Validity of the Agency for Healthcare Research and Quality Clinical Practice Guidelines
How Quickly Do Guidelines Become Outdated?

Paul G. Shekelle, MD, PhD
Eduardo Ortiz, MD, MPH
Shannon Rhodes, MFA
Sally C. Morton, PhD
Martin P. Eccles, MD
Jeremy M. Grimshaw, MD
Steven H. Woolf, MD, MPH

Context Practice guidelines need to be up-to-date to be useful to clinicians. No published methods are available for assessing whether existing practice guidelines are still valid, nor does any empirical information exist regarding how often such assessments need to be made.

Objectives To assess the current validity of 17 clinical practice guidelines published by the US Agency for Healthcare Research and Quality (AHRQ) that are still in circulation, and to use this information to estimate how quickly guidelines become obsolete.

Design, Setting, and Participants We developed criteria for defining when a guideline needs updating, mailed surveys to members of the original AHRQ guideline panels (n=170; response rate, 71%), and searched the literature for evidence through March 2000 (n=6994 titles yielding 173 articles plus 159 new guidelines on the same topics).

Main Outcome Measures Identification of new evidence calling for a major, minor, or no update of the 17 guidelines; survival analysis of the rate at which guidelines become outdated.

Results For 7 guidelines, new evidence and expert judgment indicated that a major update is required; 6 were found to be in need of a minor update; 3 were judged as still valid; and for 1 guideline, we could reach no conclusion. Survival analysis indicated that about half the guidelines were outdated in 5.8 years (95% confidence interval [CI], 5.0-6.6 years). The point at which no more than 90% of the guidelines were still valid was 3.6 years (95% CI, 2.6-4.6 years).

Conclusions More than three quarters of the AHRQ guidelines need updating. As a general rule, guidelines should be reassessed for validity every 3 years.
were still in use in 2000, when this study was undertaken. These guidelines cover a broad range of topics, including acute and chronic conditions, illnesses of children and adults, geriatric syndromes, and rehabilitation. The AHRQ guidelines were developed using a combination of systematic literature synthesis and multidisciplinary expert panels and were perceived to have advanced significantly the science of practice guideline development. These guidelines were considered to represent state-of-the-art management of the selected conditions.

Several years have elapsed since the AHRQ clinical practice guidelines were developed. At the request of and with support from the agency, we undertook a systematic assessment of the current validity of these guidelines. We used these data to provide empirical estimates of the rate at which guidelines become outdated.

**METHODS**

**Conceptual Model**

Our first task was to develop a conceptual model for assessing the validity of existing practice guidelines. No attempt to define the underlying criteria on which to base decisions about the current validity of guidelines has been published, nor does an operational method exist for doing so. A complete description of our conceptual model and its implications is described elsewhere.

In brief, we identified 6 situations that may require a guideline to be updated (or withdrawn), including changes in (1) the available interventions, (2) the evidence on the benefits and harms of existing interventions, (3) the outcomes that are considered important, (4) the evidence that current practice is optimal, (5) the values placed on outcomes, and (6) the resources available for health care.

In assessing the current validity of AHRQ guidelines, we focused on identifying when new information on interventions, outcomes, and performance justifies changing a guideline. Changes in the values placed on outcomes occur as societal norms change. Measuring these values and how they change over time is complex and is not dealt with herein, nor are changes in the availability of resources for health care or the costs of interventions, since policymakers in disparate health care systems consider different factors when deciding whether services remain affordable.

Identifying new information on interventions, outcomes, and performance that justifies changing a guideline involves 2 stages: identifying significant new evidence and assessing whether the new evidence warrants updating. Ideally, the most thorough way to identify significant new evidence would be to conduct a new systematic review. Our mandate, however, was to devise an approach that could be operationalized feasibly for larger numbers of guidelines. On that scale, conducting a systematic review for each guideline would be too costly and time consuming—it would be tantamount to completing the first step of updating, rather than determining whether updating was even necessary. The AHRQ currently spends approximately $250 000 to conduct a systematic review through its Evidence-Based Practice Center program; it would have therefore cost more than $4 million to perform one for each of the 17 guidelines that were still in use.

