What Is Too Much Variation? The Null Hypothesis in Small-Area Analysis

Paula Diehr, Kevin Cain, Frederick Connell, M.D., and Ernest Volinn

A small-area analysis (SAA) in health services research often calculates surgery rates for several small areas, compares the largest rate to the smallest, notes that the difference is large, and attempts to explain this discrepancy as a function of service availability, physician practice styles, or other factors. SAAs are often difficult to interpret because there is little theoretical basis for determining how much variation would be expected under the null hypothesis that all of the small areas have similar underlying surgery rates and that the observed variation is due to chance. We developed a computer program to simulate the distribution of several commonly used descriptive statistics under the null hypothesis, and used it to examine the variability in rates among the counties of the state of Washington. The expected variability when the null hypothesis is true is surprisingly large, and becomes worse for procedures with low incidence, for smaller populations, when there is variability among the populations of the counties, and when readmissions are possible. The characteristics of four descriptive statistics were studied and compared. None was uniformly good, but the chi-square statistic had better performance than the others. When we reanalyzed five journal articles that presented sufficient data, the results were usually statistically significant. Since SAA research today is tending to deal with low-incidence events, smaller populations, and measures where readmissions are possible, more research is needed on the distribution of small-area statistics

Address correspondence and requests for reprints to Paula Diehr, Ph.D., Professor of Biostatistics and Health Services, Department of Biostatistics, SC-32, School of Public Health and Community Medicine, University of Washington, Seattle, WA 98195. Kevin Cain, Ph.D. is Assistant Professor, Department of Biostatistics; Frederick Connell, M.D. is Associate Professor, Department of Health Services; and Ernest Volinn, Ph.D. is Research Assistant Professor, Department of Neurosurgery, all at the University of Washington. Dr. Cain was supported in part by USPHS grant number R01-CA 40644.

under the null hypothesis. New standards are proposed for the presentation of SAA results.

Small-area analysis (SAA) is a popular methodology in health services research. A typical SAA might calculate the utilization rate for a service (we will refer to this as a type of surgery) in several small areas (we will call them counties), compare the largest rate to the smallest, note that the difference is large, and attempt (using multiple regression or *t*-tests) to explain the variability in surgery rates as a function of service availability, physician practice styles, and other variables of interest. Such analyses have generated many hypotheses for further exploration. A large number of SAA studies have been performed, and several review articles have been written on this topic (Copenhagen Collaborating Center 1985; Paul-Shaheen, Clark, and Williams 1987; *Health Affairs* 1984).

The underlying hypothesis that there is "too much" variability among the counties is rarely tested formally in such research, partly because the reported variation appears large, and partly because there are no good ways to test it. The null hypothesis in this case is that the underlying surgery rate is the same in all the counties, and that the observed differences among the counties are due simply to random variation. Information is needed about the distribution of rates in this null situation to help investigators decide whether the observed variability is larger than would have been expected by chance alone.

Few statistical methods exist that allow us to assess such variability. Some tables are available on the order statistics of the standard normal distribution (Dixon and Massey 1957; Sarhan and Greenberg 1962). These give the mean and standard deviation of, for example, the highest observation of a set of five drawn from the standard normal distribution. (Students are often surprised to learn that the expected value of the maximum is not zero.) Based on these tables, the highest and lowest observations will differ, on average, by 2.3 standard deviations if five observations are drawn, and by 3.7 standard deviations if 20 observations are drawn. The difference increases with the number of observations being ordered.

In theory, if the populations are large enough, we can assume that the observed rate in each small area has a normal distribution, with a common mean and standard deviation, and the order statistic tables can be used. In practice, however, the order statistics are not very useful for several reasons. First, they are tabled only up to $N = 20$. Second, they require that each observation be drawn from the same distribution. If p is the underlying probability of surgery in a county with n_i people, the variance of that rate will be $p(1 - p)/n_i$. The variance can only be the same for all counties if they have (approximately) the same populations. This is often not the case. Finally, a statistic that is often used in SAA is the extremal quotient (EQ), the maximum rate divided by the minimum rate. The distribution for this statistic has not been tabled (primarily because its expected value is infinity), although tables that deal with some cases of interest—large, similar-sized counties—recently appeared in this journal (Kazandjian, Durance, and Schork 1989). There are also no tables available for two other SAA statistics, the coefficient of variation and the systematic component of variation. Without such tables, these descriptive statistics cannot be used to test the null hypothesis.

A statistical method that is sometimes appropriate is the 2 × k chi-square test (for k counties), which tests the null hypothesis that the surgery rate is the same in each county (Chassin, Brook, Park, et al. 1986). This is appropriate when a person can be counted in the numerator at most once and when the expected number of surgeries per county is at least five. (It is not appropriate when readmissions are possible, or for surgeries with low incidence.) This technique may be underused because it does not apply directly to age/sex-standardized rates, or because it does not provide direct estimates of the expected variability among counties. Other hypothesis-testing methods that are appropriate in some situations include analysis of variance and multiple regression; these are described in the discussion section.

SAA is a well-accepted methodology, which is being embraced by the health services research community and used in ways not foreseen by its formulators. While early small-area analysts used relatively "large" small areas, in which it could be assumed that detected variations were meaningful, the popularity of the technique and the availability of related software have encouraged investigators to apply this technique to extremely small areas. The early research was on procedures such as "ectomies," or organ removal, where a person can be in the numerator at most once. This means that the surgery rate is a proportion, which has known statistical properties. Mortality rates can also be thought of as proportions. Current researchers, however, are extending the SAA method to procedures where readmissions are possible (i.e., the same person can be counted more than once in the numerator). Rates based on these variables are not proportions, and do not have known theoretical distributions. Researchers are also interested in the variation of other nonbinary measures such as length of stay, cost of an admission, disease prevalence, or average number of fillings. The definition of a "small area" may now be a hospital or a dental practice as well as the more traditional geographical or medical service area.

Because of the popularity of SAA, and the potential for new uses of this technique, we believe that it is time for a technology assessment of SAA. While many issues need to be assessed (Diehr 1984), we will limit the scope of this research to the problem of determining when the observed variability among small areas is statistically significant (greater than expected under the null hypothesis), and to issues related to this question. This is, of course, the first step that must be taken in a small-area analysis.

METHODS

We wrote a computer program to simulate the type of surgery rates that would be observed when a given number of counties of varying sizes all have the same underlying surgery rate. The simulation program is described schematically here, with more detail in Appendix A. Let the underlying proportion of individuals who had surgery be p for each of k counties. (In the following, $k = 39$.) In any particular year, the observed proportion of people in a county who undergo surgery will vary and these proportions will tend to be normally distributed, with mean p and variance $p(1 - p)/n_i$, where n_i is the population of county i. (We are assuming homogeneous populations; see the section on age and sex adjustment further on.) Given p and n_i, we can calculate the mean and variance of the rate for each county. A flowchart for the simulation is shown in Figure 1. To simulate a set of surgery rates, we generate a random number from the appropriate normal distribution for each county (normal with mean p, variance $p(1 - p)/n_i$), resulting in 39 "surgery rates" that might have been observed in a particular year by an investigator. Although all of these rates were drawn from distributions with the same mean, the observed rates will not be the same for every county. They will vary, and the amount of variation will be larger for smaller values of n_i. We calculated the maximum and minimum of the 39 numbers, as well as the extremal quotient (maximum/minimum, or set to "missing" if the minimum = 0); a chi-square statistic for the $2 \times k$ table (surgery yes/no, by county); the coefficient of variation (CV), defined below; and the systematic component of variation (SCV), also defined below.

