Research Issues Symposium

"Does Inappropriate Use Explain Small-Area Variations in the Use of Health Care Services?" A Critique

Gestur Davidson

In a recently published article, a group of researchers (Leape, Park, Solomon, et al. 1990) report on a study they undertook to determine empirically whether Medicare enrollees' utilization rates for three procedures (coronary angiography, carotid endarterectomy, and upper gastrointestinal tract endoscopy) were significantly related to the degree to which these procedures were judged to be inappropriately performed, as determined by the appropriateness review program the researchers devised.

This is an important issue for public policy in general as well as for many private third party payers. If it were in fact found that a major cause of variation in overall use rates per capita of these and other procedures was the degree to which these procedures were performed inappropriately, then this relationship could be used to predict high inappropriate rate areas, which are not observable without expensive appropriateness reviews, from high overall use areas, which are observable. That is, this regularity, if true, could be used to focus an appropriateness review program more cost-effectively. Conversely, if this regularity did not in fact obtain, then overall use rates would not serve as good predictors of inappropriate use rates, and such rates would not be useful as a cost-effective focusing tool. Moreover, and perhaps more importantly, the absence of this regularity would call into question the very basis for the major use of small-area analysis (SAA) by third party payers, since its use rests on the untested assumption that a major determinant of variation in overall use rates is the extent

Address correspondence and requests for reprints to Gestur Davidson, Ph.D., Simulation Resource, University of Minnesota, Box 511 UMHC, Minneapolis, MN 55455.

of inappropriate use. Clearly, in the absence of this regularity it is possible to identify high levels of inappropriate use only through applying expensive appropriateness review programs, since anything else would simply be a guess and thus would yield average inappropriate use plus or minus a possibly large random component.

DESCRIPTION OF THE EMPIRICAL TEST OF THE HYPOTHESIS

Previously, these researchers reported that across three states they found no systematic relationship between the overall use rate per Medicare enrollee and the *proportion* of these same three reviewed procedures judged retrospectively to have been inappropriately performed (Chassin, Kosecott, Park, et al. 1987). Of course, given the very small sample size for this comparison ($n = 3$), no useful policy conclusions could be drawn from this earlier work.

To counter this methodological limitation, these researchers went back to their original data set and for a single state computed the overall use rates of these three procedures among the Medicare population in each of 23 counties in the state. They also calculated the proportion of each county's reviewed cases that were determined to have been inappropriately performed. They thereby increased their sample size from 3 to 23. This, however, came at a price. Since the empirical work for this study (i.e., the abstracting of hospital/medical records) had been completed years ago, their dropping to a county level of analysis now meant that they were simply dividing the same number of total reviewed procedures for this state among these 23 counties. As might be expected, the number of cases reviewed in some of the counties was quite small. (Reviewed cases per county ranged from 0 to 88 for coronary angiography, 0 to 127 for carotid endarterectomy, and 0 to 118 for upper gastrointestinal tract endoscopy.) Indeed, in some counties the overall use rate per Medicare enrollee (determined from Medicare claims data) was zero. As noted in the next section, when this occurred, the *proportion* of reviewed cases found to be inappropriate was not defined; each such occurrence effectively reduced their sample size by one.

Leape et al. used correlation analysis to test empirically the hypothesis that high overall use is importantly driven by high inappropriate use. Because of the problem that relatively few cases were

reviewed for some counties, the conventional use of correlation analysis—that is, ordinary least squares (OLS)—would be expected to result in estimates of the correlation coefficient that would likely vary more from sample to sample than would be the case if all counties had been sampled at a uniformly high level, and thus provide a less reliable test. Consequently, these researchers used an estimation technique (weighted least squares) to compensate for the varying reliability of the proportion of inappropriate cases.

The results of their use of correlation analysis on the relationship between overall use rates and the percentage inappropriate for these three procedures showed that for only one of these procedures, coronary angiography, was the hypothesized positive relationship supported by their empirical test. Specifically, for coronary angiography the correlation coefficient between overall use and percent inappropriate was .53, a result that could be expected to be found by chance alone (i.e., when the true $r = 0$) only once in every 175 samples of size 22. Moreover, if the percent inappropriate and equivocal were summed, the correlation between overall use rate and the combined inappropriate/equivocal percent was larger, $r = .63$. For both carotid endarterectomy and upper gastrointestinal tract endoscopy the empirical findings of this study do not allow one to reject the null hypotheses of no relationship between these two variables.

