PART V. ANALYZING THE PRACTICES

Chapter 56. Methodology for Summarizing the Evidence for the Practices

Chapter 57. Practices Rated by Strength of Evidence

Chapter 58. Practices Rated by Research Priority

Chapter 59. Listing of All Practices, Categorical Ratings, and Comments

Chapter 56. Methodology for Summarizing the Evidence for the Practices

Background

The Agency for Healthcare Research and Quality (AHRQ) charged the UCSF-Stanford Evidence-based Practice Center with the task of rating or grading the patient safety practices identified and evaluated in this Report. The Report is an anthology of diverse and extensive patient safety practices, grouped by general topic (Part III, Sections A-H), and then further sorted within chapters by individual practices. Synthesis was challenging, but critical in order that readers and health care decision-makers could make judgments about practices to implement and/or research further. Keeping these two audiences in mind, we set 3 goals for "rating" the practices, as follows:

- Develop a framework for rating the main elements of practices within the constraints of the available literature, providing as much pragmatic information as possible to decision-makers who might endorse practices or fund further research in the area of patient safety interventions;
- Document the limitations of the rating method so that those "taking home" messages from the report understand the inherent limitations of making comparisons of a highly heterogeneous field of possible practices; and
- Convey the results of the ratings in an organized, visually appealing, accessible way that ensures that our cautionary notes regarding oversimplifying the ratings are clear.

Ultimately, we aimed to weight the practices, based on the evidence, on a range of dimensions, without implying any ability to calibrate a finely gradated scale for those practices in between. Proper metrics for these comparisons (eg, cost-effectiveness analysis) require more data than are currently available in the literature.

Data Inputs into the Practice Ratings

For each practice, information about various inputs into the final "roll-up", as we referred to our scoring of the practices, were prospectively determined. The decision about what information to attempt to gather was based on the potential expected uses of summary tables of practices. Three major categories of information were gathered to inform the rating exercise:

- *Potential Impact of the Practice*: based on prevalence and severity of the patient safety target, and current utilization of the practice
- *Strength of the Evidence Supporting the Practice*: including an assessment of the relative weight of the evidence, effect size, and need for vigilance to reduce any potential negative collateral effects of practice implementation
- *Implementation*: considering costs, logistical barriers, and policy issues

Further clarification of the 3 categories is in order. Authors were asked to report on *prevalence and severity of the safety target* for a given practice in order to categorize the potential impact of implementing the practice. We added to this an assessment of the practice's potential impact by reviewing evidence of its *current utilization.* If an intervention is already widely used, the room for improvement, stated in terms of additional reductions in adverse events targeted by the practice that could be achieved by wider implementation, is less than if few are currently using the practice. Thus, *potential impact of implementing the practice* is a

function of the prevalence and severity of the patient safety target (eg, medical error) and the current utilization of the practice.

Of course the actual impact of any practice is assessable only if factors related to *evidence supporting the practice* are evaluated. Since the Report represents an assemblage of the evidence for patient safety practices, the instructions to authors outlined the detailed data elements related to study design and outcomes we required them to abstract from the relevant studies (see Chapter 3). This information was used to assess, in general terms, the *overall strength of the studies* for each practice. *Effectiveness* is commonly defined as the net positive effect in routine practice. Frequently the data reported in the studies related to *efficacy*, usually the net positive effect under controlled, experimental situations. The translation from efficacy to effectiveness is not straightforward if no direct evidence is available, and is therefore based on judgments about the generalizibility of the specific research studies conducted. Also of key importance, and therefore abstracted from studies for use in the ratings, was the *effect size of the intervention*.

Finally, evidence-based reviews consider the *potential for harm* from a medical intervention, and authors were asked to report on any relevant evidence, as well as reasoned concerns, gleaned from the literature or from common knowledge about a practice.

To address the real-world environment and the desire by the public for action in the area of patient safety, practice chapters were designed to include information about cost and other potential barriers to *implementation*. While authors sometimes discussed cost savings or reported cost-effectiveness analyses, the focus was on the *start-up costs and annual outlays* for ongoing use of the practice. Although initial and ongoing costs are a function of local environments (eg, size of the health care network or institution), possible cost savings are likely to be even more subject to local conditions (eg, prevalence of the patient safety target). For major investment decisions, an assessment of tradeoffs is more appropriate at the local level. Our intention was simply to report "ballpark" estimates of initial and recurring costs. Separate from economic consequence of a particular practice implementation are the *political and technical considerations*.

For all of these data inputs into the practice ratings, the primary goal was to find the best available evidence from publications and other sources. Because the literature has not been previously organized with concurrent considerations of each of these areas, most estimates could be improved with further research and some are informed by only general and somewhat speculative knowledge. Where possible, in the summaries of these elements, we have attempted to highlight assessments made on the basis of limited data.