This process was repeated 3,000 times, yielding 3,000 different simulated samples from the underlying distribution, and 3,000 different values of the maximum rate, minimum rate, maximum/minimum

Figure 1: Schematic Diagram of Simulation

```
Read in number of counties (39), county sizes $n_i$, $p$

Calculate standard deviation for each county, $s_i = (p(1 - p)/n_i)^{.5}$

For each trial ($t = 1$ to 3000)
    For each county ($i = 1$ to 39)
        Generate random number from normal distribution;
          $Y_i$ has mean 0, variance 1
        Compute $X_1 = Y_i * s_i + p$
    Compute $A_t$ = maximum of the 39 $X$s
    Compute $B_t$ = minimum of the 39 $X$s
    Compute $C_t$ = EQ = $A_t / B_t$ and other statistics

Calculate mean and standard deviation of 3000 values of $A$, $B$, and $C$, etc.

Print line on simulation table.
```

= extremal quotient (EQ), and the other statistics. We calculated the mean and variance of the minimum, the maximum, and the EQ from those 3,000 values, as well as the 95th percentile of the EQ, chi-square, CV, and SCV. This entire process was repeated for a range of values of *p*. We varied the population sizes by various factors to simulate the effect of studying, say, only females or the elderly. We also allowed the possibility of readmissions (a person appearing in the numerator of the surgery rate more than once) as explained in Appendix A. Finally we looked at the effect of studying more or fewer counties.

As an example we took the 39 counties of the state of Washington. The population of these counties varies from about 2,600 to 1.3 million, clearly violating any assumption of equal population size. The mean population per county is 114,000. Eight counties have fewer than 10,000 residents, 22 have from 10,000 to 100,000, eight have from 100,000 to 500,000 and the largest (which contains Seattle) has a population of 1.3 million. Surgery rates computed in the large counties will be relatively stable, but those in the smallest will tend to be quite variable from year to year, since the addition of one or two additional surgeries in such counties will increase the rates dramatically. The county populations are shown in Appendix B.

RESULTS

Some simulation results are shown in Tables 1–8. Each line on Tables 1–4 is the result of 3,000 iterations. Tables 5–8 used 1,000 iterations.

EXTREMAL QUOTIENT (EQ)

The EQ is the ratio of the highest observed rate to the lowest rate. It is infinite if the lowest rate is zero, which can happen with high probability if some of the counties are small and the surgery rates are low. In our simulation we have dealt with this problem by excluding those simulations in which EQ is infinite. Table 1 shows the mean and standard deviation of the components of the EQ for various surgery rates, assuming that all people in the state of Washington are eligible for the surgery. Since the distribution of the extremal quotient has a very long right tail, we also estimated the 95th percentile of its distribution. We first looked at the hypothetical condition in which each of the 39 counties had the same population, 114,000. This provides a "best case" condition for comparison, and is similar to a situation for which tables are available (Kazandjian, Durance, and Schork 1989).

The first line of Table 1 shows simulation results for an underlying rate of 50 surgeries per 100,000 ($p = .0005$). Although the underlying rate is 50, the average value of the smallest rate was 35.67, and of the largest it was 64.21. Note that the distributions of the minimum and maximum value have means that are symmetric about 50, and that the standard deviations of the two measures are approximately equal. The average value of the extremal quotient is 1.82, with a standard deviation of 0.20. The 95th percentile of the EQ is 2.19, which is about two standard deviations above the mean. The mean and standard deviation of the minimum and maximum observations increase with the surgery rate. The mean and standard deviation of the extremal quotient decrease as the rate increases, as does the 95th percentile.

An investigator who had data in which the observed average sur-

Table 1: Simulation of Minimum, Maximum, and Extremal Quotient (EQ) with all 39 County Populations Set to 114,000 (100 Percent of Population, No Readmissions)

	Minimum		*Maximum*		*EQ**		
Rate per 100K	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*95%*
50	35.67	3.26	64.21	3.16	1.82	0.20	2.19
100	79.76	4.53	120.04	4.50	1.51	0.10	1.71
250	218.47	7.04	281.58	7.25	1.29	0.05	1.38
500	455.20	9.81	545.05	10.08	1.20	0.03	1.26
1000	937.11	14.17	1062.70	13.75	1.13	0.02	1.17
2500	2400.32	21.94	2599.32	22.29	1.08	0.01	1.11
5000	4861.78	30.94	5138.37	30.70	1.06	0.01	1.07
10000	9808.92	42.49	10190.72	42.73	1.04	0.01	1.05

*EQ is not computed when the minimum is zero.

gery rate was 50 per 100,000 might compare the observed extremal quotient to the data in Table 1. Under the null hypothesis, the 95th percentile of the EQ is 2.19. An observed extremal quotient larger than 2.19 thus indicates more variability than expected by chance alone. We could reject the null hypothesis, and further analysis would be in order. An observed EQ of, say, 1.5, would not be statistically significant, and the SAA might stop at this point. An EQ of 2 or more would be statistically significant for any surgery rate in Table 1. In general, EQs that "look" large *are* significant for the experimental conditions of Table 1.

When the lowest observed rate is zero, the EQ is infinite. The tabled values of EQ are only for noninfinite values. When the lowest observed rate is zero, the investigator might instead compare the observed maximum to the expected maximum value of Table 1. A maximum value 1.645 standard deviations above its expected mean (above 64.21 + 1.645(3.16) = 69.41) would also suggest more variability than expected by chance alone. Other test statistics—for instance, (maximum-minimum)/median—could also be considered. At this time, the relative power of various possible tests for "too much" variability has not been studied.

Table 2 shows results from a simulation in which the actual population of each county was used. These varied from 2,618 to 1.3 million, as shown in Appendix B. The results for an incidence rate of 50 (first line of Table 2) are strikingly different from those of Table 1. The expected minimum value is 9.09, and the expected maximum value is 94.23, even though all counties have the same expected rate of 50. The expected value of the EQ is 6.23, with a standard deviation of 8.62,

Table 2: Simulation of Minimum, Maximum, and Extremal Quotient (EQ) Using 39 Actual County Populations (100 Percent of Population, No Readmissions)

	Minimum		*Maximum*		*EQ**		
Rate per 100K	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*95%*
50	9.09	10.61	94.23	20.64	6.23†	8.62	11.25
100	38.25	20.70	160.72	27.28	5.62	27.56	8.60
250	147.01	42.12	343.07	37.78	2.92	6.05	4.99
500	351.24	61.85	630.79	53.60	1.88	0.76	2.70
1000	794.74	86.00	1182.87	75.67	1.51	0.22	1.90
2500	2177.26	137.01	2793.43	122.29	1.29	0.10	1.48
5000	4547.87	188.96	5404.54	163.45	1.19	0.06	1.31
10000	9375.43	272.75	10552.85	225.56	1.13	0.04	1.20

* EQ is not computed when the minimum is zero.

