Achieving health gain through clinical guidelines. I: Developing scientifically valid guidelines

Jeremy Grimshaw, Ian Russell

There is increasing interest in the role of clinical guidelines in encouraging good practice and thus generating health gain. However, there is professional ambivalence towards guidelines, stemming mainly from uncertainty about their effectiveness and how best to introduce them into clinical practice. Nevertheless, a systematic review of 59 rigorous evaluations of clinical guidelines reported improvements in the process of care in 55 studies and suggested that guidelines can change clinical practice if they are appropriately developed, disseminated, and implemented. Furthermore, published research has begun to identify effective techniques for introducing clinical guidelines.

Following the recent NHS reforms, purchasers are being encouraged to introduce guidelines clarifying the roles of primary and secondary care in patient management. Purchasers can either collaborate with providers in developing local guidelines or adapt published guidelines for local use. The second option is more attractive since development of a scientifically valid guideline is resource intensive and the necessary skills are often lacking at a local level. However, the quality of published guidelines is variable. In this paper we propose a classification to help purchasers and providers to identify scientifically valid guidelines. Elsewhere we propose a complementary classification to help both parties to ensure that guidelines change medical practice.

Guidelines, criteria, and standards
Quality assurance is beset with problems of terminology; terms used as synonyms include guidelines, algorithms, consensus statements, criteria, protocols, and standards. For example, the term standard is commonly used in medical audit to describe "performance the auditors have set themselves to achieve." However, it has also been used to describe "complex aggregations of criteria built up into a series of statements describing good clinical practice." Commenting on the terminology of quality assurance, Donabedian noted that "we have used these words in so many different ways that we no longer clearly understand each other when we say them" and suggested that we need to "clarify the existing nomenclature, encrusted and mishapen though it may be." Fortunately, the US Agency for Health Care Policy and Research (AHCPR) has recently agreed distinct definitions for "practice guidelines," "medical review criteria," and "standards of quality."
of (nationally derived guidelines) for local application." The distinction between national and local guidelines based on the degree of operational detail is useful; however, the term protocol may not be acceptable to all.

Guidelines can be deterministic or branching in structure. Deterministic guidelines comprise a fixed list of elements which should be followed, irrespective of the information available to the doctor. In branching guidelines, however, the recommended course of action at each stage depends critically on the available information. Though deterministic guidelines may be appropriate to define minimum levels of care, they ignore the deductive nature of medical decision making and are therefore less appropriate to guide routine clinical practice. Branching guidelines are commonly presented either as algorithms (or "flow charts") or as prose statements. The choice between these formats usually depends upon the topic and the target audience. In the North of England study, peer groups of general practitioners showed considerable diversity in the style of their internal guidelines. In contrast, the Harvard Community Health Plan has established a quality assurance programme based exclusively on algorithms, building on their successful use as a method of information transfer in educational settings. However, doctors are often reluctant to use algorithms in everyday practice because of their apparent complexity and lack of flexibility.

MEDICAL REVIEW CRITERIA

Medical review criteria are "systematically developed statements that can be used to assess the appropriateness of specific health care decisions, services, and outcomes." Such criteria may either be developed specifically to assess appropriateness of care (see, for example, Park et al) or be derived from clinical guidelines that were introduced for contracting or quality assurance; the latter can be selected either during the development of the guidelines or later by those monitoring contracts or coordinating quality assurance. These criteria may relate to all elements of care within a guideline or to a representative sample of elements. In general, it is desirable to define a set of criteria that cover the important aspects of the guidelines. We propose that such criteria should fulfil the requirements listed in the box and be based on mandatory or, at worst, near mandatory elements. In the management of suspected AMI, for example, a useful criterion for assessing performance after the introduction of guidelines is the percentage of patients without contraindications receiving aspirin within two hours of contacting a health professional.

