**EDITORIAL COMMENT**

Can We Appropriately Measure Appropriateness?*

Paul T. Vaitkus, MD

Rockford, Illinois

“Competence, like truth, beauty, and contact lenses, is in the eye of the beholder.”

—Laurence J. Peter (1)

(And one might add “appropriateness” to the list as well).

A few years ago, at a national interventional cardiology meeting, the speaker regaled the audience with an anecdote that may have been apocryphal but was entirely believable. At some earlier meeting, this speaker had challenged his audience with 3 sequential questions to which he invited a show of hands. The first was, “Have any of you witnessed an inappropriate coronary intervention?” in response to which virtually all of the hands in the room were raised. The second was, “Have any of you witnessed someone else in this room conduct an inappropriate intervention?” The vast majority of hands shot up. And finally, “Which of you yourselves has performed an inappropriate intervention?” virtually all of the hands in the room were raised. The first was, “Have any of you witnessed an inappropriate coronary intervention?” in response to which virtually all of the hands in the room were raised. The second was, “Have any of you witnessed someone else in this room conduct an inappropriate intervention?” The vast majority of hands shot up. And finally, “Which of you yourselves has performed an inappropriate intervention?”

None of the hands was raised. For the record, I will publicly confess to having performed clinically inappropriate interventions but for a justifiable reason. As is true of most interventionalists, I have partnered with noninterventionalists for all of my career, spanning both academia and private practice. Confronted with a request by a noninterventionalist that I perform an intervention in a situation that of my own accord I would not have elected to undertake, usually because I did not think that the clinical circumstances justified it, I viewed that my responsibility to my partners was to provide a service that they could not provide for themselves. If a properly trained, thoughtful cardiologist felt that the intervention was clinically important, I acceded to their judgment. Reasonable individuals confronted with the same data set could reasonably reach very different conclusions. Therein lies the principal conundrum of all measurement of appropriateness.

The study by Chan et al. (2) in this issue of the Journal presented the data of an unusual “natural experiment” that will not likely be readily replicated. The authors, aware in advance of the development of the American College of Cardiology’s (ACC’s) appropriate use criteria (AUC) for coronary revascularization (3), provided the same clinical scenarios and questions being used by the committee developing the AUC to a broader group of practicing cardiologists and queried them for their assessments of appropriateness. The participating cardiologists were drawn from a wide range of institutions and clinical backgrounds and were entirely unaware of the impending public release of the AUC. The findings of the study by Chan et al. (2) were that although overall there was a moderately strong correlation of overall appropriateness scoring by practicing cardiologists and the AUC expert panel, individual scores frequently varied widely.

These findings would make for merely interesting academic reading were it not for the potential administrative applications of the AUC, particularly in the current environment of universal health care reform. Of the many undercurrents of health care reform that have generated anxiety for physicians, the one that potentially most directly relates to the AUC is that of seeking health care savings by elimination of reimbursement for “unproven therapies” or “inappropriate procedures.” It is also not a very large leap of faith to suggest that physician-profiling efforts by government and other payers might also seek to incorporate the AUC. It is in this context that a critical review of appropriateness measurement is essential and wherein the current study, and other lines of evidence, cast serious doubt that the AUC is ready for “prime time.”

First, we must examine how “appropriateness” criteria have been developed in the AUC and previous such attempts. All such efforts have been variations on a committee examining published clinical data. The approach was most vigorously first used more than a decade ago by the RAND Corporation (4). The assembled experts provide a numerical assessment and then seek to reach group consensus. The intrinsic methodologic limitations must be acknowledged. The committee’s decisions—like any assemblage of humans (including juries and the Supreme Court)—will reflect the biases of the participants. Furthermore, consideration of any consensus-seeking committee should raise the specter of “groupthink,” which can lead to flawed conclusions in the name of group cohesiveness. Although other dimensions on which physicians are or will soon be profiled have, at least nominally, a measurable outcome, “appropriateness” lacks a quantifiable gold standard. When the appropriateness criteria derived by different committees has been put to empiric examination, the degree of agreement has been decidedly poor (5). As conscientious, earnest, intelligent, and well informed as the members of any of the disparate committees might be, the fact that they publish contradictory guidelines is not particularly surprising given the high

---

*Editorials published in the Journal of the American College of Cardiology reflect the views of the authors and do not necessarily reflect the views of JACC or the American College of Cardiology.

From the Midwest Heart Specialists, Rockford, Illinois. Dr. Vaitkus is an employee of Daiichi-Sankyo Inc.
degree of subjectivity involved in the process. In all of these efforts, the extreme cases and scenarios are not usually in doubt. There are clear-cut situations in which interventions may be either life saving or undeniably dangerous. It is the vast, gray middle in which there is little agreement. This is the very realm in which a “gold standard” of appropriateness, if such a thing could possibly exist, would have the greatest utility.

It is important to note that although published practice guidelines and efforts such as the AUC are not intended to be static, and certainly undergo revision and updating as new data emerge, the process of revision is necessarily slow and AUC-based physician profiling and reimbursement strategies could potentially penalize physicians who either generate pioneering data or are early adapters. Two examples from the current version of the AUC—percutaneous left-main coronary intervention and the open-artery—hypothesis-driven late post—myocardial infarction intervention—illustrate the issues. The current AUC places both in the “inappropriate” category.