† The EQ was infinite (lowest rate was 0) more than half the time.

and a 95th percentile of 11.25. The investigator would have to see 11-fold differences before statistical significance could be claimed! The values in Table 2 are closer to those in Table 1 for higher surgery rates, but still differ substantially. These results show that the amount of variability among counties can be very high, even when there is no underlying difference in surgery rate among the counties. The 95th percentile of the EQ is lower than the mean plus 1.645 standard deviations because EQ has a long right-tailed distribution. Because of this, we will discuss only the 95th percentile of the EQ as a critical value. The estimated mean and standard deviation of the EQ are presented only for reference.

Table 3 shows results if the population for each county is divided by 2, which would occur if, for example, only males were being studied. The expected values, standard deviations, and percentiles are even higher in this table than in Table 2. Thus, the EQ is sensitive to the number of people in each county, as well as to the distribution of people among counties.

One assumption underlying the simulation model is that a person can be in the numerator once at the most. This is probably true for "ectomies," since an organ can be removed at most once (although a person might still be in the data base more than once if that person was readmitted for complications, or if multiple bills were submitted for the same procedure). This assumption is violated, however, when hospital admission rates for a particular *diagnosis* are analyzed. Readmission rates for many diagnoses run as high as 50 percent, and the average

Table 3: Simulation of Minimum, Maximum, and Extremal Quotient (EQ) Using 39 Actual County Populations (50 Percent of Population, No Readmissions)

	Minimum		*Maximum*		*EQ**		
Rate per 100K	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*95%*
50	1.51	5.25	115.28	32.03	‡	‡	‡
100	17.61	21.18	189.31	42.72	6.63†	11.83	12.05
250	108.62	49.52	381.12	55.33	4.33	7.30	7.88
500	295.51	83.08	685.00	73.78	3.28	16.33	4.93
1000	708.90	121.01	1262.56	104.01	1.86	0.72	2.66
2500	2048.12	191.00	2909.89	169.42	1.43	0.17	1.76
5000	4359.60	272.03	5571.86	241.20	1.28	0.10	1.47
10000	9131.65	368.35	10787.55	324.50	1.18	0.06	1.30

* EQ is not computed when the minimum is zero.

† The EQ was infinite (lowest rate was 0) more than half the time.

‡ The EQ was infinite more than 90 percent of the time, and is too unstable to present.

number of admissions of all types per person hospitalized is 1.5 in a year (U.S. Dept. of Health and Human Services 1983). To examine the effect of readmissions on SAA statistics, we made a simple assumption (for this example) that the probability of a second admission (conditional on the first admission) is .5, but the probability of a third admission is 0, adjusted to make the overall admission rate the same as in previous examples. (Details are in Appendix A.) The results of a simulation that includes readmissions are in Table 4. The variability in all the statistics is much higher here than in Table 2, where the observed surgery rates are the same but there are no readmissions. The 95th percentiles are as high as 28. Thus, readmissions can have a large effect on the variability of small-area statistics under the null hypothesis.

From this point on we use a different format for the simulation results, presenting only the estimated 95th percentiles of the descriptive statistics, based on 1,000 iterations. The first four columns of Table 5 contain the 95th percentiles of the EQ in the experimental conditions used for Tables 1-4. The tabled results differ slightly because a different number of iterations was used. The simulation results are accurate to about two or three significant digits.

Column 4 shows the results when 50 percent of those admitted were readmitted once. We next consider a situation in which only 10 percent were readmitted. This is shown in the fifth column of Table 5. The rates are not as high as those of Table 4, but are still substantially

Table 4: Simulation of Minimum, Maximum, and Extremal Quotient (EQ) Using 39 Actual County Populations (100 Percent of Population, 50 Percent Readmission Rate)

	Minimum		*Maximum*		*EQ**		
Rate per 100K	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*95%*
50	2.86	6.62	109.42	29.89	†	†	†
100	24.21	21.33	179.55	36.92	9.97	60.85	16.90
250	121.05	47.92	369.59	48.08	4.28	12.24	8.94
500	308.17	80.00	668.33	68.16	2.81	12.02	4.33
1000	734.22	112.43	1239.34	99.63	1.74	0.44	2.39
2500	2080.89	177.98	2874.77	151.55	1.39	0.15	1.68
5000	4408.56	245.96	5527.65	211.07	1.26	0.09	1.42
10000	9174.33	342.53	10727.18	296.50	1.17	0.06	1.27

* EQ is not computed when the minimum is zero.

† The EQ was infinite (lowest rate was zero) more than 75 percent of the time, and is too unstable to present.

Table 5: 95th Percentile of Extremal Quotient* (EQ) in Seven Experimental Situations for Eight Surgery Rates (Based on 1,000 Iterations per Number)

	1	*2*	*3*	*4*	*5*	*6*	*7*
Rate per 100K	*Same Pop.*	*Real Pop.*	*Male*	*50% Readmit.*	*10% Readmit.*	*Large Only*	*Same K = 10*
50	2.20	11.08†	‡	‡	14.53†	6.80	1.85
100	1.70	8.85	11.42†	23.53	11.74	3.14	1.53
250	1.40	4.65	7.75	9.79	6.08	1.99	1.30
500	1.26	2.62	5.70	4.20	3.04	1.57	1.22
1000	1.18	1.95	2.64	2.40	2.09	1.37	1.14
2500	1.11	1.47	1.77	1.66	1.52	1.22	1.09
5000	1.07	1.30	1.47	1.42	1.34	1.14	1.06
10000	1.05	1.20	1.30	1.29	1.22	1.10	1.04

Column:
1—All 39 counties have same population, 114,000.
2—Actual county populations used (2,618 to 1,344,586).
3—"Males Only" (50 percent of population used).
4—50 percent of people are readmitted once.
5—10 percent of people are readmitted once.
6—Counties with populations above 10,000 (31 counties).
7—Only ten counties, all with same population, 114,000.

* EQ is not computed when the minimum is zero.

† The EQ was infinite (lowest rate was zero) more than half the time.

‡ The EQ was infinite more than 75 percent of the time, and is too unstable to present.

higher than those of column 2 (or of Table 2). Even a relatively low readmission rate can cause serious distortions in the EQ distribution.

One way to deal with the excessive variability in the null situation is to eliminate some of the smallest counties. This is unsatisfying, since these might be the areas that had high underlying rates. However, it may be necessary in order to make the problem tractable. (Counties should be eliminated or combined based on their *expected* number of surgeries, not on the observed numbers.) Column 6 of Table 5 shows the results for Washington state if the eight counties with populations below 10,000 are removed. The values in this table are smaller than those of Table 2, although still considerably larger than those of Table 1. For example, the 95th percentile of the EQ for a rate of 50 is 11.08 for all 39 counties, but only 6.80 for the 31 largest counties. The value in the "ideal" case of Table 1 is 2.20. The chance variability in the EQ could be decreased still further by eliminating more counties, but this would make the resulting findings less and less generalizable. It is also not necessarily true that tests using this smaller subset of counties would have more power than tests using all counties.

Column 7 shows the effect of varying the number of counties. The experimental condition is the same as that of column 1, but with only 10 counties instead of 39. Comparing column 1 with column 7 shows that the EQ is lower if there are fewer counties. (The fewer the observations, the less likely an extreme value.)

In summary, except for the situation represented in Table 1 (ectomies, similar-sized large counties, high expected values), the EQ can be very misleading, as apparently large values are not significantly different from what would be expected by chance alone.