SUMMARY CRITIQUE OF LEAPE ET AL. METHODOLOGY

The way in which the Leape group conducted their empirical test of the hypothesis—that the overall use rate per Medicare enrollee of the three studied procedures is positively related to the degree to which these procedures are inappropriately performed—has four major methodological shortcomings.

First, because their test of this hypothesis was performed on an empirical data base of only 21 or 22 observations, it does not possess high enough levels of statistical power to provide an adequate test of the hypothesis. For example, if the "true" or population correlation coefficient between overall use and the percent inappropriate were thought to be .30—which as a rule of thumb is said to be a medium association (Cohen 1977)—and if erroneous rejection of the null hypothesis would be tolerated no more than one time out of 20, then any repeatedly drawn samples from a population of 21 counties would lead to the conclusion that a relationship existed between overall use

and percent inappropriate in only 39 percent of those samples (Cohen 1977). This is the probability of concluding that a positive association exists between overall use and the extent of inappropriate use for a *single* procedure. Leape et al. present the results of their test of this hypothesis for three procedures. Since all of their three tests are performed with data from the same counties, however, it cannot be safely assumed that they constitute three independent tests of the study hypothesis.[1] Consequently, it is not possible to determine a priori the likelihood of achieving the "combined Leape et al. result" of only one of three procedures exhibiting a positive association.

While one cannot make an overall a priori statement about the likelihood of achieving the Leape et al. result with their three trials,[2] it is nevertheless important for purposes of judging the policy utility of these results to examine the statistical power of a single test of their hypothesis. Four power values are provided in Table 1; they correspond to two values for the "true" r (.30 and .50), which are considered to be a medium and large association, respectively (Cohen 1977), and corresponding to two values for Type I errors (.01 and .05). All of these power values assume an n of 21, and they correspond to a one-tailed test of the hypothesis, the appropriate test in this application.[3]

There is, of course, no "right" combination of values for Type I error tolerance and the strength of the "true" relationship. Thus one cannot a priori determine a unique likelihood of committing a Type II error even for a single test. But the combination of $r = .30$ and Type I error = .05 might be considered a "fair" one with which to criticize this study. That is, with a certain evenhandedness one could say that Leape et al. could have been expected to conclude that no association existed based on a single procedure 61 percent of the time if the "true" level of association were .30. Clearly, this does not provide a strong case for concluding that "little of the variation in the rates of use of these procedures can be explained by inappropriate use" (p. 669).

A *second,* serious methodological limitation of this study is the way these researchers measure or express mathematically the inappropriate performance of these study procedures. Specifically, they express inap-

Table 1: Statistical Power of Test $R = 0$ for Different Values of Alpha and "True" R

"True" Correlation Coefficient Value	*Alpha Values*	
	.05	*.01*
.30	39%	16%
.50	77%	52%

propriate performance as a *percent* of total reviewed cases and not as a *rate* per capita. Under quite reasonable assumptions, the use of percent inappropriate could be expected to bias their empirical test of the study hypothesis in the direction of finding no relationship between overall use rate and inappropriate procedure performance.

To understand this methodological limitation as well as the next one, it is helpful to introduce the following mathematical identity or definitional relationship:

$$\frac{Overall\ Use}{POP} = \frac{N^A P^A}{POP} + \frac{N^E P^E}{POP} + \frac{N^I P^I}{POP}$$

where

$\frac{Overall\ Use}{Use}$ = the total number of cases of a procedure performed per age/sex-adjusted population in an area;

$\frac{N^A P^A}{POP}$ = the number of uses of this procedure arising from "appropriate indications" per age/sex-adjusted population, which is the product of:

- $\frac{N^A}{POP}$ = the number of people who present at a doctor's office in the area with an "appropriate indication" per age/sex-adjusted capita, and
- P^A = the proportion of the time doctors in the area perform the procedure when an "appropriate indication" is presented;

$\frac{N^E P^E}{POP}$ = the number of uses of this procedure arising from "equivocal indications" per age/sex-adjusted population, which is the product of:

- $\frac{N^E}{POP}$ = the number of people who present at a doctor's office with an "equivocal indication" per age/sex-adjusted capita, and
- P^E = the proportion of the time doctors in the area perform the procedure when an "equivocal indication" is presented;

$\frac{N^I P^I}{POP}$ = the number of uses of this procedure arising from "inappropriate indications" per age/sex-adjusted population, which is the product of:

$\frac{N^I}{POP}$ = the number of people who present at a doctor's office with an "inappropriate indication" per age/sex-adjusted capita, and

P^I = the proportion of the time doctors in the area perform the procedure when an "inappropriate indication" is presented.

