan P Brito,[7] Romina Brignardello-Petersen,[3,8] Paul E Alexander,[3] David M Rind,[9] Per O Vandvik,[5,10] Gordon H Guyatt[3]

To cite: Agoritsas T, Merglen A, Heen AF, *et al*. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. *BMJ Open* 2017;**7**:e018593. doi:10.1136/bmjopen-2017-018593

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2017-018593).

Received 10 July 2017
Revised 24 September 2017
Accepted 26 September 2017

For numbered affiliations see end of article.

Correspondence to
Dr Thomas Agoritsas;
thomas.agoritsas@gmail.com

ABSTRACT

Introduction UpToDate is widely used by clinicians worldwide and includes more than 9400 recommendations that apply the Grading of Recommendations Assessment, Development and Evaluation (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 Analytical 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, 1740 (59.0%) on moderate certainty and 366 (12.4%) on low certainty. Of the 349 discordant recommendations 204 (58.5%) were judged appropriately (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) is 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 Grading of Recommendations Assessment, Development and Evaluation (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 Grading of Recommendations Assessment, Development and Evaluation (GRADE) system 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 five seldom-occurring paradigmatic situations in which a strong recommendation is warranted despite low certainty in the evidence (table 1).[6 13] Furthermore, 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 randomised 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 9400 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 characterise discordant recommendations based on the taxonomy described above (tables 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 summarising the evidence and rationale supporting the recommendations are mostly in narrative formats and do not provide summary of finding tables or evidence profiles.[3] To assess the appropriateness of discordant recommendations according to the paradigmatic situation defined in the GRADE framework, we therefore standardised data abstraction to collect relevant information from the main text (see detailed instruction in the online 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 (1) clearly and explicitly stated; (2) not clearly and explicitly stated, but obvious; (3) not clearly and explicitly stated or obvious but relatively easy to infer; (4) 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'

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 Grading of Recommendations Assessment, Development and Evaluation (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 favour 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 favour 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 favour 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 *et al*.[13]
MALT, mucosa-associated lymphoid tissue.

Table 2 Reasons for issuing strong recommendation based on low certainty in effect estimates inconsistent with Grading of Recommendations Assessment, Development and Evaluation (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 GRADEd 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 judgement 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, antihypertensive medications)."

Elements adapted from Brito *et al*.[11]

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 (randomised 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): (1) good practice statements; (2) a misclassification of the evidence—the evidence warranted moderate or high certainty rather than low or (3) 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 prespecified response categories whenever possible and exported in SPSS V.22.0 for analysis. We analysed 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 healthcare, including 16.1% in oncology, 49.2% topics in other internal medicine specialties or primary care and 12.5% in paediatrics. 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) (see online 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

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

	Weak recommendations	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)

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 healthcare issues within each clinical field, (see online 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 behavioural or multidisciplinary interventions (10.0%). These interventions were most often compared with 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% (eg, in palliative care, dermatology or for recommendations applying specifically to the elderly population) to 7.0% in paediatrics, 8.0% in infectious disease and 10.9% in haematology (see online 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 favour 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 four 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 judgement 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 (online 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,[17] 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 characterised as expert GRADE methodologists: they were clinical epidemiologists with an in-depth understanding of GRADE methodology acquired through the use of GRADE in a large number of assessments over a period of years and were therefore well equipped to assess judgements on evidence and recommendations. This differs markedly from UpToDate authors (some with

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

	n (%)	Per cent of appropriate discordant (P value)
Clinical Specialties		(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 haemato-oncology)	43(12.3)	55.8
Paediatrics	73(20.9)	47.9
Obstetrics, Gynaecology and Women Health	19(5.4)	73.7
General Surgery	13(3.7)	69.2
Anaesthesiology	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 multidisciplinary intervention	35(10.0)	57.1
Diagnostic test, screening programmes	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 (ie, against the comparator)	274(78.5)	51.1
Against the intervention (ie, 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 and 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

Continued

Table 4 Continued

	n (%)	Per cent of appropriate discordant (P value)
Cost or resources clearly and explicitly stated	15(4.3)	86.7
Supporting 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 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
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.
RCT, randomised controlled trials; SR, systematic review.

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).[18] Consensus discussions were needed for 43% of discordant recommendations, although formal adjudication by third parties was required for only 12 discordant recommendations (3.4%).

Table 5 Summary judgements 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)

GRADE, Grading of Recommendations Assessment, Development and Evaluation.

