Suggestions for improving guideline utility and trustworthiness

Ian A Scott,1 Gordon H Guyatt2

Several underemphasised limitations of guidelines need proactive remediation in meeting the contemporary needs of clinicians

Clinical practice guideline (CPG) panels are expected to abide by standards that ensure their processes are multi-disciplinary, systematic and unbiased.1 Unfortunately, many CPGs fail to satisfy these standards. Only a third of 130 US guidelines produced by subspecialty societies between 2006 and 2011,2 satisfied more than 50% of standards set by the Institute of Medicine (IOM—see table 1),1 relating to panel composition, conflicts of interest, evidence synthesis, reconciliation of different interpretations of evidence and enumeration of treatment harms. Guidelines from other countries demonstrate similar deficiencies.3 Editorialists have identified the need for transparent measures of guideline trustworthiness;4 and some professional societies have issued rigorous standards for their guideline development panels.5 The fact that comparative studies have identified guidelines that more consistently meet most IOM standards6–7 suggests that it is possible for more guideline panels to improve the quality and rigour of their processes.

In an era when clinicians are increasingly using CPGs to inform their care and guidelines are being increasingly used as reference standards for clinical audits, pay for performance schemes, public scorecards and medical litigation, guidelines must be both rigorously developed and mindful of challenges in implementing their recommendations. In this article, we explore problematic issues that have received limited attention to date in guideline appraisal tools and commentaries.

Recommendations that conflict

A medical defence organisation in Australia recently warned doctors that conflicting guideline recommendations around prostate cancer screening using prostatic-specific antigen (PSA) testing may render them individually liable to claims of delayed diagnosis.8 In this case, CPG issued from the Royal Australasian College of General Practitioners9 stated that men aged 55–69 years should not be offered PSA testing routinely whereas CPG from the Urological Society of Australia and New Zealand10 stated they should. Similar concerns about medicolegal risk arise from conflicting guideline recommendations pertaining to breast cancer screening.11 In both cases, discordance is most likely related to conflicts of interest and differing specialty perspectives of panelists.12

Such discordance is not unusual. Even within specialties, divided opinions often exist: US and European cardiologists hold differing views about anticoagulant use in acute coronary syndromes,13 while national and international specialty groups differ in their diagnostic and management approach to diabetes14 and hypertension.15

When recommendations conflict, clinicians would benefit from explicit statements regarding how guideline authors summarised and interpreted the evidence, and what values and preferences they adopted in trading off desirable and undesirable outcomes—the two key sources of differing guidance. As disagreements in evidence interpretation may be legitimate, and because values and preferences may well differ across jurisdictions or patient groups, differing recommendations may, at times, be appropriate. Nevertheless, reducing the frequency of conflicting recommendations is desirable, and fostering greater transparency in the formulation of recommendations is clearly desirable. These goals would be facilitated if guideline groups could, at a national and international level, harmonise their methodologies and collaborate in producing evidence summaries for common use. Such partnership1 could be fostered under the auspices of multinational agencies such as the WHO or the Guidelines International Network.16

Advantages of such collaboration would arise not only from a greater likelihood of consistent interpretation, but also in efficiency: to have multiple groups searching the same literature and conducting time-consuming evaluation of that literature is hugely inefficient. Collaboration may also allow panels to review and, if satisfied, endorse each other’s recommendations,17 thus preventing further needless duplication in guideline development. At the same time, ‘globalising’ the evidence while localising the recommendations recognises the need to customise guidelines according to values and preferences of target populations or regional resource constraints or circumstances. Few guidelines will be applicable across countries of very different wealth. In the absence of demonstrable variation in characteristics and preferences of populations within a country, national borders may be the reasonable default localising boundary.

Recommendations that go beyond robust evidence

Examples exist of major guideline recommendations having to be substantively revised, even reversed, in response to an enlarging totality of evidence. In some instances, guideline panels, in formulating their original recommendations, may have been prematurely swayed by dramatic results reported in single randomised trials, many involving small samples and/or stopped early for apparent benefit.18 Such enthusiasm, possibly fanned in some instances by conflicts of interest and undue influence of industry, turned out to be misplaced when larger subsequent trials failed either to reproduce such effects or disclosed serious adverse outcomes. Examples include β-blockers as cardioprotective agents in patients undergoing non-cardiac surgery, intensive insulin therapy in critically ill patients and activated protein C infusions in...
Guideline panels need to be correspondingly conservative with their recommendations.