CHI-SQUARE

A simple way to test for differences in the surgery rate among *k* counties is to separate the people in each county into two groups (surgery, no surgery), construct a 2 × *k* contingency table, and calculate the usual chi-square statistic with $k - 1$ degrees of freedom. This is appropriate if there are no readmissions, if all people in a county have the same probability of surgery, and if the expected number of surgeries per county is at least five. We calculated the chi-square statistic for each of the seven experimental conditions that have been discussed for the EQ. (Because we did not have to estimate the population proportion since it was set by the simulation, our chi-square has *k* rather than $k - 1$ degrees of freedom.) The estimated 95th percentiles of the chi-square statistics computed are shown in Table 6. The tabled value of the 95th percentile of a chi-square distribution with 39 degrees of freedom is about 55 (54.6).

The values in column 1, the experimental condition where all counties have the same large population, is close to 55 for every surgical rate. This is what would be expected, as the null hypothesis (no variation in rates among the counties) is true. In column 2, where the actual population sizes were used, the results are very similar. This is gratifying, especially given the relatively small expected values in some of the counties; the rate of 50 per 100,000 gives an expected value of only 1.3 surgeries in the smallest county, which has a population of 2,618. Use of the Yates correction (Armitage 1973) would probably allow even smaller expected values. Column 3 shows that the chi-square statistic has approximately the correct value even when the population size is reduced by 50 percent. Thus, in general, the chi-square statistic will provide an appropriate test statistic for the first three experimental conditions.

Table 6: 95th Percentile of 2 × 39 Chi-Square Statistic* in Seven Experimental Situations for Eight Surgery Rates (Based on 1,000 Iterations per Number)

	1	*2*	*3*	*4*	*5*	*6*	*7*
Rate per 100K	*Same Pop.*	*Real Pop.*	*Male*	*50% Readmit.*	*10% Readmit.*	*Large Only*	*Same K = 10*
50	54.7	55.3	55.9	91.0	63.6	54.6	55.4
100	54.6	55.5	54.6	92.2	64.8	54.6	52.8
250	55.7	53.9	53.6	89.4	65.3	57.2	53.6
500	55.4	54.9	53.4	91.8	64.5	55.0	57.2
1000	54.0	55.1	56.4	92.0	64.8	53.5	52.4
2500	54.1	55.0	55.5	91.7	64.3	55.0	55.1
5000	53.4	53.7	55.2	91.9	64.1	55.0	56.6
10000	54.1	53.9	56.9	98.5	66.8	53.0	54.8

Column:
1—All 39 counties have same population, 114,000.
2—Actual county populations used (2,618 to 1,344,586).
3—"Males Only" (50 percent of population used).
4—50 percent of people are readmitted once.
5—10 percent of people are readmitted once.
6—Counties with populations above 10,000 (31 counties).
7—Only ten counties, all with same population, 114,000.

* 95th percentile of chi-square statistic with 39 degrees of freedom is 54.57; for column 6, 95th percentile is multiplied by 54.57/44.98 to adjust for difference in degrees of freedom. For column 7, the factor is 54.57/18.31.

Column 4 shows the 95th percentile of the chi-square statistic (under the null hypothesis) if there are 50 percent readmissions. Note that these critical values are almost twice those of column 2. If one were to perform a chi-square test using 55 as the critical value, one would reject the null hypothesis much too often, since chi-square statistics on the order of 90 are actually required for rejection. Thus, the chi-square is not appropriate when there are many readmissions. Column 5 shows the situation with only 10 percent readmissions; these percentiles are also too large, by about 20 percent. Use of the chi-square test in this situation, with a critical value of 55, would also lead to nonconservative tests. A lower readmission rate is less serious than a higher rate, but it is still a problem.

The sixth column shows that (if there are no readmissions) the chi-square statistics are quite good for all rates if the smallest counties (population below 10,000) are removed. (To make comparison easier we multiplied the simulated 95th percentiles by 55/45, as the 95th percentile of a chi-square with 31 degrees of freedom is 45.) Column 7 shows that chi-square is as appropriate for 10 counties as for 39.

In summary, the chi-square method is a good way of testing for variability if the probability of readmission is zero and the expected number of cases per county is not too small, but it may be very misleading if readmissions are possible.

COEFFICIENT OF VARIATION

The weighted coefficient of variation (CV) has been used as a descriptive statistic for SAA (Chassin, Brook, Park, et al. 1986). The CV is the ratio of the standard deviation of the rates (among counties) to the mean rate (among counties) weighted by the population in each county. No tables are available to allow us to judge what is "too large." Table 7 shows the 95th percentile of the CV in the seven experimental conditions. The CV behaves similarly to the EQ, in that its 95th percentile decreases with an increasing surgery rate and increases if the number of people in the county is lower, and if there are readmissions. Unlike the EQ, the CV is slightly higher when the number of counties is smaller (column 7 versus column 1). The CV is relatively insensitive

Table 7: 95th Percentile of Coefficient of Variation (CV) in Seven Experimental Situations for Eight Surgery Rates (Based on 1,000 Iterations per Number)

	1	*2*	*3*	*4*	*5*	*6*	*7*
Rate per 100K	*Same Pop.*	*Real Pop.*	*Male*	*50% Readmit.*	*10% Readmit.*	*Large Only*	*Same K = 10*
50	.158	.158	.226	.206	.173	.143	.186
100	.111	.111	.159	.145	.121	.102	.129
250	.071	.070	.100	.090	.076	.064	.082
500	.050	.049	.070	.064	.054	.045	.057
1000	.035	.035	.049	.045	.038	.033	.039
2500	.022	.022	.031	.029	.024	.020	.025
5000	.015	.015	.021	.020	.017	.014	.018
10000	.010	.010	.015	.014	.012	.010	.012

Column:
1—All 39 counties have same population, 114,000.
2—Actual county populations used (2,618 to 1,344,586).
3—"Males Only" (50 percent of population used).
4—50 percent of people are readmitted once.
5—10 percent of people are readmitted once.
6—Counties with populations above 10,000 (31 counties).
7—Only ten counties, all with same population, 114,000.

to variability in population sizes (column 1 versus column 2). The CV is sometimes used to compare the variability in the rate of one surgical procedure to that of another (Chassin, Brook, Park, et al. 1986). This would not be an appropriate way to compare surgeries that are performed at different rates, since the difference in the rate itself would cause the less frequent surgery type to have a larger CV than the other. There are no strong reasons to recommend the use of the CV.

SYSTEMATIC COMPONENT OF VARIATION (SCV)

McPherson, Wennberg, Hovind, et al. (1982) developed a descriptive statistic that estimates the variance among counties that cannot be accounted for by the variability within each county. This is called the systematic component of variation (SCV), and large values are indicative of true differences among the counties. As with the CV, there are no tables for its distribution. The formula for the SCV (multiplied by 1,000) is

$$SCV = (1/k)\left[\sum_{i=1}^{k}((O_i - E_i)/E_i)^2 - \sum_{i=1}^{k}(1/E_i)\right]*1000 \qquad (1)$$

where O_i is the observed number of surgeries in county i and E_i is the expected number if the null hypothesis is true. It is possible to have negative SCV values, which is unsettling but acceptable in the null case, since the true underlying variance is zero.