Rating Process

The 4-person Editorial Team developed a rating form that captured the patient safety target, practice description, and general rating categories (eg, High, Medium, Low) for some of the elements described in the section above. General heuristics were specified for each category, although individual judgment for ratings was designed into the process. The form also specified comment areas to allow raters to document their specific judgments and concerns about ratings. Each chapter was independently rated by each Editor as to the practices for which there was evidence. The Editorial Team convened for 3 days to compare scores, discuss disparities, and come to consensus about ratings–both by category and summary ratings–or the reviewed practices.

Details about Decision Rules and Judgment Considerations

Potential Impact Factor

As noted above, an assessment of potential impact considered the prevalence and severity of the patient safety target, and the current utilization of the practice being evaluated. The Editorial Team used the data from the chapters and clinical knowledge to order the potential impact as "High," "Medium," "Low," or "Insufficient Information." To qualify for the "High" score, a practice had to target a patient population of greater than 1% of hospitalized patients (about 300,000 patients per year) or target a patient safety problem that can result in death or disability. The "Low" score was used for target populations of less than 0.01% of hospitalized patients (about 3000 patient/year) who might experience reversible adverse effects if an effective practice were not available. Potential impact was deemed a "Medium" if the practice had a patient safety target that fell between the 2 other categories.

An additional decision rule was applied to the Impact rating after the initial assessment based on prevalence and severity was made. If a practice was currently widely used (>75% of hospitals), then the rating was demoted one notch (ie, from High to Medium or Medium to Low). When this situation occurred, a notation identified that the potential impact level was impacted by its high current utilization.

We reserved the "Insufficient Information" category for those cases where the prevalence and severity information was quite limited or where the patient safety target was ill-defined.

Evidence Supporting the Practice

Study strength, effect size on target(s), and need for vigilance due to potential harms were rated based more on judgment than pre-specified decision rules. In each case, raters documented their reasons for category choices.

For *study strength*, the level of study design and outcomes (see Chapter 3 for hierarchies), number of studies, numbers of patients in studies, generalizability, and other methodologic issues were specified as factors to consider in weighting the relative study strength for a particular practice. Study strength could be categorized as "High," "Medium," or "Low." The actual findings of the studies were not considered when scoring study strength because this information was captured in the assessment of effect size on target. If there was minimal or no evidence about a practice, the study strength rating was "Low" and raters did not score the remaining 2 elements of the evidence supporting the practice since that might give undue "credit" to the findings.

The assessment of *effect size on target(s)* was based on the relative risk reductions or odds ratios reported in the reviewed studies for evidence of effectiveness. The raters only used the findings reported in the practice chapters, and did not perform additional analyses (eg, meta-analysis). If all studies or, in cases where there were a large number of studies, the vast majority showed a positive and appreciable effect size (ie, greater than 15% relative risk reduction), then the positive effect size was categorized as "Robust." If there was clearly no effect or a very minimal effect (ie, less than 5% relative risk reduction), then the positive effect size was rated as "Negligible." For findings that were considered suggestive of substantive effect, but not clearly "Robust," the category used was "Modest." The final category, "Unclear," captured those practices for which the effect size results were inconsistent.

For any given practice that reduces one adverse event, it is conceivable that new problems might ensue when the practice is implemented. Thus, we subjectively rated the

concern for harm based on the *level of vigilance* necessary to ensure that the practice, if implemented, would not result in collateral negative effects. The categories available were "Low," "Medium," and "High." Thus, a practice rated as "Low" would require little to no attentiveness to potential harms, while one rated as "High" would merit heightened monitoring for potential negative effects. These ratings were made conservatively, meaning that when in doubt, a higher vigilance category was selected.

Implementation

Assuming a 3-year lead time for implementation, patient safety practices were rated for their costs and complexity. *Costs* were based on initial start-up and annual expenditures for full implementation at an average size hospital or health care organization. Potential cost savings were not considered for the rating, but were reported in the practice chapters if they were documented in the literature. If a practice was expected to require expenditures of greater than about $1 million, the rating was "High." Expenditures of approximately $100,000-$1 million were categorized as "Medium." Below this level, practices were rated as "Low" in terms of cost.

The feasibility of implementation was rated by considering potential political (eg, major shifts in who delivers care) and technical (eg, integration of legacy and newer computer systems) obstacles. Because relatively few data exist for rating implementation complexity, we used only 2 categories, "Low" and "High," meaning relatively easy and relatively difficult. In cases in which implementation could be accomplished simply with the expenditure of dollars, we gave high cost scores but low feasibility scores.

