

A Framework For Evidence-Adaptive Quality Assessment That Unifies Guideline-Based And Performance-Indicator Approaches

Aneel Advani, MD, MPH,^{a,b} Mary Goldstein, MD, PhD,^{a,b} Mark A. Musen, MD, PhD^a

^aStanford Medical Informatics, Stanford University School of Medicine, Stanford, California

^bDepartment of Medicine, VA Palo Alto Health Care System, Palo Alto, California

Automated quality assessment of clinician actions and patient outcomes is a central problem in guideline- or standards-based medical care. In this paper we describe a unified model representation and algorithm for evidence-adaptive quality assessment scoring that can: (1) use both complex case-specific guidelines and single-step population-wide performance-indicators as quality measures; (2) score adherence consistently with quantitative population-based medical utilities of the quality measures where available; and (3) give worst-case and best-case scores for variations based on (a) uncertain knowledge of the best practice, (b) guideline customization to an individual patient or particular population, (c) physician practice style variation, or (d) imperfect reliability of the quality measure. Our solution uses fuzzy measure-theoretic scoring to handle the uncertain knowledge about best-practices and the ambiguity from practice variation. We show results of applying our method to retrospective data from a guideline project to improve the quality of hypertension care.

Introduction

Clinical guidelines are increasingly being used as tools to improve the quality of medical care.¹ A recent *JAMIA* White Paper has pointed to the importance of using *evidence-adaptive* clinical decision support systems that can incorporate new evidence-based guideline knowledge as it becomes known.² In our work on the MedCritic system, we have shown that complex, multi-step case-specific medical guidelines can be used for automated quality assessment to produce consistent quantitative *adherence scores*.³ However, our previous quality assessment algorithm was based on a multi-criteria aggregation method that (1) did not unify population-wide quantitative performance measures with case-specific guideline knowledge, and (2) did not behave gracefully as new evidence about the guideline, population, or physician became known.

In this paper, we extend our work on the MedCritic system to satisfy the following three requirements that are needed for *evidence-adaptive* quality assessments usable at the population level. First, we must have a unified modeling approach that incorporates both guideline-based quality measures and validated single-step performance or outcome indicators. Second, we need to score adherence consistently with quantitative population-based medical utilities where

the evidence for absolute or relative utilities of the quality measures is available. It is often the case, however, that such evidence is unavailable in guideline documents.⁴ Thus, our quality assessment method must be adaptive to changes in *uncertainty in knowledge* or *ignorance* about the evidence base for a guideline recommendation or quality standard.

The third requirement stems from the related observation that less than 50% of daily clinical practice is supported by validated evidence that could be the basis for evidence-based guideline standards.¹ Therefore, most automated quality assessment measures that are currently used, such as the Health Plan Employer Data and Information Set (HEDIS) quality indicator standards and benchmarks,⁵ are chosen to maximize the *internal validity* or *reliability* of the quality indicators.⁶ But daily practice includes cases outside of these precise point-measures. Thus evidence-based quality assessment of daily care implies ambiguity in adherence scores due to limits in the *external validity* of evidence-based guidelines and their *customization* for individual patients.⁷ Our third requirement is therefore the need to extend our previous guideline adherence scoring method to allow for *uncertainty in belief* or *ambiguity* about the quality of physician behavior under these variances.

In our model and algorithm, uncertainty in adherence scores may arise from measured variances due to: (a) uncertain knowledge of, or variation in the strength of evidence for, the best practice;⁴ (b) guideline customization to an individual patient or particular population; (c) physician practice style variation; or (d) chance variation because of small numbers of patients in the physician's profile or other aggregation unit for the audit.⁸ This method uses interval-based *fuzzy measures* to keep track of the worst-case (giving maximum penalty for variances) and best-case scores (giving minimum penalty for variations) for adherence to guidelines.

Methods

Unified Modeling Approach. The MedCritic system for automated quality assessment scores adherence to hierarchical sets of quality-indicator *constraints* derived from guidelines. The system has been designed to work within the EON⁹ or Asgaard¹⁰ architectures for guideline-based care, and is applied to retrospective analysis of data from the ATHENA clinical decision support system for hypertension.¹¹ Since we have previously described our

language, QUIL (Quality Indicator Language) for modeling and executing queries for guideline-based quality indicators,³ we only briefly describe QUIL here.