We therefore used a combination of a focused literature search and the guidance of experts from relevant disciplines as a more pragmatic way to help identify potentially significant new evidence. We reasoned that evidence sufficient to invalidate an existing national practice guideline would, in general, be of such a magnitude that it is known to experts in the field or would have been published in significant articles in major general interest or specialty medical journals. Our model for assessing the current validity of AHRQ guidelines is outlined in **FIGURE 1**.

We requested the assistance of each of the chairs of the original AHRQ clinical practice guideline expert panels in identifying the members of their panel whom they believed were most qualified to assess the current validity of each of the individual guideline statements within their respective clinical practice guidelines. Guideline statements were considered to be those set off in the document in bold, in a box, or as a bulleted point in the “recommendations for care” section. We used this information to assign individual guideline statements to these experts, who were then sent a survey that assessed the statements’ current validity. For each guideline statement, we asked 3 questions: (1) “Are you aware of new evidence or developments in the field relevant to this guideline statement?” (2) “Is the new evidence or development of sufficient importance to invalidate the guideline statement?” and (3) “Are there new guideline statements (within the boundaries of the original guideline) that should be present?” Respondents were told to consider validity within the context of our conceptual model (eg, changes in the available interventions, changes in the evidence on the benefits and harms of existing interventions). A second round of mailings was sent to nonrespondents. Respondents were given no financial incentives to reply. For each clinical practice guideline, we also attempted to obtain an evaluation of validity from at least 1 expert not associated with the original practice guideline panel.

We conducted limited literature searches for significant new evidence that may have an effect on the validity of the guideline statements. These limited literature searches were restricted, using the “document type” Medical Subject Heading terms to retrieve only review articles, editorials, and commentaries published through March 2000 on the particular guideline topic. Our reasoning for this restriction was that new evidence that is sufficient to change practice would frequently be accompanied by an editorial or commentary, making the latter a sentinel marker for new evidence. We further restricted the search to key journals, ie, those most likely to have published evidence of sufficient magnitude to warrant the revision of an existing national practice guideline. Key
journals included the 5 major general interest publications (Annals of Internal Medicine, British Medical Journal, JAMA, The Lancet, and New England Journal of Medicine) and key specialty journals for each topic identified by local experts in the field. For some guidelines, we did not identify key specialty journals and, therefore, conducted searches unrestricted by journal.

We defined the ideal starting point for each search as the end date for the original AHRQ guideline search. Only 5 (29%) of the 17 guidelines, however, identified this date. For the remainder, we conservatively chose the starting point to be 2 years prior to the publication date of the guideline; in the 5 guidelines for which we did have data, this was the lag period between the end of the search and publication. Two reviewers (P.G.S. and E.O.) reviewed the literature searches. Titles, abstracts, and articles were reviewed sequentially, seeking new evidence regarding the guideline statements. New evidence (principally randomized clinical trials) referenced in the review articles, editorials, or commentaries was retrieved and reviewed for relevance.

To assess the current validity of the AHRQ guideline statements, we also searched for practice guidelines that had been published after the release of the original AHRQ documents and that were related to the same conditions. We used the National Guideline Clearinghouse (http://www.guidelines.gov) and the Web sites of publishing organizations to identify relevant materials, downloading either the full text or the summary statements provided by the National Guideline Clearinghouse.

We identified additional evidence from the responses to our survey. For each guideline statement judged invalid for our survey, we reviewed all of the evidence for

![Figure 1. Model for Assessing the Current Validity of Guidelines](image-url)
VALIDITY OF AHRQ CLINICAL PRACTICE GUIDELINES

Each guideline. We considered the studies identified and the responses of the experts to the survey questions about validity and new guideline statements. We also placed greater emphasis on new evidence regarding the principal diagnostic or therapeutic procedures that have a major impact on outcomes (e.g., mortality). Based on these sources and our judgment, we assigned the guidelines into 1 of 3 categories.

Major Update Required. New evidence called into question 1 or more principal diagnostic or therapeutic recommendations or new evidence suggested the need for new principal diagnostic or therapeutic guideline recommendations. A major update of the guideline or its withdrawal from circulation is warranted.