In other instances, guidelines have advocated increasingly tight control of blood glucose and blood pressure, often relying on logical inference, to extend thresholds beyond the available evidence, with later trials showing no benefits and increased hazards. Guideline panels need to be appropriately critical about existing evidence, and offer correspondingly conservative recommendations.

Recommendations that lag behind robust evidence
Guideline panels must establish procedures for regularly and systematically reviewing the evidence base and updating their recommendations in a timely manner in response to new compelling evidence. This helps prevent recently released guidelines simply ignoring such evidence, or outdated guidelines persisting long after the publication of new research.

As an example, nephrology guidelines released in 2006 and financially supported by manufacturers of erythropoietin recommended a target haemoglobin (Hb) level between 110 and 130 mmol/L in patients with end-stage renal disease undergoing dialysis. This recommendation flew in the face of trials published over the previous 5 years showing that Hb levels above 120 mmol/L were associated with increased risk of cardiovascular events and death.

US guidelines addressing stable ischaemic heart disease published in 2002 and not updated until 2012 present another example of failure to keep up to date. During the hiatus, clinicians faced recommendations that patients with stable angina and non-critical coronary artery disease be offered coronary revascularisation in addition to optimal medical therapy, despite the publication of large trials showing that optimal medical management and lifestyle modifications achieved similar outcomes of symptom control, morbidity and mortality.

Recommendations based on low-quality evidence
Guideline panels frequently confront topics for which high-quality evidence is unavailable—either because efforts to gather evidence were inadequate or because such evidence simply does not exist.

In regard to the former, systematic searching methods have evolved to ensure that all relevant published studies are available to authors of systematic reviews and, via them, guideline panels. Such methods are, however, only helpful when they are used. A recent review of Australian CPGs found that only 16% of them used systematic literature reviews as a basis for recommendations, running a risk of guidelines based on unrepresentative evidence.

With regard to the latter, recommendations have to be based on low or very low quality evidence if this is the best systematic reviews can uncover. Low-quality evidence—which runs a greater risk of subsequent reversal than evidence warranting greater confidence—underpins 85% of major CPG recommendations in cardiovascular medicine and more than 50% in infectious disease medicine.

Accordingly, recommendations based on low-quality evidence should rarely be used to generate quality metrics or legitimise established beliefs and practice styles. For instance, guidelines for systolic heart failure (SHF) have, for many years, recommended low salt diets in patients with moderate-to-severe SHF. A recent meta-analysis of randomised trials suggests, however, that such diets increase mortality. Previous intensive care guidelines for severe sepsis recommended use of either colloid solutions (starch or albumen) or crystalloids to treat sepsis-induced hypovolaemia and insulin therapy to achieve close to normoglycaemic blood sugar levels. Subsequent high-quality trials have revealed increased risk of death and need for renal replacement therapy from commonly used starch solutions, and...
raised the possibility of increased mortality from aggressive blood sugar control. A final example: a recent adequately powered trial showed no reduction in mortality from the use of intra-aortic balloons in cases of cardiogenic shock following acute myocardial infarction, despite previous recommendations based on observational data suggesting benefit in this highly lethal condition.

Guideline recommendations should always be accompanied by a systematically derived summary of best available evidence that rates evidence quality and links it with the strength of recommendations, ideally using the GRADE system. Where evidence quality is low, weak recommendations should mostly apply. The need to subject expert opinion to explicit, transparent consensus methods that minimise domination of opinion by one or a few panelists becomes even more paramount when evidence is low quality. Where panels identify recommendations for which they fail to reach consensus, they need to state the reasons why. In the rare situation in which evidence is of such low quality that, even after careful consideration of all factors that may bear on decision making, panelists regard any recommendation as being too speculative, they may refrain from making a recommendation, clearly stating the reasons why. Doing so however means that the potentially puzzled clinician is left without guidance.

Strong recommendations based on low-quality evidence carry the risks of encouraging uniform practice that may not be in patients’ best interest and inhibiting research that could clarify the magnitude of benefit and harm. In general, panelists should be very cautious in issuing strong recommendations (ie, those that apply to all or almost all patients) in the face of low-quality evidence—if one is not sure of the effects, it is difficult to be sure of the right course of action. Although strong recommendations may be justified in certain situations of low quality of evidence (see table 2), guideline users should, in general, regard strong recommendations based on low-quality evidence with circumspection.