QUIL is used to define a set of related quality indicators as individual nodes in a hierarchical guideline-based *quality constraint structure* (see Figure 1). The higher-level indicators in the hierarchy can be considered higher-level *intentions* of the guideline¹² while lower-level indicators may be more specific processes or adjusted outcomes. Individual performance criteria or outcome measures can be embedded as nodes in the guideline-based quality constraint structure. For instance, the “Reduce Co-morbidities” outcome node can be considered a high-level intention of the guideline. Although we can still conceptually assign it a utility based on reductions in mortality from the disease-specific customizations of care in the child nodes, we cannot measure this outcome *directly*. On the other hand, the “Rx β -Blocker” process node is actually a performance measure used for a quality improvement intervention hypertension care at our regional VA integrated service network.

QUIL queries consist of *goal* and *enabling* temporal constraints, preserving the form of most population rate-based performance measures. Satisfaction of the goal constraint defines a clinical execution of the medical guideline that satisfies the quality indicator, given the appropriate context defined by the enabling constraint. For example, the “Rx β -Blocker” says that the process constraint

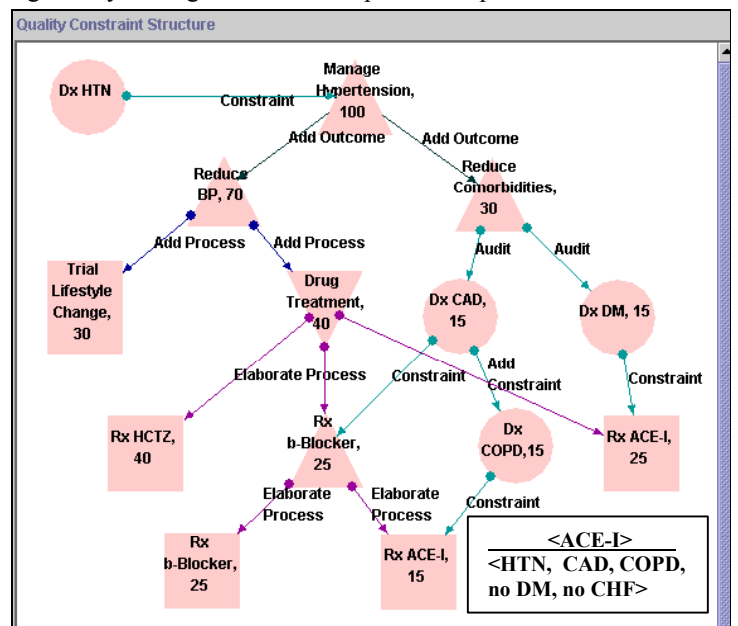
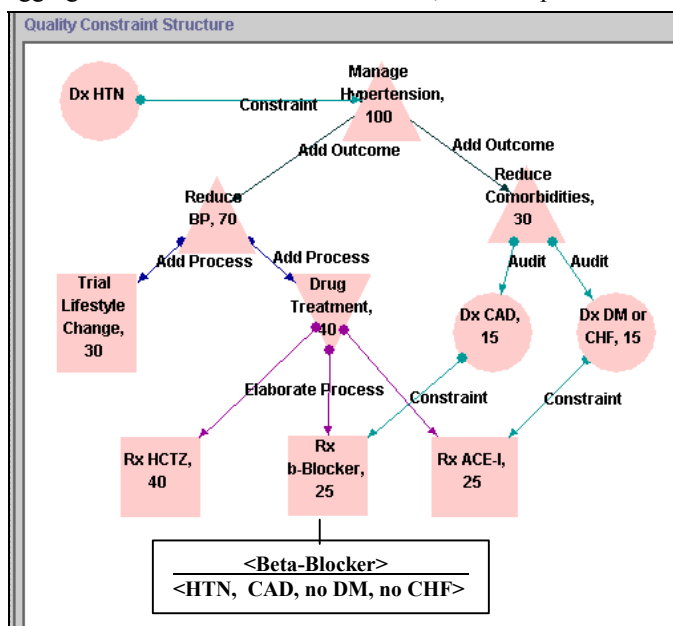
is satisfied if β -Blockers were prescribed in those patients with diagnoses of hypertension and coronary artery disease, but no diagnoses of diabetes or congestive heart failure. In the case of performance measures, these queries would be executed for all patients in a population and the adherence scores would reflect those of the entire population aggregate.

