Small Area Statistics: Large Statistical Problems

Variation in the quantities of health services used by inhabitants of small geographical areas has been examined by recent research studies,[1-7] including a paper by Roos in this issue of the Journal.[8] The usual analytic method is to calculate the utilization rates for a service in several areas, compare the largest rate to the smallest rate, note that the difference is large, and attempt (using multiple regression or t-tests) to explain this variability as a function of service availability, physician practice styles, etc. Along with the difficulties in interpreting such findings, there are many associated statistical problems which have been handled with increasing sophistication in recent articles, but may not be widely appreciated. These include:

- What are the rates actually estimating?
- Is there really a significant amount of variation among the areas?
- What is the best way to estimate the rates?

The first methodological problem is with the rates themselves: what exactly are they measuring? I will ignore the problem of judging whether the rates are too high or too low, and concentrate on the statistical properties of the estimates. The use rates are usually calculated as the number of procedures performed divided by the number of people in the small area.* The rates are analyzed further as though they represented the proportion of eligibles who received the procedure. In most cases, however, this interpretation is inappropriate. One problem occurs when an idividual can receive the procedure of interest more than once (e.g., surgery for varicose veins more than once,[1] several different types of surgery,[4] or several hospital admissions). In this case the ratio is not the proportion of people who received the procedure, since one person with many procedures could cause the rate to look high while a relatively low proportion of the residents actually had the procedure. Many of the statistical procedures used to test hypotheses about the rates also assume that the rates are proportions, which may yield results that are too liberal in assessing the statistical significance of variations among small areas.[9]

Even in a case where the procedure can occur at most once (e.g., hysterectomy) the denominator of the utilization rate may still be a problem. Clearly, the denominator should be the number of people eligible for this procedure, which should exclude, in the case of hysterectomy, women who have already had that procedure. Unfortunately, the number of women in an area who have already had a hysterectomy is rarely known, and all women in the small area are usually used in the denominator. This could provide anomalous results. Consider a hypothetical area which has such a high rate of hysterectomy that virtually all women have had a hysterectomy. Because of this, the number of hysterectomies in the year studied would be nearly zero, and the rate which used all women in the denominator would suggest that this small area had a very low rate of hysterectomy. Comparison of two areas where the fraction of women eligible for hysterectomy differed substantially could thus yield misleading results. Roos,[8] addresses this issue by including only women with a recent gynecological diagnosis (indicating the presence of a uterus) in the rate calculations. This might introduce some difficulties in the interpretation or the results, since the sample may be biased to include sicker women, but it should be applauded from a statistical point of view.

A major finding of most of the literature is that there is too much variability among the small areas, based on the difference between the highest and lowest rates. This may be an incorrect conclusion, however, since the highest rate is always, by definition, greater than the lowest rate, and the differences can be surprisingly large by chance alone. If the utilization rates can be thought of as observations from a normal distribution, the highest and lowest rates will differ (on average) by 2.3 standard deviations if five small areas are being compared, and by 3.7 standard deviations if 20 areas are being compared,[10] even if the underlying mean rate is the same in each small area. This means that if each area had 1,000 people, and the underlying procedure rate was 5 per 1,000 in all of the areas, the ratio of the largest rate to the smallest (known as the extremal quotient) would be approximately 3.2 for five small areas and 10.9 if 20 small areas were compared, by chance alone. The "chance" expected value of the extremal quotient decreases as the number of people in the small area increases, but it increases as the number of areas considered increases, and is larger for less frequent procedures. Chance variation may not be an issue in any particular analysis, but this possibility should always be addressed. A recent paper by Willemain illustrates this problem.[11]

This leads to the problem of how to test whether there is more variability among the areas than would be expected at random. Several methods have been used[1,12] which are based on the binomial properties of the estimates, that is, assuming that a person can receive the procedure at most once, or that procedures occur independently. As mentioned above, these assumptions are not always met, and may provide inappropriately liberal results (i.e., significant results where there is no underlying difference). One possible approach, if data are available, is to repeat key analyses using a true proportion, in which a person is not counted more than once, to see if the results change. More work is needed in this area.

Even if there is statistically significant variation, the size or importance of the differences among the small areas remains in question, since with large enough populations even tiny differences will produce a statistically significant result. Some research has provided estimates of the "systematic" variance of the rates, after accounting for the chance variation.[2,3,12-14] The variance estimate can be taken as a measure of the variability within one group of small areas, and can be compared across geographical regions, countries, or procedures, to see if there is more variability in one site than in another.