Table 8 shows that the SCV behaves similarly to the EQ and the CV; that is, it is sensitive to the underlying rate, to the population sizes, and to variability in the population sizes, and is very sensitive to readmissions. As with the CV, it would not be appropriate to use the SCV to compare the variability of two surgical procedures unless their rates were very similar, since there is more variability for low rates than for high rates. It would also be wrong to compare several geographic areas unless the number of counties in each was similar.

The behavior of the SCV is not surprising, given its relation to the chi-square statistic. If all of the populations are the same, then E_i will be the same for all counties, and

$$SCV = (1/E)\,(\chi^2/k - 1)*1000 \qquad (2)$$

We have seen in Table 6 that the distribution of chi-square stays constant as the rate increases (column 1); since E increases with the rate, the above formula shows that the SCV must decrease as the rate

Table 8: 95th Percentile of Systematic Component of Variation (SCV) in Seven Experimental Situations for Eight Surgery Rates (Based on 1,000 Iterations per Number)

	1	*2*	*3*	*4*	*5*	*6*	*7*
Rate per 100K	*Same Pop.*	*Real Pop.*	*Male*	*50% Readmit.*	*10% Readmit.*	*Large Only*	*Same K = 10*
50	7.05	100.34	206.49	232.25	128.23	27.29	15.06
100	3.51	41.19	99.61	115.08	57.73	13.27	6.77
250	1.49	17.36	31.44	41.48	23.40	6.14	2.78
500	.72	8.09	17.80	20.80	12.21	2.65	1.59
1000	.32	4.38	8.45	8.52	6.50	1.41	0.65
2500	.12	1.64	3.15	4.52	2.12	.59	0.28
5000	.05	.70	1.58	2.14	1.11	.25	0.14
10000	.02	.34	.72	1.04	.55	.11	0.06

Column:
1—All 39 counties have same population, 114,000.
2—Actual county populations used (2,618 to 1,344,586).
3—"Males Only" (50 percent of population used).
4—50 percent of people are readmitted once.
5—10 percent of people are readmitted once.
6—Counties with populations above 10,000 (31 counties).
7—Only ten counties, all with same population, 114,000.

increases. Examination of a chi-square table shows that χ^2/k is larger for smaller values of k, which explains why values in column 7 are larger than those in column 1. The derivation of the SCV assumed that there were no readmissions, so it should not be surprising that it is sensitive to the readmission rate.

The SCV was developed as a measure for comparing several surgery types, or the same surgery in two different regions. Its sensitivity to many factors other than true variability among small areas suggests that it does not fulfill this purpose, at least when the null hypothesis is true.

AGE AND SEX ADJUSTMENT

In the simulation, we assumed that every person in a county had the same probability p of having surgery. This is, of course, an oversimplification, as it ignores the well-known variability of most health conditions by age and sex. Different age or sex distributions among the counties might cause differences to appear, even if the rates within each stratum were identical across counties. This is typically dealt with in SAAs by age/sex adjusting the surgery rates before comparing them among counties (Armitage 1973). We did not include this type of variation because it is difficult to address in general terms: we would need to decide on not just the number of counties, their sizes, the rates to be examined, and the readmissions rates, but also on how many age/sex strata to use, what the differences in rates are for the strata, and how the distribution of these strata might vary among the communities. We used both a theoretical approach and a small simulation to study the results of age/sex adjustment, as shown in Appendix C. The results suggest that the 95th percentiles obtained from the homogeneous population (Tables 1–8) are within a few percentage points of those that would be needed to allow for age/sex adjustment. The tables here are thus adequate for a first look at the data, with the caveat that SAAs that use age/sex standardization and whose results are fairly near to the tabled 95th percentile should be studied further using a simulation that incorporates the age and sex distribution of the counties, and the age and sex distribution of the surgery rate. It is not a difficult proposition to do this for a specific set of small areas and rates.

EXAMPLES: WASHINGTON DATA

Data were available on the number of *back surgeries* in the state of Washington, which averaged 89.8 per 100,000 (Volinn 1988). The extremal quotient for the rates was 172.1/11.5 = 14.96; chi-square was 196; the coefficient of variation was 0.24; and the SCV was 95.4. The second column and second row of Table 5, which has a surgery rate of 100 per 100,000, should give approximate information about the EQ if there were no readmissions. Under the null distribution the extremal quotient has a 95th percentile of 8.85. Since 14.96 is larger than 8.85, it represents more variability than would be expected by chance alone. The null hypothesis of equal surgery rates in all counties can be rejected, based on the EQ, *if there were no readmissions*. Columns 4 and 5 of Table 5 show the critical values of the EQ are 23.53 (for 50 percent readmissions) and 11.74 (for 10 percent readmissions). The EQ is not larger than would have been expected if there were 50 percent readmissions.

Table 6 provides similar information about the chi-square statistic. Critical values from columns 2, 4, and 5 are 55.5, 92.2, and 64.8, respectively. The observed value, 196, is well above these values, indicating highly significant results if any of these experimental conditions is true. Table 7 gives three critical values for the CV (.256, .330, .287). The observed value of .24 is not larger than any of these. For the SCV (Table 8), the critical values are 41.19, 115.08, and 57.75. The observed SCV of 95.4 is significantly higher than expected unless there are 50 percent readmissions. The highest rate (172.1) was in one of the smallest counties. If the eight counties with populations below 10,000 are removed, the EQ becomes 155.9/11.5 = 13.55. This is considerably higher than the 95th percentile of 3.26 for a rate of 100 shown in column 6 of Table 5 (if there were no readmissions).

Thus, for the four tests considered, there is clear significance based on the chi-square statistic; nonsignificance based on the CV; and equivocal significance, depending on the readmission rate, for the EQ and SCV. This points out, first, that it is vitally important to find out whether there are readmissions and how many there are. Second, it shows that the four measures of variability are not measuring variability in the same ways. It is likely that some of these measures are more sensitive to true variation than the others (i.e., more powerful). The chi-square seems to be more powerful in this situation, but this may be because we have understated the variance in the simulation model. More research is needed in the non-null situation, to determine which technique should be used.

A second set of Washington data was available for *Medicare patients* with one or more hospital admissions with diabetes as a primary or related cause (Connell 1983). Approximately 10 percent of the population of Washington was eligible for Medicare, the admission rate was 2,565 per 100,000, the average number of admissions per person admitted was 1.5, and the extremal quotient was 4593/632 = 7.3. From a simulation (not shown) based on each county having 10 percent of the population in Table 2, and a 50 percent readmission rate, the 95th percentile of the EQ for an admission rate of 2,500 is 8.4. The observed EQ of 7.3 is thus not statistically significant. The observed maximum is 1.96 standard deviations above the expected maximum of 3,691, and the minimum rate is 1.15 standard deviations below its expected value. A test based on the maximum rate thus indicates that it was significantly different from that expected under the null hypothesis. This suggests that tests based on the observed maximum could have more power in rejecting the null hypothesis than tests based on the EQ in some situations.

Since we have looked at the test of variability in several different ways, it may not be surprising that some of these tests turn out to be significant but others do not. This problem of multiple comparisons could be eased if it was known which of the possible tests provided the most power against alternatives of interest, since only that test would be used.

EXAMPLES: PUBLISHED ARTICLES

We used the simulation program to evaluate the results of articles in the literature. Although we did not do a systematic search, our eclectic set of about 50 SAA articles yielded only five studies that provided the necessary data (size of each county). We compliment the authors of these articles, and suggest that future authors provide the same information.