Overall Rating for Impact/Evidence

In addition, each member of the team considered the totality of information on potential impact and evidence supporting the practice to score each on a 0 to 10 scale ("Strength of the Evidence"). For these ratings, we took the perspective of a leader of a large health care enterprise (eg, a hospital or integrated delivery system) and asked the question, "If you wanted to improve patient safety at your institution over the next 3 years and resources were not a significant consideration, how would you grade this practice?" For this rating, we explicitly did *not* consider difficulty or cost of implementation in the rating. Rather, the rating simply reflected the strength of the evidence regarding the effectiveness of the practice and the probable impact of its implementation on reducing adverse events related to health care exposure. If the patient safety target was rated as "High" impact and there was compelling evidence (ie, "High" relative study strength) that a particular practice could significantly reduce (eg, "Robust" effect size) the negative consequences (eg, hospital-acquired infections), raters were likely to score the practice close to 10. If the studies were less convincing, the effect size was less robust, or there was a need for a "Medium" or "High" degree of vigilance because of potential harms, then the rating would be lower.

Overall Rating for Research Priority

Analogously, we also rated the usefulness of conducting more research on each practice, emphasizing whether there appeared to be questions that a research program might have a reasonable chance of addressing successfully ("Research Priority"). Here, our "thought question" was, "If you were the leader of a large agency or foundation committed to improving patient safety, and were considering allocating funds to promote additional research, how would you grade this practice?" If there was a simple gap in the evidence that could be addressed by a research study or if the practice was multifaceted and implementation could be eased by

determining the specific elements that were effective, then the research priority was high. If the area was one of high potential impact (ie, large number of patients at risk for morbid or mortal adverse events) and a practice had been inadequately researched, then it also would also receive a relatively high rating for research need. Practices might receive low research scores if they held little promise (eg, relatively few patients affected by the safety problem addressed by the practice *or* a significant body of knowledge already demonstrating the practice's lack of utility). Conversely, a practice that was clearly effective, low cost and easy to implement would not require further research and would also receive low research scores.

Caveats to Ratings

For all elements assessed, divergent assessments among the 4 Editor-raters were infrequent and were discussed until consensus was reached. For each final category where differences in interpretation existed and persisted after discussion, the protocol was to document a comment about these differences (see Chapter 59). Comments were also noted when specific additional information could clarify concerns about fidelity of a specific rating. In a few cases, categories that had not been specified were created for unusual circumstances and again comments to explain the category were documented.

Rating Tables

Information Captured

Ratings were recorded on a data table, and comments were footnoted. Summarizing from the previous discussion, the various data tables appearing in Chapters 57-59 captured some or all of the following 11 elements and rating categories, ordered from strongest to weakest:

1. Chapter number
2. Patient safety target(s)
3. Patient safety practice description
4. Potential Impact: High, Medium, Low, Insufficient Information
5. Study Strength: High, Medium, Low
6. Effect Size: Robust, Modest, Negligible, Unclear
7. Vigilance: Low, Medium, High
8. Implementation cost: Low, Medium, High
9. Implementation complexity (political, technical): Low, High
10. Overall rating for impact/evidence: 0 to 10 (10 is highest), in 0.5 increments
11. Overall rating for research need: 0 to 10 (10 is highest), in 0.5 increments

Information Reported

Chapter 59 presents the detailed data tables with categorization of elements 1-9 above. Reporting specific scores for the overall ratings would imply a refinement in scoring that was neither attempted nor advisable given the nature of the information available. Where applicable, caveats to the categorizations are appended in a series of endnotes.

In rating both the strength of the evidence and the research priority, our purpose was not to report precise 1-10 scores, but *to develop general "zones" or practice groupings*. As noted earlier, better methods are available for comparative ratings *when the data inputs are available*. The relative paucity of the evidence dissuaded us from using a more precise, sophisticated, but ultimately unfeasible, approach.

Chapter 57 summarizes the overall ratings for the “Strength of the Evidence" regarding their impact and effectiveness score, and subdivides the practices into 5 zones. Practices are listed from highest score to lowest score for each rating zone. The zones are “greatest strength” (score of 8-10), “high strength” (score of 6-7.5), “medium strength” (4-5.5), “lower impact/evidence scored practices” (score of 2-3.5), “lowest impact/evidence scored practices” (score of 0-1.5). Practices near the bottom of one zone may be just as appropriate to list near the top of the adjacent lower zone. Similarly, practices at the top of a zone may actually be more comparable to those in the adjacent higher zone. The cut-offs between zones are somewhat artificial, but allow a general and reasonable synthesis of the data on impact and evidence supporting (or negating) the effects of the practice. Readers can be confident that practices that fall in the highest zone do not belong in the lowest zone.

Chapter 58 summarizes the overall ratings for the “Research Priority” score, and provides examples of types of research that may be helpful. Practices are categorized in 3 zones: “Further Research Likely to be *Highly* Beneficial” (scores of 7 and higher), “Further Research Likely to be Beneficial” (scores of 4 to 6.5 inclusive), and “Low Priority for Research” (below 4). The chapter lists practices for each of the top two categories of research priority.