Thus we see that our first design requirement is solved because we can hierarchically combine a case-specific guideline with performance measures that are embedded within it. This structure and, as we shall see below, the propagation of scores upwards, is well-defined with respect to further customizations or adjustments in the process or outcome indicators. We model these customizations as *elaborations* of the graph structure. Thus, in the right-hand structure in Figure 1 the “Rx β -Blocker” node has been elaborated into a more specific set of alternatives, adjusted for the presence or absence of a diagnosis of chronic obstructive pulmonary disease (COPD) in the patient.

Representing Quantitative Evidence under Uncertainty.

To solve our second requirement, we add utility weights to the nodes in our constraint structure. These weights can correspond to real-world utilities to patients of satisfying the part of the guideline represented by the given node. The utilities then give us the *maximum adherence score* for concordance with that quality indicator. For instance, we may have absolute utilities in quality-adjusted life-years saved (QALYs) for the outcome measure of “Reduce BP”.

Figure 1. QUIL Model of Quality Constraint Structure for Hypertension Guideline. The screenshots show the *quality constraint structure* for a hypertension guideline. Each leaf node contains a QUIL query that represents the quality indicator for that part of the guideline. A schematic of the query for the “Rx β -Blocker” node is shown. The nodes have logical relationships with their children, with downward-pointing triangles as OR nodes, and upward-pointing triangles as AND nodes. The structure to the right is an *elaboration* of the structure to the left, where the β -Blocker treatment indicator has been *customized* to patients without COPD. The numbers in the nodes refer to the utilities to the patient of satisfying the quality measure. Parents receive utility as evidence-based aggregates of the utilities of child nodes, with the particular function given by the logical relationship between parent and child.



However, we may only have relative utilities from relative risk reduction information for different alternative process constraints. For instance, the nodes for “Rx β -Blocker” and “Rx ACE-I” may only be relatively comparable to one another in the context of a diagnosis of coronary artery disease (CAD). So we could assign these nodes with utility weights that consistently reflect both their relative risk reductions in QALYs end-points and the fact that these drugs do lower the blood pressure and hence provide an absolute benefit in utility as well.

The design third requirement relates to representing uncertainty due to variance in our knowledge or belief. In the first case, we need to model variations in the strength of evidence regarding the utility of a node. Because of the relative paucity of detailed evidence for every part of a guideline, we may not be *confident* of our utility assignment. This form of uncertainty is different from *risk*, in that we are not trying to take an expectation between alternative utilities but rather we are unsure of the probabilities of each utility value.

It has been shown that this form of uncertainty can be represented by interval-based *non-additive* probability measures,¹³ also called *capacities* or *fuzzy measures*. The measure defines two levels of subjective probability or *belief* in an outcome, which we denote as *[worst-case, best-case]*, where we give one utility weight at each end of a real interval. Note, however, that in terms of our expectation of a utility, we are completely indifferent for values within the interval. In fact, in general, our expected value for this outcome can be calculated as any convex combination:

$$[1] EU = \gamma(\text{best} - \text{case}) + (1 - \gamma)(\text{worst} - \text{case})$$

where $\gamma \in [0,1]$ refers to our *degree of confidence* or *aversion to uncertainty* in our belief. This is a formal way to model variations in adherence scores due to each of the cases (a)-(d) we outlined for the third design requirement. Thus we can take γ to be a function of each of these sources of variation:

$$[2] \gamma = f(\text{strength evidence, reliability of quality measure, guideline customization, practice variation, number of patients in an aggregation unit})$$

For example, the guideline may reflect an uncertainty about the best practice for the drug treatment of hypertension (see the node “Drug Treatment” in Figure 1). Thus the utility for a node, and hence the maximum adherence scores, may in fact be the interval [25,40], since we are indifferent as to which drug prescription is appropriate (“Drug Treatment” is an “OR node”). Of course, if we adjusted the guideline, by including utilities for the covariate morbidities that allow us to select which guideline step is appropriate, then we might be more confident about how to make our decision. Thus, in our example, all the child nodes in the disjunction under “Drug Treatment” would be scored at 40 reducing our

interval to [40,40]. Thus, with complete confidence in how to select alternatives, that is, with complete knowledge about appropriate guideline customization, our measure simply reduces to an additive point-probability. This probability then gives us a well-defined expected utility and hence a scalar adherence score for our quality measure. We have therefore outlined an *evidence-adaptive* measure that can adapt to increasing evidence-based knowledge in our decision support system by converging to an expected utility.