In the case of left—main intervention, large amounts of data have been and continue to be published that this is a viable clinical alternative (6). Had AUC-based profiling and reimbursement already been in place, would the clinical investigators who generated the new data, facing lack of payment for their clinical efforts and adverse consequences of negative personal “appropriateness” ratings, have been inhibited in their clinical research efforts? Do we Americans thus cede all future efforts at “pushing the envelope” to our international colleagues?

Just proximate to the publication of the AUC, a well-performed meta-analysis of the open-artery clinical trials suggested that late intervention may confer a survival advantage that was not identified in any of the individual trials, each of which had less power than pooling of all available data (7). Would the “appropriateness profile” of the physician who now accepted the results of this newer publication suffer adverse effects until the AUC are revised several years hence?

If we examine the limitations of physician profiling on other dimensions such as clinical outcomes or economics, any such efforts based on “appropriateness” criteria become all the more suspect.

Profiling physician performance on “hard” clinical end points, such as revascularization procedure—related myocardial infarction or mortality, has required decades of rigorous empiric development and continuous ongoing revision and recalibration of models (8). These efforts underscore just how difficult it is to get it right, a rigor that is simply not present in the “appropriateness” arena. As was recently highlighted, other “clinical performance” measures that lack a long history of comprehensive case-mix adjustment serve to emphasize that they risk misleading more than informing (9).

But even with perfect case-mix adjustments, public profiling on clinical outcomes such as the ongoing public dissemination of revascularization outcomes data in New York state, the likelihood of unintended consequences is often ignored in the zealous 1-dimensional efforts of bureaucrats. One can reasonably argue that such initiatives serve to undermine appropriate patient care as the score-carded physicians become risk averse (10).

Profiling on economic outcomes has thus far proven to be a morass. Substantial errors in the classification of physicians have been identified (11). Some of the sources of the erroneous profiles may not be easily amenable to improvement. For example, although insufficient case-sample size is one source of unreliability, adjustment of sample size by a substantial amount reduces the probability of misclassification only to a small degree (11). How to increase sample size could be an issue without a solution. Although each payer may initially be hampered by having data only on cases for which they pay, even if all payers pooled their data into one master database, a particular physician’s data are limited by the number of cases he performed in a given span of time. Increasing the span of time to cover several years then may defeat one of the goals of such a profiling effort, that is, encouraging improvement over time. What is to be gained by scoring a physician on cases performed several years in the past? Cost attribution is another nettlesome issue. A cardiologist is profiled on the costs of caring for an inpatient with a cardiac diagnosis, but another physician involved in the case may order cardiac tests that I would have deemed superfluous. How is this captured? To whom are the added costs attributed? These issues have led RAND investigators to conclude that “current methods of physician cost profiling are not ready for prime time” (12).

Finally, there has been a modest quantity of data on the application of previous “appropriateness” criteria as a management tool. But first, we must expose an unsubstantiated bias of the regulators, financiers, and politicians. Undeniably, geographic variations in usage of procedures exist (13). These differences are all too frequently and without qualification interpreted as evidence of inappropriate “overutilization,” and the possibility of substantial underutilization in areas of undersupply of physicians and other health care resources is simply ignored (14). Thus, while inside the Beltway, the discussion will focus on how to use “appropriateness” criteria to quash overutilization, the available empiric data suggest that “appropriateness” criteria may very well be more sensitive to exposing underutilization, not overutilization (15—17). But the bias on the part of payers will be distinctly one-directional. The Centers for Medicare & Medicaid Services auditor, charged with containing costs, is not likely to investigate cases of medically managed coronary artery disease and question whether the patient should more appropriately have been transported to the catheterization laboratory.

So where does this leave us? The ACC expended resources on developing the AUC and should be commended for taking the lead before a federal agency decided it was
their purview. But if they developed an officially ACC-sanctioned tool that the membership may not endorse, as the current research suggests, perhaps the initiative needs to be revisited and refined. Additional research may focus on to what degree ACC members agree with the criteria or whether a broader consensus may lead us to more convincing measures of appropriateness. In this era of electronic data gathering, it would not be such an onerous effort to reach out to the entire membership to voice their agreement or disagreement with each of the appropriateness criteria. Set it up on the ACC Web site and let all of us vote on it. If 5,000 ACC members say an indication for percutaneous coronary intervention is appropriate, maybe it really is appropriate. And if 5,000 potential expert witnesses agree that in a particular clinical setting, I should not have performed a percutaneous coronary intervention, perhaps I will refrain from doing so. Replicate the AUC experiment. Set up half a dozen more committees and see if they come up with concordant or contradictory criteria. Survey the prospective committee members for their attitudes and biases before they enter the discussion phase. If we intend to profile the practicing interventionists, why not also profile the criteria generators? Check any tendencies of the Centers for Medicare & Medicaid Services or others to misapply the AUC as a physician-profiling or reimbursement policy tool. Run some physician-profiling experiments and determine the variability of measurement of case appropriateness in the hands of potential auditors. If the AUC cannot be operationalized, then it should not be allowed to be.

Whatever methods the ACC decides upon, it is imperative to gather more data to assess the degree to which the membership does or does not support the AUC, to demonstrate the strengths and flaws of how the AUC might be applied in practice, and to demonstrate whether similar appropriateness scores would be developed by alternate methodologic approaches. All of these questions need to be addressed before payers begin to adapt this tool to their ends.

Reprint requests and correspondence: Dr. Paul T. Vaitkus, Midwest Heart Specialists, 1340 Charles Street, Suite 300, Rockford, Illinois 61104. E-mail: pvaitkus@dsi.com.

REFERENCES


Key Words: appropriate use criteria ● concordance ● PCI.