From the perspective of policy relevance, one would like to know how much of the variance in the inappropriate use per capita is accounted for by the variance across those areas in the overall use per capita. If the answer is "a lot," then it is possible to "predict" with reasonable accuracy areas of high inappropriate use per capita by areas with high overall use. It is important to emphasize that policy is interested in identifying high *absolute* levels of inappropriate use per age/sex-adjusted population, which may be quite imperfectly measured by *relative* or percentage levels of inappropriate use.[4]

For example, compared to the average for a state, one area's (e.g., county's) inappropriate rate per capita may be high, but if that area's appropriate use and equivocal use rates per capita are even higher relative to their statewide means, then the percentage of inappropriate care may be below the statewide average. Conversely, inappropriate use per capita may be lower than the state mean, but if appropriate and equivocal use are even lower relative to their statewide means, percent inappropriate may be higher than the statewide mean.

Clearly, the relative measurement of inappropriate care has the potential to mislead. Moreover, and from the perspective of this empirical test of the study hypothesis, such a way of measuring (relative) inappropriate use could easily bias the estimated degree of association between inappropriate and overall rates of use. Specifically, compared to the correlation coefficient obtained between the overall procedure use per capita and the inappropriate rate per capita, the correlation coefficient between overall use per capita and percent inappropriate can be biased toward no association. It can easily be shown to be so as the examples provided in this critique would suggest, when the covariances of the per capita appropriate and inappropriate and the covari-

ance of the per capita equivocal and inappropriate use are both positive (Kendall 1969).

It remains, then, to determine whether it is plausible for these covariances to be empirically positive. This would seem to be quite reasonable. First, one can, as was done in the identity equation provided earlier, "decompose" all three of the component use rates per capita into their two respective parts. Considering the first set of components first, it can be argued that COV(N^A/POP, N^I/POP) and COV(N^E/POP, N^I/POP) are at worst zero, and much more likely to be positive. All of these terms might broadly be called "access factors." That is, they are all likely to be functions of three broad sets of variables: age/sex distribution in an area; the health status of individuals, and a third category including health insurance coverage, health beliefs, and income, all of which determine the propensity to seek care for any given health status. It thus seems likely that these two "first component" covariances would be positive. And the covariance among the second components, that is, COV(P^A,P^I) and COV(P^E, P^I), are, if anything, more likely to be positive. That is, if doctors in one area are predisposed to recommend a high proportion of inappropriate care relative to a statewide mean, it seems quite reasonable that they will have higher than average levels for P^E and P^A as well.

Given this plausibility, the correlation coefficient between overall use per capita and percent inappropriate is biased downward as an estimate of the correlation between overall and inappropriate use per capita. And again, this is important because policy interest focuses largely on this absolute rate of inappropriate use per capita.

A *third* methodological limitation of this study concerns the omission from the model of additional variables that are very likely to be important determinants of the overall use rate and the rate of inappropriate procedure performance per capita. When using regression/correlation analysis to estimate the quantitative relationship between one variable (dependent variable) and several additional covariates, it is important to include all covariates that a theoretical model suggests influence the dependent variable. If all such variables are not included, the estimated regression/correlation coefficients of the covariates that were included can be biased (Maddala 1977). Indeed, they will be biased unless the included and excluded covariates are statistically independent.

The importance of this omission is more clearly appreciated by observing that policy is particularly interested in how overall use rates per capita are related to just the P^I component of the "full" rate of inappropriate procedure performance per capita. That is, doctors can-