However, in many cases, we may not have absolute confidence in the appropriateness of the guideline. As we discussed above, we may have valid alternatives that may be followed, or as we shall see below, there may be reasons that the quality measure is not perfectly reliable. In these situations, we can also use an interval-based belief measure, and use an appropriate γ -value to calculate our expectation.

Principles for Aggregating Single-Indicator Assessments

One additional aspect of the problem remains to be solved, however. That is the method by which we can propagate adherence scores upwards from the queries present in the leaf nodes. (The temporal constraints in higher-level nodes are inherited downwards so the leaf nodes contain all the constraint information from their ancestors in the structure.) In our previous work, we used a generalized mean function, the L_p norm, given by,

$$[3] \|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}, 1 \leq p \leq \infty, \text{ where } x = \text{parent's}$$

score in terms of those adherence scores, x_i , of the child nodes, and a value of $p=1$ for AND nodes (for a sum of the child scores) and of $p \rightarrow +\infty$ for OR nodes (for the max of the child scores). In our new approach, each score for a leaf node is now the fraction of patients in the given population satisfying the quality measure, as opposed to just an all-or-none score. We also extend the previous work to allow propagation of intervals, for which we now let $-\infty < p < +\infty$.

In this case we keep track separately for the aggregation of the worst-case and best-case extremes. For an OR node, the best-case or most optimistic case is given by assuming that we get the maximum score from the possible child scores no matter which alternative we actually expect. In the worst case, the OR node parent will expect only the least valuable node to be satisfied. The OR node takes an interval-based aggregation with *[worst-case, best-case]* defined by the norms $[L(p \rightarrow -\infty), L(p \rightarrow +\infty)] = [\min, \max]$. For OR nodes whose children also have interval-based scores, the $[\min, \max]$ are respectively calculated from the $[\text{childrens' intervals' worst-cases, childrens' intervals' best-cases}]$. For an AND node, the score is given by only giving credit for the utilities actually measured for the child nodes. However, if there is any element of “OR”-ness either from the children’s intervals’ min-max differences or from “siblings” which are not yet known if the current choices are not exhaustive of the

alternatives. Thus, the AND node takes the following form $[worst-case, best-case] = [L(p \rightarrow -\infty) + L(p=1), L(p \rightarrow +\infty) + L(p=1)] = [\min + \text{weighted sum}, \max + \text{weighted sum}]$. We have chosen this mathematical aggregation operator because the method of preserving extreme points on the aggregation at each level can be proved as correctly preserving the best- and worst-case value throughout the aggregation. We omit the proof here.

Results

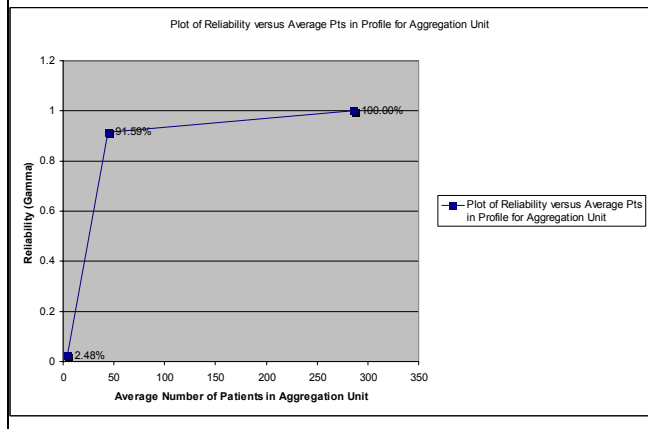
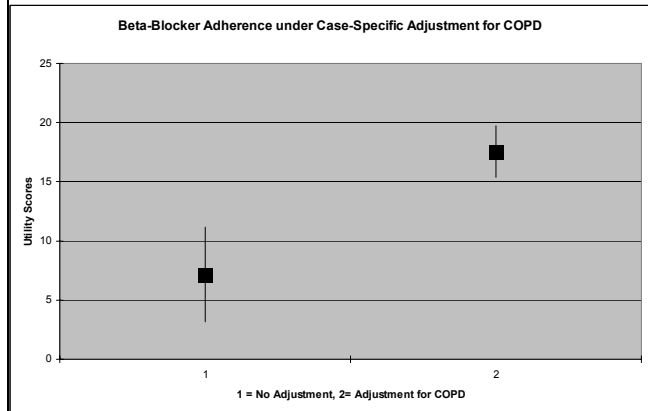
Now that we have outlined our method, we present the results of applying our scoring method to a dataset comprising records of hypertension care for approximately 1000 patients in the seven clinic divisions of the Palo Alto VA Health Care System, California. We present the results of two analyses of interval-based scoring relating to the performance measure represented by the node “Rx β -Blockers” in the quality structure above (see Figure 2). First, we show what the intervals for adherence scores are for the node with and without customization of the node for the presence of COPD as a comorbidity with its own utility (for which we have used a value of 10). Secondly, we show the effect on the degree of confidence (γ) of changing the aggregation unit for measuring adherence to the measure.