STANDARDS OF QUALITY

Standards of quality are "authoritative statements of (a) minimum levels of acceptable performance or results, (b) excellent levels of performance or results, or (c) the range of acceptable performance or results." Irvine and Donaldson proposed alternative terms for the United Kingdom: ideal standards reflect "the best attainable under the very best conditions, where money and resources are unlimited"; optimal standards are "the best which conscientious practitioners can achieve under (normal) working conditions with the resources that are available to them"; and minimal standards are "those below which care should not be allowed to fall without causing the possibility of harm to patients." Standards relate to medical review criteria rather than to guidelines. Once criteria have been selected, a desired standard of performance can be defined. The standard chosen will depend on both the relative importance of the element from which the criterion is derived and the desired "level" of performance. The ideal standard may be interpreted as the estimated level beyond which iatrogenesis occurs, and the minimal standard as the estimated level below which avoidable mortality or major morbidity occurs. We believe, however, that it is the optimal standard at which both purchasers and providers can agree that this is best interpreted as the level at which the resources devoted to fulfilling the guideline give value (to patients) for money equal to that if they were devoted to any other NHS activity.

In the management of suspected AMI, for example, the standard for a mandatory element (for example, the percentage of patients without contraindications receiving aspirin within two hours of contacting a health professional) could be as high as 99% for ideal performance, as low as 75% for minimal performance, and between 85% and 95% for optimal performance. The use of a confidence-like interval for the optimal standard acknowledges the difficulty of

Requirements for medical review criteria

1 Each criterion should be easy to define*
2 Each criterion should relate to morbidity that is amenable to improvement by medical care*
3 There should be a sound scientific basis for discriminating between good and less than good performance as judged by each criterion*
4 The effects of non-medical factors on this performance should be adequately understood*
5 The criterion should yield data on enough patients for valid statistical analysis+
6 Together these criteria should span the range of morbidity covered by the guideline†
7 Together these criteria should span the range of skills required by the guideline†
8 Together these criteria should span the range of resources specified by the guideline†

*Derived from Kesselner et al
†Derived from Irvine
developing scientifically valid guidelines

Table 1 Desired attributes of clinical guidelines

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validity</td>
<td>Guidelines are valid if, when followed, they lead to the health gains and costs predicted for them</td>
</tr>
<tr>
<td>Reproducibility</td>
<td>Guidelines are reproducible if, given the same evidence and methods of guideline development, another guideline group produces essentially the same recommendations</td>
</tr>
<tr>
<td>Reliability</td>
<td>Guidelines are reliable if, given the same clinical circumstances, another health professional interprets and applies them in essentially the same way</td>
</tr>
<tr>
<td>Representative development</td>
<td>Guidelines should be developed by a process that entails participation by key affected groups</td>
</tr>
<tr>
<td>Clinical applicability</td>
<td>Guidelines should apply to patient populations defined in accordance with scientific evidence or best clinical judgement</td>
</tr>
<tr>
<td>Clinical flexibility</td>
<td>Guidelines should identify exceptions to their recommendations and indicate how patient preferences are to be incorporated in decision making</td>
</tr>
<tr>
<td>Clarity</td>
<td>Guidelines must use unambiguous language, precise definitions, and user friendly formats</td>
</tr>
<tr>
<td>Meticulous documentation</td>
<td>Guidelines must state when and how they are to be reviewed (under two separate circumstances – the identification or not of new scientific evidence or professional consensus)</td>
</tr>
</tbody>
</table>

*Adapted from Agency for Health Care Policy and Research.15*

estimating the value for money of a range of alternative uses of NHS resources given the current state of the knowledge base about health service effectiveness and efficiency. It is this knowledge base that the NHS research and development programme is designed to enhance.

Factors influencing the validity of guidelines

To maximise the potential health gain from guidelines it is important to ensure that they are rigorously developed and thus consistent with the available scientific evidence or, in the absence of such evidence, best clinical judgement. If those developing guidelines fail to overcome the many potential biases inherent in that development, the resulting guidelines may recommend ineffective or even dangerous clinical practice. In this paper, therefore, we propose a classification which links validity to the way in which guidelines are developed. However, this classification does not cover dissemination and implementation strategies as these are not intrinsic properties of a guideline despite their importance in determining whether guidelines achieve health gain.16

Table 1 summarises nine desirable attributes for clinical guidelines proposed by the AHCPR.15 Although only the first of these is labelled “validity,” our systematic review of clinical guidelines4 suggests that the next three attributes – reproducibility, reliability, and representative development – also affect whether or not guidelines lead to health gains when they are followed. (In other words, as others have noted in other contexts, validity, reproducibility, reliability, and representativeness are not independent attributes.)

