MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

An Intelligent Case-Adjustment Algorithm for the Automated Design of Population-based Quality Auditing Protocols

Aneel Advani[a,d], Neil Jones[b], Yuval Shahar[c], Mary Goldstein[a,d], Mark A. Musen[a]

[a]*Stanford Medical Informatics, Stanford University, CA, USA* [b] *SCHIN, University of Newcastle, UK*
[c] *Department of Information Engineering, Ben-Gurion University, Israel*
[d] *GRECC, VA Palo Alto Health Care System, Palo Alto, CA, USA*

Abstract

We develop a method and algorithm for deciding the optimal approach to creating quality-auditing protocols for guideline-based clinical performance measures. An important element of the audit protocol design problem is deciding which guideline elements to audit. Specifically, the problem is how and when to aggregate individual patient case-specific guideline elements into population-based quality measures. The key statistical issue involved is the trade-off between increased reliability with more general population-based quality measures versus increased validity from individually case-adjusted but more restricted measures done at a greater audit cost. Our intelligent algorithm for auditing protocol design is based on hierarchically modeling incrementally case-adjusted quality constraints. We select quality constraints to measure using an optimization criterion based on statistical generalizability coefficients. We present results of the approach from a deployed decision support system for a hypertension guideline.

Keywords:

Quality Assessment; Clinical Audit; Medical Guidelines; Case-Adjustment; Performance Measures

Introduction

Evidence-based guidelines have now become an increasingly important starting point for the design of quality indicators for clinical audits and to measure the quality care [1]. In our work on the Quality Intelligence Language (QUIL) system, we have addressed some of the challenges involved in creating a language and system to model and evaluate the quality indicators that can be derived from a medical guideline. We have previously outlined a framework for modeling quality indicators from the intentions of the guideline authors [2], a method for scoring adherence to these intention-derived quality indicators [3], and an algorithm for deriving coherent "global" quality measures and auditing queries from formal guideline specifications [4]. In this paper, we focus on how to design an auditing protocol from the set of quality indicators derived from a guideline specification.

The audit protocol design problem, even when a coherent set of guideline-based quality indicators has been developed, requires further refinement and analysis. First, the costs to a medical care organization of preparing data sources and conducting large-scale queries for quality indicators can be as high as $28,000 to $76,000 per measure [5]. In practice, therefore, measuring all the quality indicators derived from the guideline may be prohibitive. Moreover, it has been shown that reporting quality indicators that are too complex is counter-productive for their adoption by purchasers and consumers of medical services [6].

Second, as with most statistical or data-analytic model selection problems, there exists a validity-reliability (bias-variance) trade-off in the selection of quality indicators for the auditing protocol. The better the auditing protocol fits the full guideline-derived and fully case-adjusted model of the guideline, the greater the danger of overfitting the data and reducing the reliability of the model and preventing generalizability to other data sets. Similarly, the more global the quality indicators used in the auditing protocol, the greater the reliability over a larger population, but the greater the potential for a biased estimate because key covariates may not be incorporated in the global quality indicator.

Thus, the main challenge in choosing an auditing protocol from a model of guideline-based quality indicators is defining a design objective function that is sensitive to the validity-reliability trade-off, that can model costs of measurement, and that finds the optimum query-set of quality indicators and level of case-mix-adjustment. An algorithm that automates the appropriate inclusion and level of case-mix-adjustment of quality indicators in an auditing protocol also leads us to a solution to the more general problem of how to obtain population-based quality indicators from individually case-specific guideline interpretations.

In the rest of the paper, we discuss the specification of an objective function and algorithm for automated optimal auditing protocol design. We present some results of the approach as applied to the development of auditing protocols in relation to a guideline used in a decision support system for hypertension care [7].