The first graph shows the result of measuring the adherence to the higher-level “Rx β -Blocker” in the hypertension guideline. We plot the worst-case and best-case ends of our adherence interval along with the means for the “Rx β -Blocker” aggregated over all patients in the region. The first series is the done without adjusting for consideration of any case-specific modification for clinician adherence to the performance measure. The second series includes the adjustment for clinician’s concomitant consideration of a *relative contraindication* to β -Blocker use (the presence of COPD) in the patient. The variance (difference between upper and lower bound) decreases by factor of 1.42 as the adjustment is carried out. Moreover, we also gain in utility, since we have labeled the “Dx COPD” node as transferring utility to the “Reduce Comorbidities” node as well. Thus, as we get more confident of our adjustment, we increase our γ by a factor of $1.42/(1.42-1) = 3.38$ in comparison with any older value.

The second graph shows the results for adherence to the “Rx β -Blocker” quality measure for various levels of aggregation of the analysis. In this case, we run the QUIL queries over (1) each physician, (2) each clinic division, and (3) the entire Bay Area VA Region. To show the variance with aggregation level, we have plotted the reliability against the number of patients within each aggregation unit. Note that we can take the reliability as a direct measure of the *minimum* γ value that we must keep arising solely from the chance variation due to the imperfect reliability of the measure. In our case, the reliability of the measure is a function of the number of patients in physician or clinics

Figure 2 Results of Case-Specific and Aggregation-Level Adjustment for Beta-Blocker Treatment Quality Measure

The first graph shows that both the validity (mean) and reliability (variance) of the β -Blocker performance measure increase with adjustment for specific case mix due to COPD. The second graph shows that the coefficient of reliability depends on the aggregation level of the measurement. In this case, the aggregation level affects the average patient bin size for the region, clinic division, and physician.



panel as a fraction of the total panel size for the region that satisfies our enabling constraint for the quality measure to be applied. The reliability also depends on the proportionate variance due to the physician-level or the clinic-level as a fraction of the total variance included that of the entire region. To compute the reliability in comparison to that of the region as a whole, we use the Spearman-Brown prophecy formula⁸ for obtaining the reliability as a coefficient of precision.

Note that given this tension between reliability and validity, to explain any variances in a specific case for a given physician, the reliability value of 80% only occurs when the panel size of patients is around 40 for that *type of case combination*. In our case, the physician’s effect on their own panel of patients would only explain about 2.5% of the variation in adherence to the β -Blocker quality measure. So

whether or not we do case-adjustment of the type in the previous analysis, we will not change our belief interval by much since our starting γ will be 0.025.

We can observe therefore that there will be an optimal level of aggregation and an optimal level of case-adjustment for any given quality measure and patient population. As the aggregation level increases in the number of patients, the reliability would increase, but our ability to accurately reflect case adjustment in the *reliability* of the aggregate will decrease. Moreover, the trade-off will be different for each node in the guideline. However, as the first graph shows, the validity or absolute utility for our measurement may still increase with case-specific adjustment since we add or subtract utility from the score as appropriate for each case. Thus we retain the ability to improve our external validity even if there are limits to the improvement of reliability.

Discussion

We have outlined a method for quality assessment that seeks to solve some of the main problems with using medical guidelines as the basis for quality indicators. We have extended our previous work by incorporating ability to embed existing performance indicators in guidelines and include evidence-based guidelines and performance indicators by relating scoring with utilities. Our work exhibits the quality of being *evidence-adaptive*. As we acquire more knowledge about the guideline evidence-base, we can quickly incorporate it in our model in a graceful way and continue with on-going quality assessments.