Pasley et al. (1987) present hospitalization and surgery rates for the elderly in 62 counties of New York state. Since the article gave the population for each county, we were able to perform the same simulation as for Washington state. Table 9 shows the observed EQs and the 95th percentiles of the EQ under the null hypothesis. We simulated both the no-readmission and the 50 percent readmission cases. For example, for cholecystectomy, the rate was 400 per 100,000; the observed EQ was 5.00, which is lower than the 95th percentile assuming no readmissions (5.23) and assuming 50 percent readmissions (9.09). The observed coefficient of variation (.306) is larger than the 95th percentile assuming no readmissions (.197) and that assuming 50 percent readmissions (.251). Five of the eight EQs reported were statistically significant. Three (for cholecystectomy, prostatectomy, and herniorrhaphy) were not significantly different from what would be expected by chance. These three conditions where the EQ was not significant involved the lowest rate procedures, and prostatectomy is applicable to only half of the populations. The expected numbers of cases in the smallest county were 3.3, 8.0, and 3.5 for the three procedures, respectively. There was also considerable variability in the population sizes (837 to 281,328). All of these conditions should lead us to be wary of accepting the EQ on its face value. Although the EQ was not always significant, the observed CVs were significantly higher than the 95th percentile in every case, indicating that there is significant variability. Here, the CV seems to be more powerful than the EQ, again suggesting that there may be differences in power for the various tests.

Wennberg and Gittelsohn (1982) presented data on hospital admissions for 13 areas in Vermont. The smallest area had a population of 7,960, and the observed rate was 16,335 per 100,000. The observed extremal quotient of 1.63 was highly significant, based on simulations assuming a 50 percent readmission rate as above (the 95th percentile was 1.12). As the expected number of cases in the smallest county is 1,300, and all counties are relatively large, these results might have been expected. On the other hand, our simulation model perhaps

Table 9: Reanalysis of Discharge Rates for the Elderly

		Extremal Quotient			*CV*		
	Discharge	*Observed**	*Expected 95%*		*Observed**	*Expected 95%*	
Type of Discharge	*Rate per 100K*		*No Readm.*	*50% Readm.*		*No Readm.*	*50% Readm.*
All	35300	2.38	1.13	1.20	.172	.017	.024
Surgical	13700	1.88	1.26	1.37	.123	.032	.042
Selected elective	2600	2.87	1.80	2.25	.177	.078	.102
Selected nonelective	2300	2.75	2.02	2.45	.153	.083	.107
Cholecystectomy	400	5.00	5.23*	9.09*	.306	.197	.251
Prostatectomy (males)	1900	2.48	3.49*	4.59*	.240	.128	.170
Herniorrhaphy	400	2.62	5.34*	9.53*	.241	.198	.252
Lens extraction	1200	5.40	2.72	4.03	.274	.115	.148

Source: Data from Pasley, et al. 1987.

*The observed EQ is not higher than the 95th percentile.

did not allow enough variability in the readmission rate—there were surely people with more than two admissions. This might still be examined in more detail.

Knickman and Foltz (1984) looked at hospitalization rates from the National Health Interview Survey in four regions of the country, the smallest having a sample 10,921 respondents. The observed EQ for admission rates falls right at the 95th percentile of the EQ (assuming a 50 percent readmission rate), and so is just statistically significant. The smallest number of admissions expected was 1,229. This article also looked at variation in the number of hospital days per person, clearly not a binary variable. We cannot evaluate that analysis at this time, but it would be possible to address it with the simulation technique if the underlying distribution of hospital days per person were known.

Lewis (1969) presented data on six surgical procedures for 13 regions. The smallest region had a population of 41,000 and the smallest rate was 42 per 100,000. All of the regional differences shown were statistically significant. The smallest expected number of admissions was 17.

Chassin, Brook, Park, et al. (1986) analyzed the use of services by Medicare patients. Thirteen areas of the United States were studied, with the smallest having 83,000 Medicare enrollees. Sixty-seven of 123 procedures studied showed at least threefold differences between sites. All of the 30 "selected" medical and surgical procedures presented in the article showed statistically significant variation based on the 95th percentile of the EQ and on the CV. The minimum expected number of cases was about 17.

This review did now show any large problems in the published articles. Probably the most significant finding is that so few articles provided the necessary data for the comparison to be made.

SUMMARY AND DISCUSSION

Based on these results, the amount of chance variability in the descriptive statistics used in SAA is higher than intuition might have suggested. There is more variability for low-incidence surgeries and for smaller subgroups of the population. The variability also depends strongly on the probability of readmission, which is rarely considered in small-area studies. The inclusion of counties with small populations is also a major determinant of the variability. Some descriptive statistics are more sensitive than others to these factors. When "real" data on the counties of Washington and New York states were tested against the null distributions, some extremal ratios that seemed to be high were not significantly different from the null situation. Results from larger data sets seemed more stable.

Our findings suggest strongly that investigators should explore the null hypothesis, rather than assuming that observed variability is significant or important. If the null hypothesis cannot be rejected, either there is not excess variability or the data at hand are not adequate to assess the existing variation.

We are unable to recommend a single good descriptive statistic for SAA. A descriptive statistic should have some intuitive meaning, and should be sensitive to important variations in the data but insensitive to unimportant variations. Of the four statistics examined, the EQ is most intuitively satisfying, followed perhaps by the SCV and the CV. The chi-square is not very intuitive. On the other hand, all of the other measures vary considerably due to such factors as the number of counties, the size of the counties, the variation in county size, and the underlying rate being evaluated. Comparisons of several surgical procedures, or of several counties for the same surgical procedure, might be deceptive if these statistics are used. Further, the distributions of these statistics are not known in the null situation. The chi-square statistic, although it lacks intuition, is tabled and is independent of most of these factors. If the statistic is divided by its degrees of freedom, χ^2/dof is relatively independent of the number of counties, as well. (The 95th percentile for 5 counties is 2.2; for 10, 1.8; for 20, 1.7; for 30, 1.5; and for 100, 1.2.) Future researchers might consider using this statistic (among others) in SAAs of "ectomies." We have no recommendations on descriptive statistics for the situation where readmissions are possible. More information about the true underlying distributions, and more simulations, are necessary to address this problem.

How should an investigator proceed to test the null hypothesis? The results above indicate that if the counties are large and about the same size, there are no readmissions, and the expected values are fairly large, the EQ may be used, either at face value or with tables newly developed by Kazandjian et al. (1989). If there is variability in the county sizes, however, the only way to evaluate the significance of the EQ is a simulation similar to ours. The SCV and the CV present similar problems, in that there are not tabled values, and also because they are sensitive to the underlying surgery rate as well as to the variability. Only the chi-square statistic is uniformly good in these situations, as it is tabled and does not vary with the rate; we recommend using it if the expected numbers of cases are five or more (and lower if the Yates correction is used). If age/sex adjustment is required, two variations of the chi-square test—the Mantel-Haenszel approach (Armitage 1973) or a logistic regression that tests for the effect of the dummy variables for county, after controlling for age and sex—would also be reasonable.