If significant and large differences are established, the next step is to explain the observed differences. Comparisons which use the small area as the unit of analysis, with the utilization rate as the dependent variable, may incur statistical problems. The major consideration is that the number of residents in each area varies, and with it the reliability of the calculated utilization rate for that area. Very small areas might appear to have excessively high rates because of a few

*Some analyses use instead the ratio of the observed number of procedures to the expected number of procedures. The statistical properties of these estimates are not substantively different from those of the use rate. Some related research, which compares rates among hospitals, rather than small areas, is included in this discussion.

additional procedures, while rates for larger areas will be more reliable. One possibility is weighted analysis, which places more emphasis on the rates from larger areas.[12] In addition, recent statistical methods have been introduced which provide better over-all estimates for groups of rates than is achieved by using only the area's observed rate to represent that area.[12–16] In effect these methods estimate the rate for an area as a weighted average of the average rate for all areas and the observed rate for that area, thus "shrinking" all of the observed rates toward the over-all mean. Smaller areas, for which the rate is more variable, are shrunk further toward the mean than are the larger areas. These estimates are not necessarily optimal estimates of any one rate, but over-all they have better characteristics than the unshrunken estimates, and can help to protect the researcher from spurious results caused by sampling variability.

Given the large number of unresolved, and sometimes unrecognized, problems in the statistical analysis of variations among small areas, it seems clear that future analyses should continue to address the methodological, as well as the substantive, issues in this field.

Paula Diehr, PhD

Address reprint requests to Paula Diehr, PhD, Associate Professor, Department of Biostatistics SC-32, University of Washington, Seattle, WA 98195.

REFERENCES

1. Lewis CE: Variations in the incidence of surgery. N Engl J Med 1969; 281:880–885.
2. McPherson K, Strong PM, Epstein AE, Jones L: Regional variations in the use of common surgical procedures: within and between England and Wales, Canada and the United States of America. Soc Sci Med 1981; 15A:273–288.
3. McPherson K, Wennberg J, Hovind OB, Clifford P: Small-area variations in the use of common surgical procedures: an international comparison of New England, England, and Norway. N Engl J Med 1982; 307:1310–1314.
4. Roos N: High and low surgical rates: risk factors for area residents. Am J Public Health 1981; 71:591–600.
5. Wennberg J, Gittelsohn A: Variations in medical care among small areas. Sci Amer 1982; 246:120–134.
6. Wennberg JE, Barnes BA, Zubkoff M: Professional uncertainty and the problem of supplier-induced demand. Soc Sci Med 1982; 16:811–824.
7. Wennberg J, Gittelsohn A: Small area variations in health care delivery. Science 1973: 18:1102–1108.
8. Roos N: Hysterectomy: variations in rates across small areas and across physicians' practices. Am J Public Health 1984; 74:327–335.
9. Diehr P: Statistical measures for admission rates. Seattle: Department of Biostatistics, University of Washington, Technical Report No. 24, 1978.
10. Dixon WJ, Massey FJ: Introduction to Statistical Analysis. New York: McGraw Hill, 1957.
11. Willemain TR: On the comparison of highest and lowest surgery rates in small-area studies. *In:* I Rothberg (ed): Regional variation in hospital use. Lexington MA: Lexington Books, 1982.
12. Williams RL: Measuring the effectiveness of perinatal medical care. Med Care 1979; 17:95–110.
13. Williams RL, Cunningham GC, Norris FD, Tashiro M: Monitoring perinatal mortality rates: California, 1970 to 1976. Am J Obstet Gynecol 1980; 136:559–568.
14. Stanford Center for Health Care Research: Comparison of hospitals with regard to outcomes of surgery. Health Services Res 1976; 11:112–127.
15. Efron B, Morris C: Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 1975; 70:311–319.
16. Stanford Center for Health Care Research: Staff study of institutional differences in postoperative mortality. Springfield, VA: National Technical Information Services, 1974; PB 250 940/LK.

ERRATUM

In: Christoffel KK: Homicide in childhood: a public health problem in need of attention. Am J Public Health 1984; 74:68–70. Figures 1 and 2 (p 69) lacked sufficient black and white contrast to show the decimal points in the numbers along the vertical axes, when the Figures were reduced in size for use in the Journal.

For clarification, in Figure 1: the Homicide, Death Rates/100,000 column should read in descending order 7.00, 6.00, 5.00, 4.00, 3.00, 2.00, 1.00, and 0; the Homicides as Percent of all Deaths column should read in descending order 7.0, 6.0, 5.0, 4.0, 3.0, 2.0, 1.0 and 0.

In Figure 2: Homicides, Death Rates/100,000 column should read 2.82, 2.46, 2.10, 1.74, 1.38, 1.02, 0.66, and 0.30; Homicides as Percent of all Deaths column should read 6.5, 5.6, 4.7, 3.8, 2.9, 2.0, 1.1, and 0.2.