Methods

Hierarchical Elaborations of Quality Indicators

One strategy to solving the problem of creating quality measures that have greater reliability from model-simplicity, without losing validity, is to impose a particular semantic structure on the set of quality indicators derived from a guideline. We do this by modeling the quality indicators in a *quality constraint structure* (a directed acyclic graph or DAG) so that more heuristically general but valid properties of the guideline are defined higher in the hierarchy. These more general quality indicators represent the *intentions* of the guideline authors. For example, on the left side of Figure 1, we see a quality constraint structure for a hypertension guideline. In the model, the node "Drug Treatment" is a quality indicator representing "appropriate treatment with drug therapy". The concept is not observable directly; however, we can measure it by defining the concept in terms of quality indicator queries that are themselves measurable as well as directly observable. Therefore, the node "Rx HTCZ" for "treatment with thiazide diuretics" is both measurable and can be directly observed with a database query. We note that the quality constraint structure also immediately distinguishes the case-adjustment relationships between indicators. For instance, the higher "Rx b-Blocker" node is defined for those patients who have diagnoses of hypertension (HTN) and coronary artery disease (CAD) and no diagnoses of diabetes (DM). These comorbidities are *enabling constraints* cumulatively inherited from the circular nodes higher up in the structure. The enabling constraints specify the entry conditions that any patient case must satisfy in order to be considered part of the population for the quality indicator. The lower "Rx b-Blocker" indicator is therefore an adjustment to the parent indicator, with the additional enabling constraint that the patient have no diagnosis of chronic obstructive pulmonary disease (COPD). Thus, we have a monotonic relationship between the height of quality indicator in the intention structure and the amount of case-adjustment for the indicator. Since these enabling constraints inherit conjunctively downward, the set of patient cases that satisfy a node are always a subset of those that satisfy any of its parent nodes. What this means for the audit design problem is that the more case-adjusted the enabling constraints in node, the fewer patients in the population the node selects for. Therefore, we have a monotonic relationship between the size of *query populations* of the nodes and the amount of case adjustment. From the two properties of node height above, we note that the validity of an audit will increase with increasing case-adjustment, and therefore an audit that stops at a higher-level node has the potential to be less valid. However, the higher audit, the greater the reliability of the audit since the size of the query population varies with the height of the node. Thus, the audit protocol design problem consists of deciding how far down in the structure to go to include nodes for the auditing protocol [see Figure 1].

Audit Protocol Design Objective Function

Objective Function

With the given relationships between node height, case-adjustment, and the query population size of nodes, we can construct an objective function that will allow us to optimize the validity-reliability trade-off. In general, we would like our objective function G for any given node i, to be of the form:

G[node i] = reliability(data-->node i sub-structure)
** validity (node i sub-structure --> node i parents* (1)

For the actual form of the function G, we use a formulation called generalizability theory, which is an extension of classical reliability theory that incorporates multi-factorial analysis of variance [8]. In our case, we use a generalizability coefficient for G based on a fully crossed two-facet mixed design. The variation-causing *facets*, or factors, include (A) the selection from a fixed set of node query ratings (r) that are descendant of node i in the quality constraint structure and (B) random sampling from a large set of potential patient case observations for each node (o). Lastly, each of these is crossed with the selection of the professional person whose behavior we that we are auditing (p). The *generalizability coefficient* for an P x R x O design is given by:

$$G[i](X_{pro}, \mu_p) = E\rho^2(X_{pro}, \mu_p) * E\rho^2(X_{pro}, \mu_p) \qquad (2)$$

where,

$$E\rho^2(X_{pro}, \mu_{pr}) = \frac{\sigma^2(p) + \sigma^2(pr)}{\sigma^2(p) + \sigma^2(pr) + [\sigma^2(po) + \sigma^2(pro)]} \qquad (3)$$