**Recommendations that do not reconcile individualised estimates of benefits and harms**

Despite poor reporting of care-related harm in many clinical trials, guideline authors should, as far as possible, express the benefit-harm trade-offs in terms of absolute risk of patient-important events.

Recommendations for universal prophylaxis for venous thromboembolism (VTE) in hospitalised general medical patients overlook the fact that symptomatic VTE occurs in only 2/1000 untreated patients without risk factors, that prophylaxis has no impact on mortality, and that for every 1000 average-risk patients treated with prophylactic heparin, three episodes of pulmonary embolism are prevented at the cost of four major bleeding episodes.

Different guidelines also vary in the levels of absolute disease risk at which they recommend initiation of preventive treatments such as statins. In other instances, treatment is recommended solely on the basis of single risk factors—such as cholesterol or blood pressure levels—exceeding certain threshold values, in the absence of a multifactorial estimate of patients’ overall disease risk. Both approaches discourage prioritisation...

### Table 2  Situations for which currently available low-quality evidence is associated with strong recommendations that could change as new high-quality evidence emerges*

<table>
<thead>
<tr>
<th>Situation</th>
<th>Strong recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) When low-quality evidence suggests benefit in a life-threatening situation (evidence regarding harms can be low or high)</td>
<td>Fresh frozen plasma or vitamin K in a patient receiving warfarin with elevated INR and an intracranial bleed. Only low-quality evidence supports the benefits of limiting the extent of the bleeding</td>
</tr>
<tr>
<td>(2) When low-quality evidence suggests benefit and high-quality evidence suggests harm or a very high cost</td>
<td>Head-to-toe CT/MRI screening for cancer. Low-quality evidence of benefit of early detection but high-quality evidence of possible harm and/or high cost (strong recommendation against this strategy)</td>
</tr>
<tr>
<td>(3) When low-quality evidence suggests benefit equivalence of two alternatives, but high-quality evidence of less harm for one of the competing alternatives</td>
<td><em>Helicobacter pylori</em> eradication in patients with early stage gastric MALT lymphoma with <em>H pylori</em> positive. Low-quality evidence suggests that initial <em>H pylori</em> eradication results in similar rates of complete response in comparison with the alternatives of radiation therapy or gastrectomy; high-quality evidence suggests less harm/morbidity from <em>H pylori</em> eradication</td>
</tr>
<tr>
<td>(4) When high-quality evidence suggests benefit equivalence of two alternatives and low-quality evidence suggests harm in one alternative</td>
<td>Hypertension in women planning conception and in pregnancy. Strong recommendations for labetalol and nifedipine and strong recommendations against ACE inhibitors and angiotensin receptor blockers (ARB). All agents have high-quality evidence of equivalent beneficial outcomes, with low-quality evidence for greater adverse effects with ACE inhibitors and ARBs</td>
</tr>
<tr>
<td>(5) When high-quality evidence suggests modest benefits and low/very low-quality evidence suggests possibility of catastrophic harm</td>
<td>Testosterone in males with or at risk of prostate cancer. High-quality evidence for moderate benefits of testosterone treatment in men with symptomatic androgen deficiency to improve bone mineral density and muscle strength. Low-quality evidence for harm in patients with or at high risk of prostate cancer</td>
</tr>
</tbody>
</table>

*Adapted from reference 42.*

INR, international normalised ratio; MALT, mucosa-associated lymphoid tissue.
of treatment for patients at higher absolute risk which could potentially prevent more adverse events and at a lower cost within a given population. As a final note, recommendations for add-on incremental therapy that target multiple risk factors in the same individual may be inadvisable when very small benefits do not warrant exposure to associated harms and burdens.

Recommendations that focus on single diseases and ignore comorbidities

Most patients with a chronic disease have multiple comorbidities that single-disease guidelines do not address. Most CPGs do not consider competing risks and rarely consider applicability to individuals with limited life expectancy who are unlikely to benefit from long-term preventive treatments. Hardly ever do they address when to stop one or more of multiple chronic treatments, despite high-quality evidence that define circumstances where antihypertensive, hypoglycaemic and psychotropic medications can be safely discontinued.