not and should not be held responsible for the rate at which patients present themselves at doctors' offices with inappropriate indications (i.e., the N^I/POP or access factor). The decision to perform the procedure with an inappropriate indication is clearly the doctor's and patient's, however, and is thus a concern of policy. Of course, although P^I and N^I/POP are conceptually distinguishable, one cannot empirically identify and measure them individually; even with an appropriateness review program one observes only the procedures *proposed,* or the product of P^I and N^I/POP. Although these two components are not individually measurable, if one is really interested in obtaining a good estimate of the relationship between P^I and overall use per capita, one can help the cause by controlling for the foregoing factors that are likely to account for variation in N^I/POP, namely, age distribution, health status, and those factors influencing the propensity to seek care such as health insurance. Of course, Leape et al. do directly control for age-distributional differences by age/sex-adjusting their overall use rate. However, in the Medicare population one might expect substantial variation in health status beyond that accounted for by age/sex. And although their data pertained only to the Medicare population, variation in Medicare supplemental insurance coverage and income levels could also be expected to vary substantially across a state. Health status data are not easy to come by and neither are Medicare supplemental insurance coverage data. But both sets of variables clearly belong in a fuller specification of the model explaining the relationship between overall use per capita and inappropriate use per capita. Given this, their exclusion from the Leape et al. model is likely to have imparted additional bias to their estimated correlation coefficients, although it is not possible to determine a priori the magnitude or even the direction of that bias.

Finally, the *fourth* methodological limitation of this study is the presence of measurement error in both the overall procedure use rate and the percent inappropriate variables—measurement error that the authors acknowledge. Random error, in part contributed by measurement error, is of course assumed to be an inherent characteristic of regression/correlation analysis. The authors' greater concern for the varying size of the error (heteroscedasticity) in their percent inappropriate variable over their sample data points (counties) led them implicitly, at least, to assume the percent inappropriate variable as their dependent variable and to use weighted least squares rather than OLS. (As weights they used the square root of the number of cases that were used to estimate percent inappropriate.) But significant amounts of measurement error in an *independent* variable—in their case, implic-

itly, the overall use rate—can give rise to significant bias in an estimated regression/correlation coefficient (Maddala 1977; MacMahon and Peto 1990). While neither the size nor direction of this bias can be determined with certainty a priori, under perhaps plausible enough conditions the measurement error in the overall use variable can be expected to bias their correlation coefficients toward zero.[5]

RECOMMENDED APPROACHES AND SUMMARY

The extent to which variation in the use of procedures over geographic areas is accounted for by variation in their inappropriate performance is an important issue for policy at all levels. Certainly, its importance clearly calls for the best empirical investigation we can devise. For the reasons we have indicated, we believe that the empirical test of this policy concern, as provided by Leape et al., falls short of this desired standard.

In an effort to contribute to the achievement of that standard in future empirical work, we offer our thoughts on how to overcome the limitations of the test presented by Leape et al. First and most important, a larger number of small areas must be sampled to achieve an acceptable level of statistical power for the test. Related to this, the small areas themselves should be drawn up in as meaningful a way as possible for this test. Defining small areas as counties is unlikely to capture the full amount of meaningful variation in either rate. Algorithms for delineating hospital market areas can be relatively easy to implement and would undoubtedly provide a more meaningful test of this question (Caper 1988). Further, and for its obvious policy utility, the small areas should have a wide geographic dispersion—wider than a single state.

Second, inappropriate procedure use (i.e., procedure use associated with an inappropriate indicator in an appropriateness review program) should be measured as a rate per age/sex-adjusted population. Appropriateness reviews are likely to be performed on a sampling basis, and even if not, inappropriate use could have quite large chance variability and hence low reliability due to the small population sizes of individual age/sex strata. While Leape et al. address the problem of heteroscedasticity with weighted least squares, this may be too extreme an adjustment (Pocock, Cook, and Bevesford 1981). In conjunction with multiple regression or a multiple logistic model, an alternative,

intermediate weighting scheme would seem more desirable (Pocock, Cook, and Bevesford 1981).

As noted previously, what is particularly important for policy is the relationship between overall use per capita and the decision-making process of an area's physicians concerning use of a procedure, summarized by the empirically unobserved variable P^I. This relationship would be more precisely estimated if factors are controlled for—factors that are important in determining the per capita rate at which individuals with inappropriate indications present themselves to physicians. These latter factors are likely to include the extent of health insurance and indicators of health status. In future empirical work, the effort should be made to obtain data on these covariates. Finally, by acquiring data from procedure use and from appropriateness reviews for more than a single year, an instrumental-variable estimator could be used that would likely reduce the bias due to measurement error.

APPENDIX

Consider the following simple cost model of a one-procedure appropriateness review program.