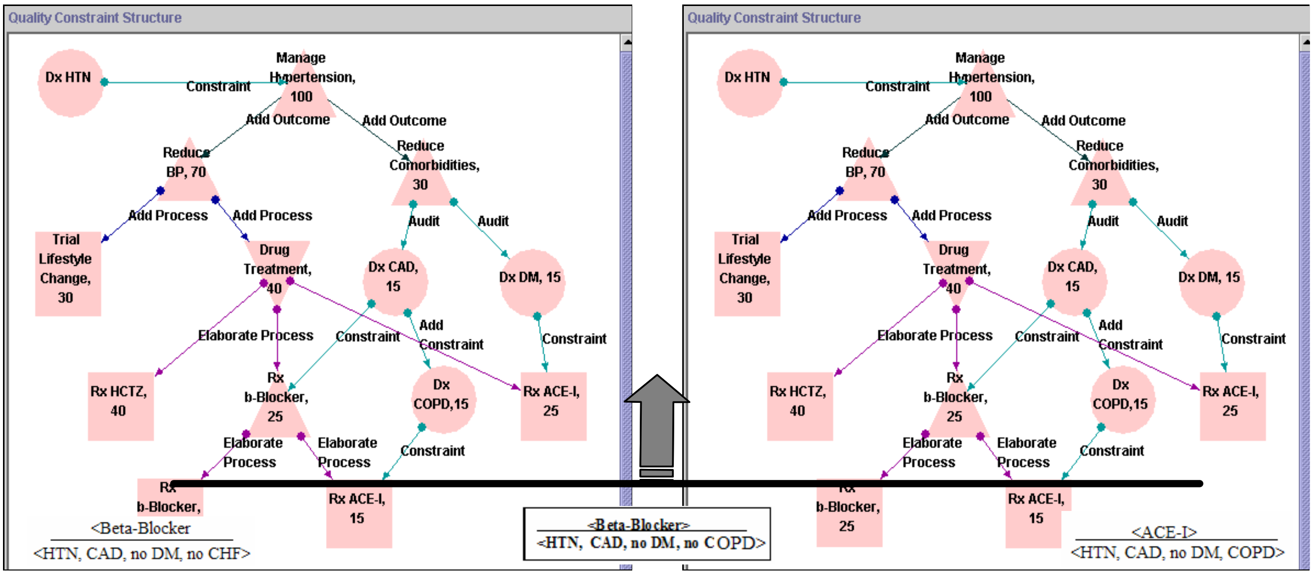
and,

$$E\rho^2(X_{pro}, \mu_{pr}) = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(pr)} \qquad (4)$$

Each of the equations (3) and (4) defines an intraclass correlation coefficient in the form $\sigma_g^2 / \sigma_g^2 + \sigma_e^2$giving the fraction of variance in the true generalization (σ_g^2) that is explained by the expected observations ($\sigma_g^2 + \sigma_e^2$). Equation (3) is a generalization coefficient that gives an estimate of the reliability of generalizing from individual observations of clinician scores for a particular node query and a particular patient case (X_{pro}) to a mean score for a clinician's performance rating for a given node query (μ_{pr}). Similarly, Equation (4) is a generalization coefficient that gives the squared correlation between clinician performance scores observed with a perfectly reliable fixed set of query nodes in the sub-structure of node i (μ_{pr}) and a broader universe of generalization (the more general quality indicators that are parents of node i). Note that as the variance factor σ^2(pr) increases, the reliability (from Equation (3)) *increases* and the validity (from Equation (4)) *decreases*. Increasing the variance σ_g^2(pr) can be done by freeing the query node i from a more standardized set of sub-queries, that is, from stopping the auditing at a node without case-adjusting using lower-level nodes. Thus, if we case-adjust a quality indicator with more fixed query nodes required to measure it, then the variance σ^2(pr) decreases and the quality indicator becomes more valid but less reliable. The objective function *G*[*i*] for a quality indicator node i thus encapsulates the validity-reliability trade-off for case-adjustment as required.