While those developing guidelines should aim at achieving all nine attributes and focusing especially on the first four, we are conscious that these four are difficult to bring into operation, especially for purchasers and providers. Table 2 therefore identifies four practical questions that, published work suggests, have a direct, though variable, effect on validity: how evidence was identified and synthesised; how many users of guidelines and how many key disciplines were included in the guideline group; and how the guidelines were developed.

Even with the help of this table, it is difficult for potential users to assess published guidelines; contrary to the penultimate attribute of table 1 (meticulous documentation), few guidelines include details of their development. Of the 59 rigorous evaluations of guidelines in our review,4 none gave sufficient information to enable us to judge with confidence the rigour of their development. Fortunately, Hayward and colleagues have recently proposed structured abstracts for guidelines, which would encourage developers to publish details including methods used to gather, select, and synthesise evidence; methods used to assign values (relative importance) to the potential outcomes of alternative practice options; and validation (for example, results of any external review of the resulting guidelines, comparison with guidelines developed by other groups, and clinical testing of guidelines in practice).26

Methods of identifying and synthesising evidence

Methods of identifying and synthesising evidence include expert opinion, unsystematic literature review, ungraded systematic review, graded systematic review, and formal meta-analysis (table 2). In the United Kingdom many guidelines are currently developed by expert groups without formal literature reviews. This approach relies heavily upon the group’s knowledge of published work and their clinical experiences in non-evaluative

Table 2 Factors influencing the validity of guidelines

<table>
<thead>
<tr>
<th>Likelihood of scientific validity</th>
<th>Method of synthesising evidence</th>
<th>Composition of guideline group</th>
<th>Method of developing guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>Formal meta-analysis</td>
<td>Low (&quot;national external&quot; guideline group)</td>
<td>Evidence linked guideline development</td>
</tr>
<tr>
<td></td>
<td>Graded systematic review</td>
<td>Low (&quot;local external&quot; guideline group)</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Ungraded systematic review</td>
<td>Medium (&quot;intermediate&quot; guideline group)</td>
<td>Formal consensus development</td>
</tr>
<tr>
<td>Low</td>
<td>Expert opinion</td>
<td>High (&quot;internal&quot; guideline group)</td>
<td>Informal consensus development</td>
</tr>
</tbody>
</table>
conditions. However, clinicians' knowledge of published work is often incomplete, through poor presentation of research findings, lack of time to search for information, and dispersion of scattered results across a large number of journals. Furthermore, clinicians have difficulty in interpreting published evidence. The value of clinicians' experiences have also been questioned. Sackett identified four reasons why clinical experience may be biased and overestimate effectiveness: correlation between patient compliance and response to treatment; regression to the mean; the placebo effect; and the desire of both patients and clinicians for a successful response. Thus guidelines developed without a literature review may be biased towards reinforcing current practice rather than promoting evidence-based practice.

Traditional unsystematic literature reviews (without explicit research strategies, explicit inclusion criteria, or formal methods of synthesising evidence) overcome some of these problems by expanding the evidence base. However, important studies may be missed if the implicit search strategies are inappropriate or incomplete. Furthermore, the lack of explicit inclusion criteria can overwhelm the reviewer with evidence of variable quality, and the lack of formal methods of synthesis create difficulties in interpreting the evidence. Thus guidelines developed after an unsystematic literature review may suffer from selection bias while providing false reassurance.

Systematic reviews overcome these problems by defining search strategies and inclusion criteria – for example, for research designs. Ungraded systematic reviews use explicit standards to judge the scientific validity and clinical applicability of evidence and thus draw conclusions only from studies that are judged methodologically sound. For example, Haynes and colleagues reviewed the literature evaluating the effectiveness of continuing medical education. Before starting they defined an explicit search strategy and inclusion criteria specifying a minimum standard of methodological rigour. They identified 248 potential studies published between 1970 and 1983, of which only seven studies fulfilled their inclusion criteria. This allowed them to ensure that their conclusions were based on the best scientific evidence.

Ungraded systematic reviews typically review only randomised trials, but this approach may result in important evidence from other research designs being ignored. Sackett therefore described a method of grading evidence in which studies are ranked by design – for example, level I indicates a large randomised trial with low risk of error and level V a series of cases with no control data. This allows those developing guidelines to grade recommendations according to the strength of evidence – for example, a grade A recommendation is supported by at least one level I study.