If the same person can be counted more than once, however, none of these methods can be trusted, as is shown clearly in columns 4 and 5 of Tables 5 through 8. These so-called "readmissions" can occur because of rehospitalizations for the same problem or for complications, because of billing errors in claims files, or because the variable of interest was not binary to start with (e.g., number of hospital admissions of any type, total health care costs, number of fillings). In such situations it is crucial to determine the distribution of the number of admissions per person. If data are available *at the person level*, and if they can be considered normally distributed, an analysis of variance would be appropriate. (For instance, if the dependent variable were log visits or log costs, they might be approximately normal. Preliminary analyses suggest that ANOVA would be well behaved even in the experimental conditions used in Tables 1–8.) If not, a modification of this simulation program could be used to generate the appropriate critical values. Analysis of variance may be difficult, as data bases often contain no records for the people who did not have surgery; either dummy records would be required or special software would be necessary.

Another approach to determining whether there is excess variability has been suggested. This is to regress the observed rates on some relevant covariates (e.g., number of surgeons per capita). Under the null hypothesis that all underlying rates are the same, there should be no significant association between the rates and the various regressors. This approach is particularly appealing because it does not require that the distribution of the dependent variable be known at the individual level; we need assume only that the rates themselves are normally distributed and independent, which might be approximately true. Unfortunately, this approach may be more appropriate in theory than in practice due to the nature of small-area analysis. There are usually relatively few data points (one per small area) and the variables tend to be "per capita" rates. The underlying variability of rates of this type has been demonstrated in this article, and there will be as much variability in the regressors as in the dependent variable, suggesting that spurious relations will abound. The small number of observations means that outliers will tend to have a large influence on the estimated regression coefficients and significance levels. Variability in size among the counties means that the assumption of homoscedasticity will not be met. The usual weighted least-squares approach with weights proportional to size will inappropriately assume that there is no variability in rates among counties, and will give most of the weight to the largest counties. There has been work on the "proper" weights to use in such analyses (Breslow 1984; Pocock, Cook, and Beresford 1981; Tsutakawa 1988), but these depend on knowledge of the distribution of the dependent variable *at the individual level*. Finally, SAA regressions usually require adjusting the observed variables (e.g., number of surgeons) for population size. In our experience, simply dividing the observed counts by the population size yields a variable that is still correlated with population size, leaving the strong probability that "significant" associations are really due to correlation of the variables under examination with a third variable (population). In a recent study of back surgery, 30 regressors were examined, and many of the "significant" results could be explained by a single outlier, usually one of the smallest counties (Volinn, Mayer, Diehr, et al. 1988).

After significance is established, what is the next step? Two questions are usually asked. The first is, "which counties are significantly different?" The second is, "what types of counties are significantly different?" It is possible to calculate a 1 degree of freedom chi-square statistic for each county $[(0 - E)^2/E]$. However, this will not follow the chi-square distribution if readmissions are possible. Further, unless this particular county has been specified in advance, the statistics should be adjusted for multiple comparisons. In our situation, where 39 such comparisons would be made, $\alpha = .05/39 = .0013$ is the adjusted level, corresponding to the critical value of 9.0 rather than the usual value of 3.84. Such a procedure might have low power. In situations where analysis of variance is appropriate, the usual multiple comparisons methods can be used to determine which counties are different from the others. However, if the number of counties is large, it may be very difficult to achieve significance. The simulation method could be adapted to calculate the distribution of, for example, the second-highest rate, or the biggest difference between two adjacent rates, under a specific type of (non-null) variability (e.g., the smaller counties have higher rates). The regression approach is the only approach we have to determine characteristics of counties with high rates, but it must be used with great care, for reasons just mentioned. Clearly there is room for methodologic work in this area.

The implications of these findings are limited somewhat by the form of the simulation, the lack of data to improve the simulation, and the generalizability to other settings of findings about the state of Washington. As mentioned above, age and sex variation are not included explicitly in these simulations, but the general findings are true with or without age/sex adjustment. Another simplification was the model for readmissions, which lets individuals have either one or two admissions, but no more. This was done primarily for illustrative purposes. In real situations, people may have many more than two admissions, which would increase the amount of variability above that shown. We reasoned that the example given made our point—that readmissions are important. In future work, if the real distribution of readmissions is known, it can be incorporated directly into the simulation. We hope that future researchers will concentrate on obtaining such data.

We have not addressed the power of these tests. It may well be that some of the statistics or methods are better than others in detecting true variation. We presented some examples where one type of test rejected the null hypothesis, but another did not. This area remains to be explored. The relative performance of the descriptive statistics may well be different in the non-null situation.

This discussion has noted several methodological issues that should be addressed. One factor inhibiting such research is the lack of detailed SAA data. The publication of data at the patient level, showing the number of patients with 0, 1, 2, admissions, for instance, would permit estimation of variances that could be used in the simulations or in other methods where the binary (no readmissions) assumptions are incorrect. It is often difficult to obtain person-level data because there are no individual patient identifiers on claims records. Another possibility is to obtain several years of data for each small area. This would permit a direct estimate of the variation for each county which might, after smoothing, be used in simulation programs or other methods.

Reporting standards for SAAs should be established. At a minimum, an investigator should provide the actual sizes of all of the small areas, and indicate whether there are readmissions in the data base. If there are no readmissions, then the chi-square statistic or its variants can be used to establish excess variation. An alternative is to use other statistics, but their null distributions must either be tabled (Kazandjian, Durance, and Schork 1989) or obtained from simulation. If there are readmissions, then the investigators must present information about the distribution of the readmissions at the person level. Based on that information, it may be possible to determine a statistical technique that is appropriate. Or, the simulation method can be used to find the 95th percentile of the desired test statistic under the null hypothesis, after the variation has been built into the model.

The existence of large variation under the null hypothesis for counties in Washington state does not necessarily mean that published studies of small-area variation are inaccurate. In general, published studies, including those cited here, have been based on "ectomies" and on very large "small areas," for which the rates are fairly stable. However, we know of investigators who plan to perform analyses similar to the Washington simulations. And the article that analyzed counties (Pasley et al. 1987) did have some results that are not statistically significant. These findings are probably most meaningful for investigators who plan to perform SAAs in new situations.

Researchers should be wary of findings based on small populations; lower-incidence surgeries; procedures that may involve readmissions; "nonstandard" procedures that are not binary, such as hospital admissions or patient days; and on new types of "small areas," such as hospitals or dental practices, whose variability has not been studied. New SAA research should make an effort to study the null hypothesis and to present data useful for its evaluation, such as the sizes of populations and the distribution of readmissions. These new types of small-area analysis may well be misleading if attention is not paid to the null distribution.

Further methodologic work is needed with respect to the null hypothesis. Research to identify the study designs and descriptive statistics that are most powerful in detecting small-area variation is also important. We believe that the simulation approach can be useful in providing ways to test for significant variation among small areas, in the null and the non-null situation. For "ectomies," only the county sizes have to be provided to the program. For non-ectomies, the researcher must also determine the distribution of readmissions (or whatever the variable is) at the patient level, so that the program can be set to simulate this. We hope to work with other SAA investigators in exploring these areas.

APPENDIX A

SIMULATION MODEL

The simulation program is written in FORTRAN and runs on an IBM AT. The description of the program in the text made some oversimplifications. We do not actually assume that all of the surgery rates have a normal distribution. This assumption would be defensible for large values of p (the probability that an individual has surgery) and n_i, based on the normal approximation to the binomial. However, if the expected number of cases in a county is less than five, the approximation is not good and can even yield negative values of the surgery rate. For this reason we use the exact binomial calculation for counties whose expected number of surgeries is less than five. (The Poisson model would also be appropriate if the counties are sufficiently large.)