Our optimization algorithm for finding the optimal level of case-adjustment works from the "bottom-up". As Algorithm 1 shows, the optimization starts by calculating the generalizability coefficients from the leaf nodes in the sub-structure of a given quality indicator node. Then it goes up one level in the structure and recalculates the generalizability coefficients, with the case-adjusted lower nodes omitted from the query. The process is repeated to the top-most node until an optimum level, L(i), for the omission of lower-level nodes that maximizes G[i] is found. Now we have a set of levels, L[i], one for each node in the structure which each independently are optimal levels of case-adjustment for that node. Our algorithm sweeps unidirectionally from the bottom up.

Figure 1 - QUIL Quality Constraint Structure for JNC VI Hypertension Guideline. The screenshots show the quality constraint structure for a hypertension guideline. Each leaf node is contains a QUIL query that represents the quality indicator for that part of the guideline. A schematic of the query for the "Rx beta-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 node "Rx b-blocker" is elaborated in the structure to the right into the nodes "Rx b-Blocker" and "Rx ACE-I", where the beta-Blocker treatment indicator has been adjusted for patients without and with COPD, respectively. The numbers in the nodes refer to the utilities to the patient of satisfying the quality measure. The black bar represents the process of limiting the case-adjustment level in an auditing protocol. For our structure, the auditing protocol desing problem consists of finding what level the case-adjustment bar is optimally placed.



Therefore, the algorithm makes the assumption is that we would rather query higher-level nodes than particular lower level nodes at the expense of higher-level nodes in some other branch. Therefore, the algorithm minimizes the query-related costs associated with querying an extra set of case-adjusted nodes when they would not improve the generalizability. This algorithm minimizes the total number of nodes queried, however, it doesn't consider the utilities or differential costs of adjusting one set of nodes versus another.

Algorithm 1. Automated Case-Adjustment for Auditing Protocols in QUIL.

```
-----------------------------------------------------------------------
Q[j] ← Sort nodes {i} in QUIL quality constraint structure
in reverse topological order

For i = {1, … , n(Q) d
        G max = 0
        For l = {1, …, i} do
                If Gmax < G[I] then
                        Gmax ← G[I]
                        L[i] ← height of node l
                                in structure
                l ← l + 1
        Output L[i]
-----------------------------------------------------------------------
```

We can extend the algorithm to include weightings by the utility or importance weight of a quality indicator node. To do this we need to define the utility of improving the generalizability of a node in comparison to that of its parent using the differences in utility weights defined for each of the quality nodes (see 1). Thus, if we have a set of utility weights U(i) (normalized to be between zero and one) associated with each node i, we can define a global generalization function G* given by the weighted average and repeat Algorithm 1 using G* for each node i and taking the level L*[i] that gives the final maximal G[i] over all i.:

$$G^* = \sum_{i \in \{nodes\}} U(i)G[i] \qquad (5)$$

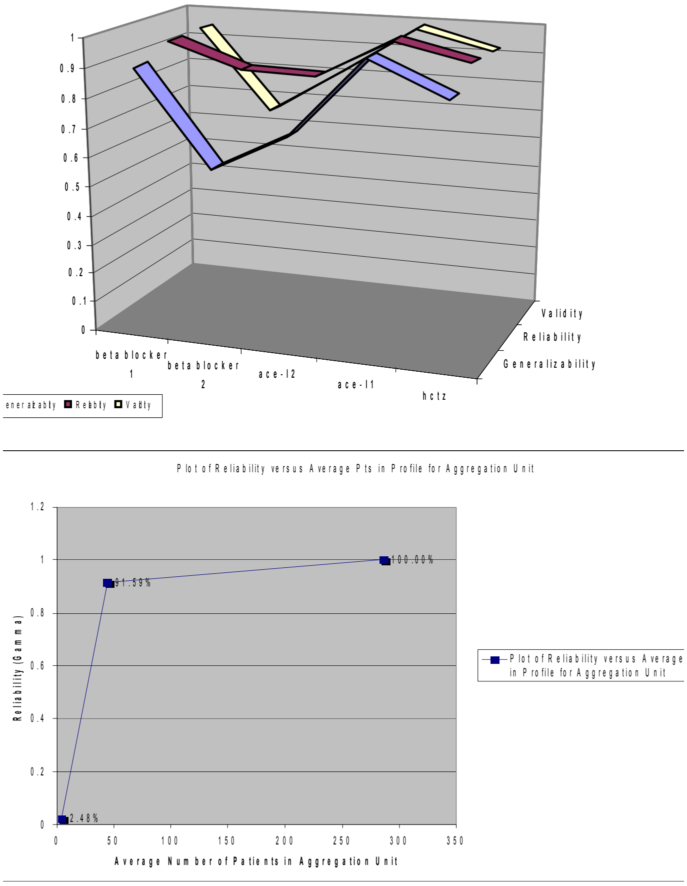
This algorithm would allow us to pick one level of case-adjustment for the entire quality constraint structure at the expense of the generalizability of particular nodes.