Minor Update Required. The principal diagnostic or therapeutic recommendations were still valid, but new evidence supported changes to other recommendations or greater refinement of existing recommendations. A minor update of the guideline is warranted.

Still Valid. The recommendations remain valid; no update is warranted.

Next, we conducted a survival analysis of the AHRQ guidelines to estimate the rate at which guidelines become outdated. For each guideline, we determined a date of “birth”; that is, when work on the guideline was completed. We also determined whether the guideline had “died” by the time we reviewed it (April 2000). If so, we determined the date of “death,” and if not, we treated the observation as right-censored with a lifetime at least as long as the time from birth to April 2000.

For all 19 original guidelines, we defined birth as 12 months prior to the publication date of the guideline (which approximates the date when the completed guideline was delivered to the AHRQ). Three guidelines had dates of death corresponding to when they were withdrawn or had updates initiated by the AHRQ or the US Public Health Service (human immunodeficiency virus infection, smoking cessation, and urinary incontinence). For guidelines judged to require updating, we assumed conservatively that their deaths occurred in April 2000. This assumption is conservative because their actual lifetimes were of that length or less; hence, our lifetime estimates are biased positively.

To estimate the survival curve, we first fit a nonparametric Kaplan-Meier curve to this lifetime data set of censored and uncensored observations. We also fit a parametric model assuming that the survival function followed the Weibull distribution. Both of these methods take into account that some observations are censored in that the true lifetimes of 3 of the guidelines are unknown. This analysis was conducted using S-PLUS statistical analysis software.

RESULTS

Of the 19 clinical practice guidelines for which development was facilitated by the AHRQ, we assessed the 17 shown in Table 1. The urinary incontinence guideline underwent an update in 1996 to provide more specific recommendations; we consider only this 1996 update in our assessment of current validity. The AHRQ withdrew the human immunodeficiency virus guideline in October 1999 because the AHRQ judged it to be invalid. Similarly, the smoking cessation guideline recently underwent a Public Health Service–sponsored update.

We received replies from the panel chairs for 15 of the 17 clinical practice guidelines identifying whom among their respective panel members should be contacted to assess the current validity of their practice guideline statements. For 1 of the 2 practice guidelines for which we did not receive this information, we contacted the evidence-based practice center that had participated in the AHRQ-sponsored literature review that was used in the specialty society–sponsored guideline update process and asked the appropriate member of the task order team to complete the survey. For the remaining clinical practice guideline, 1 of the authors coincidentally had participated in the original guideline development process and decided which guideline statements to send to which experts.

We were unable to attain current contact information for 4 of the recommended experts, and in 2 instances, the panelists were deceased. We then sent out the individual guideline statements to the respective experts (n = 175). Five surveys were returned with the notation that the panelist was no longer at the given address and no forwarding address was known. Of the remaining 170 surveys, 121 (71%) were returned. Four panelists explicitly declined to participate in the survey. The number of responses varied by guideline. For all but 3 guidelines, more than 60% of the surveys were returned. We were able to obtain assessments of the validity of the practice guideline statement from 8 nonpanel experts for 7 of the guidelines. Among all questions about validity or the need for new recommendations across guidelines, there was complete agreement among respondents for 71% of questions, and only 5% of questions had more than 1 dissent.

Our focused literature searches identified 6994 titles. From these titles, 610 were selected for abstract review based on their potential for providing new information relating to the respective guideline areas. Of these, 173 full-text articles were identified and reviewed. In addition, 159 guidelines were also retrieved and reviewed. Experts identified an additional 156 references, so a total of 766 abstracts and 208 articles were reviewed.

Current Validity of AHRQ Guidelines

Table 1 lists our classification for each AHRQ guideline along with a succinct description of the key evidence on which we based our decision. The new evidence in Table 1 is not a complete recapitulation of all that was identified but shows the evidence that was most influential in making classifications. In all, we found that 7 guidelines needed a major update; 3 guidelines were still valid; and 6 guidelines needed a minor update. For 1 guideline (quality determinants of mammography), we reached no conclusion because we had an insufficient response rate to our survey and
were unable to adequately interpret the highly technical literature regarding possible improvements in mammographic imaging.

