On Evaluating the Efficacy of Anesthetic Practices: Need for Attention to Clinical Details

To the Editor.—The finding by Moller et al. that pulse oximetry could not be shown to affect anesthetic outcomes raises disturbing methodologic questions. If pulse oximetry, an innovation widely regarded as essential for modern safe anesthesia, cannot be shown to improve outcomes, then what anesthetic innovations or practices can?

The authors suggest that the study was too small and thereby lacked sufficient power to detect a significant difference with high probability. If this explanation is the answer, it bodes poorly for attempts to validate the majority of innovations or practices whose contributions are still less profound.

However, the authors also speculate that there may be a need for improved methods of evaluation. Three examples illustrate that, indeed, the problem is not primarily in the large sample sizes mandated by statistical theory but instead in the experimental logic used to evaluate new technologies for clinical practice.

First, failure to achieve a statistically significant result does not necessarily imply that an innovation has no effect on outcome.

Nineteen of 104 responding clinicians stated that there were instances in which pulse oximetry helped avert an anesthetic mishap. Determining whether the circumstances in these instances occur often enough to make the benefits of pulse oximetry cost-effective is a valid judgment. Only one well documented instance may be sufficient to demonstrate that oximetry can have value. In fact, 83% of the responding physicians felt that pulse oximetry would be of value in their practice.

These clinicians must have been taking into account factors not included in the necessarily restricted set of outcomes used in the study. Experimental logic that finds no effect of pulse oximetry without first exploring in greater detail the basis for these clinical judgments needs to be reexamined.

Second, even if, with great effort and expense, we increase the sample size to a point at which theory states there is a high probability of rejecting the null hypothesis, in practice the null hypothesis is often of little clinical interest.

Suppose the study had the necessary sample size—the authors estimate it would have required half a million patients—needed to reject a null hypothesis that the incidence of, e.g., myocardial ischemia is identical in patients monitored with and without pulse oximetry. What have we gained? The summary conclusion of an overall statistical difference in rates between groups of diverse patients is often not much help. It does not tell a policy-maker, for example, whether it is cost-effective to insist on oximetry in all ASA physical status 1 patients having regional anesthesia. Nor does it help a clinician decide how to use oximetry in a particular ASA 4 patient with weak pulses due to peripheral vascular disease.

Clinicians are concerned less with whether an innovation is, on average, "effective" than with identifying the situations in which the use of an innovation is warranted. Although clinicians continually strive to make more refined clinical distinctions and increasingly more accurate clinical predictions, the logic of randomized clinical trials often fosters a tendency to eliminate the detail needed by clinicians to make these distinctions. Sir Ronald Fisher, the eminent statistician, said that randomization "relieves the experimenter of the anxiety of considering and estimating the magnitude of the innumerable causes by which his data may be disturbed." Randomization allows us to compare the average response in treatment groups. However, the clinical details that we are averaging out are the crux of medical practice.

The authors incorporated some clinical details in their logistic regression equations, but they used this technique to eliminate the effects of these variables from estimates of the efficacy of oximetry rather than to focus on those effects.

Third, increasing the sample size does not necessarily lead to increased power to detect the effect of an intervention on rare outcomes.

To illustrate, suppose that, in ASA 3 and 4 patients, the incidence of postoperative atelectasis is 1% in patients monitored by oximetry and 3% in those managed without oximetry. In the approximately 2,000 such patients in the study by Moller et al., such a difference would be statistically significant at the P < 0.005 level. However, in this study, the sample size was even larger because the sample included an additional 18,000 ASA 1 and 2 patients. Suppose that, in this large subgroup of relatively healthy patients, there is a lower incidence of postoperative atelectasis, say 0.5%, regardless whether pulse oximetry is used. Lumping together all the patients in the study without distinguishing these clinically relevant subgroups gives a much larger sample size but does not increase the power of the study; rather, it leads to an overall nonsignificant result. Failing to preserve the clinically relevant distinctions recognized by physicians can obscure important effects while appearing to increase statistical power. Indeed, this could provide another reason why the study was unable to demonstrate statistically significant results.

Of course, dredging the data to find potentially important subgroups involves the risk of uncovering false-positive associations due merely to chance. However, such associations need not be accepted at face value; the methods of epidemiology and clinical investigation for assessing clinical plausibility, strength of association, and replicability can be brought to bear as well. Drawing conclusions based upon these additional considerations is less automatic and frequently less clear-cut than simply relying on the verdict of a P value but may lead to results that are clinically more informative.

To make useful distinctions in patient care, clinicians require a research paradigm that focuses on details rather than eliminates them. If, as the authors suggest, "Improved methods for evaluation of new standards and monitoring equipment are needed," the foundation for such improved methods should be based on attention to clinical detail.

Alan R. Shapiro, M.D.
Department of Anesthesiology
St. Luke's-Roosevelt Hospital Center
428 West 59th Street
New York, New York 10019
References


(Accepted for publication July 15, 1993.)

In Reply---Shapiro addresses pertinent issues with regard to some of the many methodologic problems when evaluating innovations in modern anesthesia.

The randomized evaluation of pulse oximetry was designed to include 20,000 patients to obtain a minimum statistical power of 90% for the overall number of postoperative, cardiovascular, and respiratory complications. In that regard, the study was not too small. However, as we have stated, if one was to look at one specific complication, e.g., myocardial infarction, the sample size was not sufficient. To study a specific complication with a very low incidence, no matter how clinically relevant, one would need an impractically large sample size.

Stratification and logistic regression analysis was used not only to eliminate the effects of the known confounders but, equally important, to look for significant subgroups of patients or situations in which pulse oximetry monitoring might show significant results. Unfortunately, no such subgroups were identified, i.e., no significant difference or even trend in the figures was found when evaluating patients in ASA physical status 3 and 4.

The answers to the questionnaire of the participating anesthesiologists and the previous findings in the “American Society of Anesthesiologists closed claim study” make a strong argument for pulse oximetry monitoring but must be seen with the objective findings from the randomized evaluation suggesting that perioperative pulse oximetry monitoring was not a “break-through” that could reduce the number of postoperative complications substantially. However, anesthesiologists all over the world already have adopted pulse oximetry in clinical practice as a tool that guides them in the daily management of their patients, in teaching situations, and in emergencies. Future research concerning standards, patient care, and monitoring perhaps ought to focus more on clinical details, as stated by Shapiro; however, until we have identified and evaluated new research methods, the prospective randomized study remains the “gold standard” if practically and clinically possible. The randomized evaluation of pulse oximetry has illustrated that the problems with such studies are numerous. Nevertheless, we believe these problems can and must be mastered. Limiting the inclusion criteria to specific risk groups (e.g., ASA higher than 3, age older than 65 y, upper abdominal surgery), isolating more rigorous outcome variables, and establishing a wide degree of cooperation between departments and countries, new prospective anesthesia safety studies must be performed in the future.

Jakob Trier Moller, M.D.
Department of Anaesthesia
Rigshospitalet
University of Copenhagen
Blegdamsvej 9
DK-2100 Copenhagen
Denmark

References


(Accepted for publication July 15, 1993.)