Let

CRA = the unit (average) cost of conducting a minimum review of the procedure, that is, the average cost of reviewing a request that is accepted;

CRD = the unit (average) cost beyond *CRA* of reviewing the procedure when it is denied;

B = the expected net unit (average) benefit associated with denying a request to perform the procedure when it is rated inappropriate;

I/POP = the per capita rate of inappropriate indications reviewed for the procedure; and

USE/POP = the overall per capita use of the procedure.

Then if

CS = the per capita cost savings from operating an appropriateness review program,

$$CS = B * I/POP - CRD * I/POP - CRA * USE/POP$$

$$= [B - CRD] * I/POP - CRA * USE/POP$$
$$> 0 \text{ if } I/USE > CRA/[B - CRD].$$

NUMERICAL EXAMPLE

Let

CRA = *$50*
CRD = *$500*
B = *$2500*

	Use POP	*Appropriate Use POP*	*Equivocal Use POP*	*Inappropriate Use POP*	*Percent Inappropriate*
State average	.13	.07	.02	.04	31
Case 1	.22	.12	.04	.06	27
Case 2	.08	.04	.01	.03	38

Then

CS = ($2000) * I/POP - $50 * Use/POP
State average = $73.50
Case 1 = $109.00
Case 2 = $56.00

As is apparent for this particular numerical example, cost savings per capita and percent inappropriate are inversely related.

NOTES

1. If they could be considered independent trials, then the cited power value of .39 could be used as the binomial parameter to determine a priori the probability of achieving the "combined Leape et al. effect" of no more than one positive association in three trials. In this particular case, the combined probability would be .66.
2. This would provide a very interesting and potentially very useful application of the data resampling technique known as the bootstrap. Specifically, simulation techniques could be used on a joint distribution of regression equation residuals from the three procedures to arrive at an empirical estimate of the statistical power of this combined three-procedure test of the study hypothesis.
3. Power for a two-tailed test would, of course, be lower yet: to attain the power of a one-tailed test at an alpha of .05 would necessitate accepting an alpha of .10 for a two-tailed test.
4. A simple cost/saving model easily illustrates that the net savings per capita from an appropriateness review program is a function of both the inappropriate rate per capita and the overall rate per capita. The latter determines the cost of reviewing every occurrence of the procedure. There is, in addition, a simple constraint that must be satisfied in order to make it worthwhile to enter an area at all, and that constraint does include the percent inappropriate. Nevertheless, using the assumptions made in the

text concerning covariances, it is easy to construct a simple numerical example to illustrate how focusing on relative inappropriateness can mislead policy from maximizing cost/savings per capita. This is shown in the Appendix to this critique.

5. Specifically, the measurement error must be uncorrelated with the non-error components of both variables and the regression equation error itself.

REFERENCES

Caper, P. "Geographic Variations in the Use of Health Care Services: A Comment." *Journal of the American Medical Association* 259, no. 13 (1 April 1988): 1947.

Chassin, M. R., T. Kosecott, R. E. Park, C. M. Winslow, K. L. Kahn, N. J. Merrick, J. Kessey, A. Fink, D. H. Solomon, and R. H. Brook. "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services? A Study of Three Procedures." *Journal of the American Medical Association* 258, no. 18 (13 November 1987): 2533–37.

Cohen, T. *Statistical Power Analysis for the Behavioral Sciences.* Orlando, FL: Academic Press, 1977.

Kendall, M. G., and A. Stuart. *The Advanced Theory of Statistics.* Vol. 1. New York: Hafner Publishing Company, 1969.

Leape, L. L., R. E. Park, D. H. Solomon, M. R. Chassin, J. Kosecoff, and R. H. Brook. "Does Inappropriate Use Explain Small-Area Variations in the Use of Health Care Services?" *Journal of the American Medical Association* 263, no. 5 (2 February 1990): 669–72

Maddala, G. S. *Econometrics.* New York: McGraw-Hill, 1977.

MacMahon, S., and R. Peto. "Blood Pressure, Stroke and Coronary Heart Disease: Part 1. Prolonged Differences in Blood Pressure: Prospective Observational Studies Corrected for the Regression Dilution Bias." *The Lancet* 335 (1990): 765–74.

Pocock, S. J., D. G. Cook, and S. A. A. Bevesford. "Regression of Area Mortality Rates on Explanatory Variables: What Weighting is Appropriate?" *Applied Statistics* 30, no. 3 (1981): 286–95.