The Rate at Which Guidelines Become Outdated

**Figure 2** shows the Kaplan-Meier survival curve for the AHRQ guidelines. The solid line shows the estimated proportion of guidelines that survive at each point (i.e., that have lifetimes at least as long as the horizontal time in years), with dashed lines representing the 95% confidence interval (CI) bounds. The Weibull parametric model produced estimates similar to the nonparametric approach, and we used the former to construct estimates and 95% CIs for certain survival deciles. **Table 2** shows the Weibull model estimates for lifetimes of certain proportions (e.g., 50% of the guidelines will survive at least 5.8 years [95% CI, 5.0-6.6 years]). A sensitivity analysis using only guidelines judged as requiring a major update (death) yielded a 90% survival estimate of 5.5 years (95% CI, 4.5-6.5 years) and a 50% survival estimate of 7.1 years (95% CI, 6.4-7.8 years).

**Table 1. Evaluation of Agency for Healthcare Quality and Research (AHRQ) Clinical Practice Guidelines**

<table>
<thead>
<tr>
<th>AHRQ Guideline</th>
<th>Assessment of Important New Evidence</th>
<th>Judgment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute Pain Management: Operative or Medical Procedures and Trauma (February 1992)</td>
<td>New analgesics and routes of administration have superseded some recommendations</td>
<td>Minor update</td>
</tr>
<tr>
<td>Urinary Incontinence in Adults: Acute and Chronic Management (March 1996 update)</td>
<td>A new antimuscarinic drug for incontinence, tolterodine, is now available and is similar in efficacy to recommended drug (oxybutynin) but with less cholinergic adverse effects</td>
<td>Minor update</td>
</tr>
<tr>
<td>Pressure Ulcers in Adults: Prediction and Prevention (May 1992)</td>
<td></td>
<td>Still valid</td>
</tr>
<tr>
<td>Depression in Primary Care Volume 1: Detection and Diagnosis, and Volume 2: Treatment of Major Depression (April 1993)</td>
<td>Evidence exists to support expanded use of psychotherapy, cognitive therapy, and selective serotonin reuptake inhibitor pharmacotherapy in select patients, contradicting guideline recommendations</td>
<td>Major update</td>
</tr>
<tr>
<td>Cataract in Adults: Management of Functional Impairment (February 1993)</td>
<td>A key factor recommended for making the decision to operate, visual acuity, has been shown to not be a predictor of good outcomes Routine preoperative evaluation has been shown to be of no value, contradicting a guideline recommendation Significant changes in surgical techniques have occurred</td>
<td>Major update</td>
</tr>
<tr>
<td>Sickle Cell Disease: Screening, Diagnosis, Management, and Counseling in Newborns and Infants (April 1993)</td>
<td>Two specific therapies for sickle cell disease, hydroxyurea and bone marrow transplantation, have become available</td>
<td>Major update</td>
</tr>
<tr>
<td>Benign Prostatic Hyperplasia: Diagnosis and Treatment (February 1994)</td>
<td>One recommended therapy (balloon dilation) is now rarely used, and another (finasteride) has been shown in a randomized trial to be less effective than α-blockers</td>
<td>Major update</td>
</tr>
<tr>
<td>Management of Cancer Pain (March 1994)</td>
<td>New analgesics and routes of administration have supplanted some recommended approaches</td>
<td>Minor update</td>
</tr>
<tr>
<td>Unstable Angina: Diagnosis and Management (May 1994 [amended])</td>
<td>A new class of therapy agents, glycoprotein IIb/IIIa inhibitors, has demonstrated lifesaving benefits Low-molecular-weight heparin has been shown to be superior to unfractionated heparin for select patients</td>
<td>Major update</td>
</tr>
<tr>
<td>Heart Failure: Evaluation and Care of Patients With Left Ventricular Systolic Dysfunction (June 1994)</td>
<td>Several principal recommended therapeutics have been superseded by new evidence regarding β-blockers, spironolactone, and the emergence of angiotensin receptor antagonists as the preferred alternative for patients intolerant of angiotensin-converting enzyme There are expanded indications for intracardiac defibrillators</td>
<td>Major update</td>
</tr>
<tr>
<td>Otitis Media With Effusion (July 1994)</td>
<td>New data exist about the effects of otitis media with effusion on receptive and expressive language development, in turn affecting recommendations for timing of a principal therapeutic intervention, myringotomy tube placement</td>
<td>Major update</td>
</tr>
<tr>
<td>Quality Determinants of Mammography (October 1994)</td>
<td>Not enough information</td>
<td>Uncertain</td>
</tr>
<tr>
<td>Acute Low Back Problems in Adults (December 1994)</td>
<td>Some preventative and therapeutic statements need more refined estimates of effects, mostly tempering recommendations for limited effectiveness of back schools, lumbar corsets, and epidural steroid injections</td>
<td>Minor update</td>
</tr>
<tr>
<td>Treatment of Pressure Ulcers (December 1994)</td>
<td></td>
<td>Still valid</td>
</tr>
<tr>
<td>Poststroke Rehabilitation (May 1995)</td>
<td>More evidence exists for the benefit of stroke units in improving outcomes, probably sufficient to support a more specific statement with a stronger level of recommendation</td>
<td>Minor update</td>
</tr>
<tr>
<td>Cardiac Rehabilitation (October 1995)</td>
<td></td>
<td>Still valid</td>
</tr>
<tr>
<td>Recognition and Initial Assessment of Alzheimer’s Disease and Related Dementias (November 1996)</td>
<td>Clinical trial evidence exists of a modest benefit of cholinesterase inhibitors in select patients</td>
<td>Minor update</td>
</tr>
</tbody>
</table>
VALIDITY OF AHRQ CLINICAL PRACTICE GUIDELINES