When contradictory results are reported by trials addressing similar issues or several trials need to be combined formal meta-analysis is recommended. This combines the data or results of multiple trials and summarises all the reviewed evidence by a single statistic, typically a pooled relative risk of mortality (see for example Yusuf et al). Technologies successfully evaluated in this way include β blockade after myocardial infarction and prophylactic antibiotics in caesarean section; in contrast, technologies on which meta-analysis has thrown doubt include spinal manipulation for back and neck pain and lipid lowering drugs in low risk patients. Most meta-analyses consider only randomised trials and are thus open to the same criticism as ungraded systematic reviews. However, a generalisation of Sackett's proposal should in due course enable reviewers to derive pooled estimates from graded systematic reviews.

**COMPOSITION OF GUIDELINE DEVELOPMENT GROUP**

Both the validity and acceptability of guidelines depend on whether they are developed by an internal group comprising the clinicians who are going to use them, an intermediate group including representatives of those clinicians, or an external group excluding those clinicians. External guidelines may be produced either locally or nationally. Though guidelines developed internally are more likely to change medical practice, they are less likely to be scientifically valid through lack of technical skills in guideline development, clinical expertise, and resources.

The balance of disciplines within a guideline development group also has considerable influence on the recommendations. Leape and colleagues observed differences in recommendations for carotid endarterectomy between a surgical panel and a "balanced panel" comprising surgeons, neurologists, and other specialists: the surgical panel found more indications for surgery than the balanced panel. Scott and Black observed similar differences between surgical and mixed panels working on cholecystectomy. Lomas identified three reasons why guideline development groups should include representatives of all "stakeholders": firstly, the limited information available for guideline development needs to be supplemented by the interpretations of these stakeholders; secondly, legitimate conflicts over values need to be resolved; and thirdly, the successful introduction of a guideline requires that all key disciplines contribute to its development to ensure "ownership" and support.

**METHODS OF DEVELOPING GUIDELINES**

There are three main methods of developing guidelines: peer groups, Delphi techniques, and consensus conferences. However, all three suffer from biases and there is little evidence on their relative merits. Thus it is inappropriate to use the primary method of developing guidelines as a classification factor.

Woolf identified four secondary methods of development distinguished by the nature of decision making: informal consensus development, formal consensus development,
Developing scientifically valid guidelines

...it is free of complex analytical procedures and therefore relatively easy and fast. However, the guidelines are often of poor quality because the lack of explicit consensus method can lead to the group being dominated by forceful personalities or interests. Furthermore, guidelines produced by this method are often poorly documented, with little methodological information to reassure the user about scientific validity.

Formal consensus development methods (used by many consensus development conferences and Delphi groups) provides "greater structure to the analytical process" but fails to provide "an explicit linkage between recommendations and quality of evidence." Evidence-linked guideline development requires the explicit linkage of recommendations to the quality of the supporting evidence (see, for example, Woolf et al.48), and thus enhances the scientific validity of the guideline. In the absence of rigorous evidence it is even possible to include expert opinion within this approach if the resulting recommendations are clearly documented. Explicit guideline development (perhaps better described as "evaluative" guideline development) evaluates the benefits, risks, and costs of potential interventions by estimating the probability and value of each potential outcome and thus develops a "balance sheet" to facilitate patients' decisions about alternative treatments. While this approach is attractive in theory, its use in practice is so limited that it is not yet appropriate to include it with other methods of synthesizing evidence in table 2.

Conclusion
Within the United Kingdom the development and use of clinical guidelines is relatively new. As a result, few guidelines have been developed rigorously. If the introduction of guidelines into contracts is to generate health gain, however, purchasers and providers need to identify rigorously developed guidelines. In this paper we have proposed a classification of factors influencing the scientific validity of guidelines designed to inform choice about which guidelines to be integrated into contracts. Greater validity is likely to follow from the use of systematic literature reviews, independent (external) guideline development groups including representatives of all key disciplines, and the explicit links between recommendations and scientific evidence. Those developing future guidelines should also adopt these strategies whenever possible.

We thank the Chief Scientist Office of the Home and Health Department of the Scottish Office for funding the Health Services Research Unit; however the opinions expressed in this paper are those of the authors, not of the funding body.


42. Eddy DM. Practice policies: where do they come from? JAMA 1990;263:1265-75.