Results

We present results relating to our optimization criterion and our algorithm applied to auditing protocols used in our work on a guideline-based decision support system for hypertension care, called the ATHENA system [9]. Our optimization criterion is applied 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 analyze the behavior of the optimization criterion G, for five nodes in the sub-structure under the "Drug Treatment" node in our hypertension quality structure from Figure 1. In the upper graph in 2, we plot the reliability, validity, and generalizability coefficients for the five queryable nodes in the quality structure. Note that in this region of bin sizes and adjustment levels, the relationship between validity and reliability is clearly correlated.

If we take the set of nodes in a sub-structure of the higher (adjusted) Beta-Blocker node we can see the effect of adjusting the node. The "Rx b-blocker" is adjusted to "Rx ACE-I" for the presence of COPD. We can see that when the node is adjusted, as embodied in the higher-level “Rx b-blocker” node (referenced as “beta-blocker 1” ordinate in the graph in 2) the validity goes up much more than the reliability when compared to the “beta-blocker 2” node. Thus, when the node can be adjusted to discriminate between special cases of the patient population, the measurement it represents is more valid and generalizable to larger sets of cases. In contrast, the nodes that are closest to the data have their relative strengths from their reliability rather than validity.

Figure 2 - Results of Case-Specific and Aggregation-Level Adjustment for Hypertension Drug Treatment Quality Measures. The first graphs shows the validity, reliability, and generalizability coefficients for the nodes in the sub-graph under the «Drug Treatment» node from Figure 1. The graph shows that in this case the reliability and validity give the same rank ordering of generalizability. 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.



We said above that stopping an audit at a higher node would reduce the validity. This is still a true statement since stopping an audit and not adjusting a node where it could be adjusted by its descendants would decrease the validity. This reasoning is consistent with the observation that leaf nodes, which are equivalent to nodes that are not considered for further specialized case-adjustment, have the lowest validity coefficient in the graph in Figure 2 .

Figure 2 also illustrates the solution to our algorithm for optimizing the level of case-adjustment in the auditing protocol. It is clear from the results of the generalizability coefficients that the optimum level of case-adjustment would be at the level of the black bar shown in Figure 1. We also show in Figure 2 that the within-group reliability of the quality indicators varies with their population bin size as well. The reliability coefficient thus depends not only on the group interaction effects of the patient case and physician decision facets but also on the single-node variability for a given patient case, physician (or clinic) decision. The effect is most apparent if the higher-level indicators of the quality constraint structure are in fact applied to larger organizations such as hospitals or HMOs. Then the query population effect and the effect of standardizing a larger proportion of the audit protocol both increase the reliability. However, the bin size effect does not do this at the expense of the validity of generalization, unlike the standardization effectDiscussion

The use of generalizability theory in the context of guideline-based quality indicators and guideline-based audit protocols affords us a path to solving some other challenges in research on guidelines. For instance, an important issue in guidelines is the question of how best to model higher-level goals or intentions in the relation to the guideline model. Some authors [10] [11] have suggested that intentions and goal knowledge should be closely tied to the guideline planning and execution language itself. This analysis leads to the view that a guideline interpreter or execution engine must be available for intentions to be meaningfully in quality assessment analyses.