Figure 2. Kaplan-Meier Survival Curve for AHRQ Clinical Practice Guidelines

The solid line represents the Kaplan-Meier curve for the Agency for Healthcare Research and Quality (AHRQ) guidelines. The dashed lines represent the 95% confidence interval.

Table 2. Estimated Time of Survival of Clinical Practice Guidelines

<table>
<thead>
<tr>
<th>% of Guidelines</th>
<th>Time (95% Confidence Interval), y</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>3.6 (2.6-4.6)</td>
</tr>
<tr>
<td>90</td>
<td>4.4 (3.5-5.3)</td>
</tr>
<tr>
<td>50</td>
<td>5.8 (5.5-6.6)</td>
</tr>
</tbody>
</table>

COMMENT

Of the 17 clinical practice guidelines developed under the auspices of the AHRQ between 1990 and 1996, we determined that more than three quarters need updating. This finding provides the first empirical evidence of the rate at which a collection of well-constructed clinical practice guidelines go out-of-date. Our survival analysis estimated that half of the guidelines became obsolete in 5.8 years. This estimate must be considered an upper bound since some of the guidelines in this cohort certainly became obsolete earlier than April 2000, the date of the judgment in our study. It is tempting to try to more precisely estimate the date when these guidelines became outdated by using the publication date of the new evidence used to reach the decision about validity. However, determinations of guideline validity require both evidence and judgment, and we did not collect judgments earlier than March 2000. We refrained from making assumptions about judgment at earlier points.

Our study has implications for the developers and users of clinical practice guidelines. Guidelines should be reviewed regularly to assess whether they are up-to-date. Given that our estimates of the lifetime of validity represent upper bounds, we believe that the best interval to recommend for this assessment is a conservative one—3 years. The lower bound of the 95% CI for the period when 90% of the guidelines were estimated to be still valid. A shorter interval may be indicated for topics noted for rapid scientific advances and a longer interval for fields that are more stable. Given the limited sample size in our study, we cannot provide estimates of survival stratified this way.

Our study also yielded important lessons about the practical methods to accomplish guideline review. Our combination of focused literature searches and expert opinion was a feasible way to review the validity of guidelines without expending considerable resources (approximately $100,000 to assess 17 guidelines). We consider our work to be only a first step, however, and further work is needed to better establish the optimal method of limited literature searches and how this should be integrated with expert judgment.

We found that we were able to identify relevant information from our limited literature searches in a more expeditious manner for topic areas with which we were clinically familiar. For topics with which we were unfamiliar, selecting and retrieving a larger proportion of irrelevant articles (lower specificity) became necessary to maximize sensitivity. Having the limited literature searches conducted by groups already familiar with the topic, such as relevant Cochrane Review groups or evidence-based practice centers, is likely to be more efficient.