Several strategies may render CPGs more useful in the context of multimorbidity: (1) cross-referencing guidelines dealing with other complaints commonly associated with the index condition (e.g., depression, pain, cognitive impairment and falls in patients with heart failure, most of whom are elderly); (2) emphasizing benefits (or harms) of disease-specific treatments with regard to other commonly co-occurring diseases that may be ‘concordant’ (diseases such as diabetes, hypertension or coronary artery disease which share a common management plan) or ‘discordant’ (diseases such as diabetes, asthma and depression in which management plans differ and may interact); (3) estimating the time at which slowly accruing treatment-related benefits outweigh immediate or constant rate harms, and if this time exceeds expected lifespan, recommending that treatments be discontinued or not initiated; and (4) liberalising and customising treatment targets (such as the desired levels of glucose and blood pressure control) according to age and homeostatic reserve.

Recommendations that are insensitive to patient preferences

Patients may value outcomes differently from guideline authors, including critically important outcomes such as death or serious morbid events. Many patients place as much, if not more, emphasis on avoiding treatment-related short-term toxicity, even if uncommon, than on primary effects in lowering future disease risk. Ideally, guideline developers should systematically review available evidence regarding risk perceptions and care preferences of their target populations and develop recommendations accordingly. Antithrombotic guidelines from the American College of Chest Physicians demonstrate this approach. Guideline panels should state explicitly the value and preferences structure underlying their recommendations, with statements attached to recommendations in which patient preferences are likely to be particularly salient to decision-making.

Guideline developers should promote informing and empowering patients to share in decision making. Strategies include guideline chapters that encourage clinicians to adopt a shared decision making approach, patient versions of guidelines, risk communication tools (graphs and pictograms), values clarification tools and complete patient decision aids and, perhaps most usefully, tools for use during time constrained patient-physician interactions tied to specific recommendations.

Recommendations that ignore implementation challenges

Limited resources, organisational and cultural barriers, and clinicians lacking skills to optimally implement recommendations constitute barriers to guideline uptake. Although guideline developers cannot deal with every contingency, they should ideally survey a representative sample of end-users, identify likely roadblocks to implementation and proffer potential solutions.

Resource implications of guideline recommendations, likely to be an increasing challenge to implementation in fiscally tight healthcare systems, were not explicitly considered in almost half of 30 US specialty society guidelines, and when they were, only half consistently used a formal method. Although no universally agreed method for incorporating economic analyses into CPGs currently exists, guideline panels should ideally consider cost-effectiveness when determining the direction and strength of their recommendations. They should also revise their recommendations if needed in response to formal economic evaluations of adherence to guideline recommendations and costs of implementation strategies in target populations.

Recommendations that are poorly responsive to a changing environment

Once CPGs are released, guideline panels should adopt procedures that allow recommendations and their methods of implementation to be revised, in a timely fashion, in response to user feedback and evaluations of guideline impact from qualitative research and clinical audit. Guidelines need to become living documents capable of rapid updating as important new evidence or suggestions for improving content and format emerge. Recently released cancer CPGs from the Cancer Council of Australia use an online ‘wiki’ format to allow readers—whether they are patients, carers or clinicians—to submit suggestions that a working group then considers.

Are we asking too much of guideline panels?

Guideline panels may feel that they have enough with which to contend without imposing the additional requests discussed in this article. They may bemoan the cost and effort of complying with onerous standards. However, the issues discussed above encompass the challenges that clinicians have to grapple with in everyday practice while caring for individual patients and for which clinicians look to clinical guidelines for assistance. Fortunately, various electronic support systems are being developed that may greatly assist panels in retrieving evidence and authoring actionable guidelines. Considerable time and resources are currently expended worldwide on guideline development. Fewer
but better resourced and rigorous guideline panels may be able to implement the entire range of strategies that maximise guideline trustworthiness.

Contributors IAS conceived the ideas, gathered data and wrote the draft manuscript; GHG assisted in gathering data, critically appraised the manuscript, and revised the text. Data were obtained from published research kept on file by IAS and supplemented by additional references supplied by GHG. IAS is a general physician and clinical epidemiologist with long-standing interest in guideline development and dissemination, has coauthored cardiology guidelines, and has published research articles pertaining to guideline methodology. GHG is a general internist and clinical epidemiologist who has published extensively on guideline methods, developed with others the GRADE clinical epidemiology, and is a lead methodologist for the American College of Chest Physicians guidelines on antithrombotic treatment.

Competing interests None.

References


52. Braithwaite RS, Fiellin D, Justice AC. The payoff time: a flexible framework to help clinicians decide when patients with comorbid disease are not likely to benefit from practice guidelines. *Med Care* 2009;47:610–17.