However, we can however answer this question more definitively using generalizability theory as we have intimated in our results above. Using generalizability theory we can look at the effect of individualized case-adjustments on the final reliability and validity of scoring adherence to the higher-level indicators, including guideline intentions. If the population-aggregate measures give answers and have reliability and validity similar to the individually case-specific measures that the guideline interpreters and planners allow us to use, then retrospective intelligent quality assessment can still be done without the presence of a full guideline interpreter. Thus, modeling intentions can be done in a “model-theoretic” instead of “proof-theoretic” fashion, without the use of a guideline interpreter or planning language

The methods we have presented above can be extended in many ways. One major extension required to the algorithm above is to allow hidden intermediate nodes in the auditing protocol design. This problem could be extended to the general problem of finding the best subset of nodes – no matter where in the quality structure – to use in order to reduce measurement costs as well as to maximize guideline adherence. The solution would require a more sophisticated heuristic search algorithm than the unidirectional sweep that we have outlined above. It would also need a theory of imputation for missing node data given a limit to the amount spent on making additional queries. That would help design lower cost auditing with an understanding of how same objective function changes with real-world pressures. Ultimately, the goal of this approach is to reach a point where a rotational strategy can be developed for mandated reporting of much more sophisticate population-based quality measures than are used now. That would allow algorithms such as ours above to be used directly for current HMO report card survey strategies.

Conclusion

We have presented a formulation and an algorithm for helping to automate the audit protocol design problem in quality assessment. To do so, we had to add semantics such as modeling case-adjustment to the statistical aspects of quality measure scoring and at the same time add quantitative design to the logical methods used guideline-based quality assessment of agent behavior.

Acknowledgments

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 Samson Tu, Susanna Martins, Robert Coleman, Martin O'Connor, and Ravi Shankar for valuable contributions to the ATHENA decision support system. Views expressed in this paper are those of the authors and do not necessarily reflect those of the VA.

References

[1] Using Clinical Practice Guidelines to Evaluate Quality of Care. Vol 2: Methods. Washington: US Dept. of HHS; 1995. (AHCPR/AHRQ Pub. No. 95-0046.)

[2] Advani A, Goldstein MK, & Musen MA. A framework for evidence-based quality assessment that unifies guideline-based and performance-indicator approaches. In: *Proc of the 2002 AMIA Fall Symposium*, San Antonio, TX

[3] Advani A, Shahar Y, & Musen MA. Medical quality assesment by scoring adherence to guideline intentions. In: *J Am Med Inform Assoc* 2002; 9(9):S92-S97.

[4] Advani A, Goldstein MK, Shahar Y, & Musen MA. Developing quality indicators and auditing protocols from formal guideline models: knowledge representation and transformations. In: *Proc of the 2003 AMIA Fall Symposium*, Washington, DC

[5] Ahmed F, Elbasha EE, Thompson BL, Harris JE, Sneller V. Cost-Benefit Analysis of a New HEDIS Performance Measure for Pneumococcal Vaccination. *Med Decis Making* 2002; 2(Suppl):S58-S66.

[6] Hibbard JH, Jewett JJ, Legnini MW, & Tusler M. Choosing a health plan: do larger employers use the data? *Health Affairs* 1997; 16(6):172-180.

[7] 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. 2000 AMIA Fall Symposium*, Los Angeles, CA; 2000

[8] Brennan RL. *Generalizability Theory.* New York: Springer-Verlag, 2001

[9] 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

[10]Miller PL. Goal-directed critiquing by computer: ventilator management. *Comp.Biomed.Res*. 1985; 18:422-38

[11]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.

Address for correspondence

Dr. Aneel Advani
Stanford Medical Informatics
MSOB x215, 251 Campus Dr West
Stanford, CA 94025