We have also attempted, as David Eddy has suggested, to solve a long-term goal for automated quality assessment methodologies in medicine.¹⁴ Other work, such as guideline critiquing systems like VQ-ATTENDING¹⁵ or automated utilization review systems like QFES¹⁶ systems have taken either the case-specific or the population-based approach. In our work on the MedCritic system, we have attempted to combine case-based and population-based quality assessment methods using a consistent methodology. We do this by using a representation that unites concepts from automated planning and those from decision theory based on utilities. The central idea in joining these two areas that we exploit is that of decision-making under *ambiguity*, with non-additive belief measures. We have shown how this formulation can encapsulate a lot of features of the automated quality assessment problem.

We are currently extending our work to consider (1) *automated* search for possible explanations in terms of case-adjustments or valid physician variances, (2) automatic search for the best auditing protocol given the relationship between aggregation, case-adjustment, and reliability, (3) modeling strategic behavior on the part of physicians and health plans in relation to their own utilities and to gaming the auditing protocol.

Acknowledgements. This work was supported by the NIH grants LM07033, LM06806, LM05708 and VA grant HSR&D CPI 99-273. We would also like to thank Yuval Shahar, Martin O'Connor, and Samson Tu for valuable discussions. Views expressed in this paper are those of the authors and do not necessarily reflect those of the VA.

References

- ¹ Fields MJ, Lohr NK, eds. Institute of Medicine (US). Guidelines for Clinical Practice: From Development to Use. Washington: National Academy Press; 1992.
- ² Sim I, Gorman P, et al. Clinical decision support systems for the practice of evidence-based medicine. J Am Med Inform Assoc. 2000; 8:527-534.
- ³ Advani A, Shahar Y, Musen MA. Medical quality assessment by scoring adherence to guideline intentions. Proc. AMIA Annual Symposium, Washington, DC; 2001.
- ⁴ Shaneyfelt TM, Mayo-Smith MF, Rothwangl J. Are guidelines following guidelines? The methodological quality of clinical practice guidelines in the peer-reviewed medical literature. JAMA. 1999;281:1900-1905.
- ⁵ National Committee on Quality Assurance. Health Plan Employer Data and Information Set 2000. Vol. 2: Technical Specifications. Washington: NCQA; 2000.
- ⁶ McGlynn EA, Ach SM. Developing a clinical performance measure. Am J Prev Med. 1998;14(3 suppl):14-21.
- ⁷ Graham RP, James, PA, Cowan, TM. Are clinical practice guidelines valid for primary care? J. Clin. Epidemiol. 2000; 53:949-954.
- ⁸ Hofer TP, Hayward RA, et al. The unreliability of individual physician "report cards" for assessing the costs and quality of a chronic disease. JAMA 1999;281:2098-105.
- ⁹ Musen M, Tu S, Das A, Shahar Y. EON: a component-based approach to automation of protocol-directed therapy. JAMIA 1996;3:367-88.
- ¹⁰ Shahar Y, Miksch S, Johnson P. The Asgaard project: a task-specific framework for the application and critiquing of time-oriented clinical guidelines. Artificial Intelligence in Medicine 1998; 14:29-51.
- ¹¹ Goldstein MK, Hoffman BB, et al. Operationalizing clinical practice guidelines amidst changing evidence: ATHENA, an easily modifiable decision-support system for management of hypertension in primary care. Proc. AMIA Annual Symposium, Los Angeles, CA; 2000.
- ¹² Advani A, Lo K, Shahar Y. Intention-Based Critiquing of Guideline-Oriented Medical Care. Proc. AMIA Annual Symposium, Orlando, FL; 1998.
- ¹³ Augustin T. On decision making under ambiguous prior and sampling information. Proc of the 2nd Int'l Symp on Imprecise Probabilities. Cornell Univ.: Ithica, NY, 2001.
- ¹⁴ Eddy DM. Performance Measurement: Problems and Solutions. Health Affairs 1998;17(4):7-25.
- ¹⁵ Miller PL. Goal-directed critiquing by computer: ventilator management. Comp.Biomed.Res. 1985; 18:422-38
- ¹⁶ Balas EA, et al. An Expert System for Direct Delivery of Published Clinical Evidence. JAMIA;3:56-65