Our search for new evidence would have been more efficient if we had been able to better target our search. Developers of guidelines could greatly ease the resources required for subsequent updating if future guidelines described the type of evidence that they anticipated would play a pivotal role in requiring revision of a guideline statement.

The relatively high response rate to our survey (without financial incentives) suggests that it might have been important to the members of the original guideline expert panels to keep their respective guidelines up-to-date. The lower response rate we achieved to our survey of experts who were not on the original guidelines panel indicates that more effort (e.g., financial incentives) may be needed to obtain an acceptable response rate from experts without this personal investment. Having periodic review by experts who were not involved with the initial guideline may help protect against perpetuating underlying bias introduced by the members of the original guideline panel, but it would come at increased cost.

All of these difficulties may be overcome by considering development of guidelines an ongoing process rather than a discrete event. Organizations developing guidelines should consider implementing guideline groups for a fixed period, with a rotating membership analogous to a National Institutes of Health study section. After guideline development, automated literature searching could be conducted at regular intervals (e.g., every 6-12 months) and relevant citations reviewed by subgroups of the panel. Relevant articles could then be distributed to the entire panel for consideration of the impact of new evidence on the validity of the existing guideline, using the criteria we have articulated. Areas of the guideline judged to be invalid could then be changed expeditiously if the guidelines were available in electronic form on the Web. In addition, Web-based availability would facilitate easy replacement of only those sections of the guidelines judged to be in need of change. Furthermore, this could be linked through list servers to electronically alert interested parties that a change has been made. In this way, guidelines would always be up-to-date and readily available.

Our study has several limitations. First, we assessed a relatively small number of existing clinical practice guidelines. The National Guideline Clearinghouse currently contains...
more than 1050. However, the guidelines we assessed represent the entire output of a high-profile program to develop national guidelines on a broad range of topics, which increases the generalizability of our estimates of shelf life. Second, these guidelines were not selected randomly; they came from 1 developer with a particular approach to guideline development. Guidelines that are less rigorously developed may become outdated at a different rate than what we report here. Third, we could not determine the actual date that guidelines became outdated, only that they had done so within a certain interval. This probably overestimates the useful life of these guidelines and is the reason we used the lower bound of the 95% CI to recommend when a review should be initiated. Fourth, our method of assessing the current validity of the guidelines has not been validated against the ideal gold standard: a simultaneous full update of the guidelines. Until such a comparison is undertaken, the methods we report here, which appear to have face validity, offer a practical way to make these determinations with limited resources.

In summary, we found that more than 75% of the AHRQ guidelines developed between 1990 and 1996 need updating. Half of these guidelines were certainly outdated in 5.8 years. To keep practice guidelines up-to-date, we recommend that as a general rule they be reviewed no later than 3 years after completion. This 3-year rule should be moved forward or backward in time if there is an expectation that the topic area of a particular guideline is evolving quickly or slowly.

Author Contributions: Study concept and design: Shekelle, Ortiz, Eccles, Grimshaw, Woolf. Acquisition of data: Shekelle, Ortiz, Rhodes. Analysis and interpretation of data: Shekelle, Ortiz, Morton, Eccles, Grimshaw, Woolf. Drafting of the manuscript: Shekelle, Ortiz. Critical revision of the manuscript for important intellectual content: Shekelle, Ortiz, Morton, Eccles, Grimshaw, Woolf. Statistical expertise: Morton. Obtained funding: Shekelle. Administrative, technical, or material support: Rhodes. Study supervision: Shekelle.

Funding/Support: This research was performed by the Southern California Evidence-Based Practice Center/RAND under contract with the AHRQ (contract 290-97-0001) and was supported by contract 290-97-0001, task 1000, level 1 from the AHRQ to the Southern California Evidence-Based Practice Center. Dr Shekelle was a senior research associate of the Veterans Affairs Health Services Research and Development Service.

Disclaimer: The authors of this article are responsible for its contents. No statement in this article should be construed as an official position of the AHRQ or the US Department of Health and Human Services.

REFERENCES