The necessity for frequent consensus discussions reflects the substantial judgement required in categorising 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,[3] 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 categorisation 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.[19]

Another element contributing to the challenges in making categorisations 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.[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 only 29% of

discordant recommendations were consistent with one of the five 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 ninth edition American College of Chest Physicians (ACCP) guideline on Antithrombotic Therapy and Prevention of Thrombosis, after it implemented GRADE.[17 20] 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.[19] Panel members also expressed scepticism regarding 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, the authors identified both financial and intellectual conflicts of interest among panel members as an explanation for discordant recommendations.[19 22] 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 summarise 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 1000) 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 organisations using GRADE, our results suggest the desirability of such training for those involved in formulating recommendations to optimise 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.[3 23 24] Future avenues for research should also look at optimal presentation formats of Evidence-Based Medicine textbooks and guidelines, to ensure clinicians actually understand both the rationale and potential implications of all recommendations for clinical practice.[8 25–28]

Author affiliations
[1]Division of General Internal Medicine, Department of Internal medicine, Rehabilitation and Geriatrics, University Hospitals of Geneva, Geneva, Switzerland
[2]Division of Clinical Epidemiology, University Hospitals of Geneva, Geneva, Switzerland
[3]Department of Health Research Methods, Evidence, and Impact, Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada
[4]Division of General Pediatrics, Faculty of Medicine, University Hospitals of Geneva, University of Geneva, Geneva, Geneva, Switzerland
[5]Department of Internal Medicine, Innlandet Hospital Trust-division Gjøvik, Gjøvik, Norway
[6]Department of Internal Medicine, Pontificia Universidad Catolica de Chile, Santiago, Chile
[7]Division of Endocrinology, Diabetes, Metabolism and Nutrition, Department of Medicine and Knowledge and Evaluation Research Unit, Mayo Clinic, Rochester, USA
[8]Faculty of Dentistry, University of Chile, Santiago, Chile
[9]Institute for Clinical and Economic Review, Boston, Massachusetts, USA
[10]Faculty of Medicine, Institute of Health and Society, University of Oslo, Oslo, Norway

Contributors TA and GHG designed the study. DMR provided the list of all recommendations and grading from UpToDate. PEA helped structuring data abstraction. TA, AM, AFH, AK, IN, JPB, RB-P, and POV reviewed there commendations in duplicate and classified them according to GRADE taxonomy. TA and GHG wrote the first draft of the manuscript. All authors have read the manuscript and made improvements of the content and wording.

Competing interests TA, AK, IN, RB-P, PEA, DMR, POV and GHG are active members of the GRADE working group. DMR, 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.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement There were no additional unpublished data from this study.

REFERENCES

1. Institute of Medicine (US) Committee. Standards for developing trustworthy clinical practice guidelines. In: Graham R, Mancher M, Miller Wolman D, eds. *Clinical practice guidelines we cantrust*. 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;336:924–6.
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;64:383–94.
4. Alonso-Coello P, Schünemann HJ, Moberg J, *et al*. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ* 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: 14. Going from evidence to recommendations: the significance and presentation of recommendations. *J Clin Epidemiol* 2013;66:719–25.
7. Balshem H, Helfand M, Schünemann HJ, *et al*. GRADE guidelines: 3. Rating the quality of evidence. *J Clin Epidemiol* 2011;64:401–6.
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;67:629–34.
11. Brito JP, Domecq JP, Murad MH, *et al*. The endocrine society guidelines: when the confidence cart goes before the evidence horse. *J Clin Endocrinol Metab* 2013;98:3246–52.
12. Vigersky RA, Bhasin S, Martin KA. The endocrine society clinical practice guidelines: a self-assessment. *J Clin Endocrinol Metab* 2013;98:3174–7.
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. Wolters Kluwer. *Smarter decisions. Better care*. Waltham, MA: UpToDate. http://www.uptodate.com (accessed 7 Jul 2017).
16. Agoritsas T, Vandvik PO T, Neumann I, *et al*. Chapter 5. Finding current best evidence, in JAMA users. *Guides to the medical literature: a manual for evidence-based clinical practice*. 3rd edn: McGraw-Hill Medical, 2015.
17. 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;81:141–3.
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–74.
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;141:53S–70.
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;68:703–7.
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;64:395–400.
24. Guyatt GH, Oxman AD, Kunz R, *et al*. Going from evidence to recommendations. *BMJ* 2008;336:1049–51.
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;144:381–9.
26. Kristiansen A, Brandt L, Alonso-Coello P, *et al*. Development of a novel, multilayered presentation format for clinical practice guidelines. *Chest* 2015;147:754–63.
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.