The method for allowing readmissions is as follows. Let the proportion of the population with no admissions be $1 - p$, the proportion with exactly one admissions be $p - a$, and the proportion with exactly two admissions be a. The expected number of admissions per person is $1^*(p - a) + 2a = p + a$. The variance is $v = p + 3a - (p + a)^2$. (If $a = 0$ there are no readmissions, and the mean reduces to p and the variance to $p(1 - p)$ as in Figure 1. The value of p used is adjusted so that $p + a$ in the readmission case is the same as p in the no-readmission case.) The variance of the proportion with an admission is v/n_i. For the larger counties, we simulated the rate as shown in Figure 1, but using mean $p + a$ and variance v/n_i. For the smaller counties, in which the exact binomial distribution was used, we used a two-step process. After determining the number of people with at least one admission, we "flipped a coin" for each person to decide whether that person had a second admission.

Tables 1-4 are based on 3,000 iterations per line. We found that this produced very stable estimates of most of the parameters. However, the mean and standard deviation of the EQ were still unstable with 3,000 iterations. For other parameters, a number of iterations as low as 500 produced fairly good estimates. It is likely that the mean and standard deviation of the EQ are not worth computing, because of the long right tail of the distribution, and that the simulation should instead be used to estimate percentiles of the distribution of the EQ (or other statistics such as the CV).

APPENDIX B

POPULATION SIZES FOR WASHINGTON STATE COUNTIES
(Total = 4,447,315)

1.	2,618	9.	13,817	17.	33,530
2.	3,636	10.	16,505	18.	36,156
3.	4,022	11.	17,528	19.	36,730
4.	6,129	12.	18,003	20.	41,637
5.	7,501	13.	18,053	21.	48,548
6.	9,035	14.	24,326	22.	49,003
7.	9,076	15.	25,042	23.	50,661
8.	9,583	16.	32,135	24.	52,821

Continued

APPENDIX B: Continued

25.	53,639	30.	111,382	35.	210,349
26.	57,733	31.	114,016	36.	359,467
27.	63,899	32.	145,058	37.	384,458
28.	70,851	33.	175,986	38.	527,559
29.	79,202	34.	183,035	39.	1,344,586

APPENDIX C

THEORETICAL AND SIMULATION APPROACH TO AGE/SEX STANDARD RATES

Age and Sex Adjustment

In the simulation, we assumed that every person in a county had the same probability p of having surgery. This is, of course, an oversimplification, as it ignores the well-known variability of most health conditions by age and sex. Different age or sex distributions among the counties might cause differences to appear, even if the rates within each stratum were identical across counties. This is typically dealt with in SAAs by age/sex-adjusting the surgery rates before comparing them among counties (Armitage 1973).

We did not include this type of variation for several reasons. First, it would be difficult to address in general terms, since we would need to decide not just on the number of counties, their sizes, the rates to be examined, and the readmission rates, but also on how many age/sex strata to use, what the differences in rates were for the strata, and how the distribution of these strata might vary among the communities.

In this appendix, we evaluate the effect of age/sex standardization on the variance of the estimated rates, and hence its effect on the variability of statistics such as EQ, using both a theoretical approach and a small simulation. In our other work, we assumed a uniform value of p for all people in each county. It is more common to have data in which the p's vary by strata, and the strata vary by county. What is the effect of using age/sex-standardized data with overall rate p, rather than homogeneous data with rate p?

Define the following:

n_i = population of county i, $i = 1,2,. . .k$.
n_{ij} = number of people in age/sex stratum j ($j = 1,2, . . . n$) in county i.
f_{ij} = n_{ij}/n_i (fraction of people in county i who are in stratum j).
π_j = fraction of the standard population who are in stratum j.

$\hat{p}_{ij}$ = estimated surgery rate in county i, stratum j (i.e., number of surgeries divided by n_{ij}).

p_j = true surgery rate in stratum j (under the null hypothesis, assumed to be the same in all counties).

Under the null hypothesis that surgery rates do not differ across counties, the true age/sex-adjusted surgery rate for all counties is

$$p^* = \sum_{j=1}^{n} \pi_j p_j$$

Under the case of no readmissions, the number of surgeries in stratum j in county i has a binomial distribution, and hence the variance of $\hat{p}_{ij}$ is

$$Var\,(\hat{p}_{ij}) = p_j(1 - p_j)/n_{ij}$$

The age/sex-standardized estimate of surgery rate for county j is

$$\hat{p}_i^* = \sum_{j=1}^{n} \pi_j \hat{p}_{ij}$$

and the variance of this estimate is

$$\begin{aligned} Var\,(\hat{p}_i^*) &= \sum_{j=1}^{n} \pi_j^2 \quad Var\,(\hat{p}_{ij}) \qquad \text{(C.1)} \\ &= \sum_{j=1}^{n} \pi_j^2\, p_j(1 - p_j)/n_{ij} \\ &= \frac{1}{n_i}\sum_{j=1}^{n} \left(\frac{\pi_j}{f_j}\right) \pi_j p_j(1 - p_j) \\ &= \frac{1}{n_i}\sum_{j=1}^{n} \left(\frac{\pi_j}{f_j}\right) \pi_j p_j - \frac{1}{n_i}\sum_{j=1}^{n} \left(\frac{\pi_j}{f_j}\right) \pi_j p_j^2 \end{aligned}$$

This variance can be compared to the variance of the unstandardized rate estimate if the population of county i were homogeneous, with all persons having probability of surgery equal to p^*:

$$\frac{1}{n_i} p^*(1 - p^*) = \frac{1}{n_i}\sum_{j=1}^{n} \pi_j p_j - \frac{1}{n_i}\left(\sum_{j=1}^{n} \pi_j p_j\right)^2 \qquad \text{(C.2)}$$

If the surgery rate is low, the second term in both (C.1) and (C.2) will be small and can be ignored. If the distribution of people across age/sex strata in county i is similar to the age/sex distribution of the standard population, the factor π_j/f_j will be close to one for all j and the variance in (C.1) will be similar to that in (C.2). In this situation the variance (C.2) will give a good approximation to the true variance (C.1). The simulations presented in this article can be thought of as being based on this approximation. Suppose, however, that the age/sex distribution does differ substantially across counties. In this situation, the factor π_j/f_j will be greater than one in those strata that are underrepresented in county i relative to the standard population, and less than one in overrepresented strata. If all of the strata with a high surgery rate (p_j) are underrepresented while those with a low surgery rate are overrepresented, then the actual variance for that county (C.1) will be bigger than the approximate variance (C.2). If this is true for many of the smaller counties, the distribution of the EQ will have greater variability than shown in the simulations. On the other hand, if all of the small counties have reduced variance, due to overrepresentation of all of the strata with high surgery rates, then the variability of the EQ will be smaller than shown in the simulations. We therefore conclude that in the null case, age/sex standardization can either increase or decrease the variability of the EQ (as well as the other statistics), depending on the age/sex distribution in all the counties, relative to how the surgery rate varies across strata.

We tried two examples, dividing people in the state of Washington into two strata, for over and under age 65. The percent of people over 65 varies from 8 percent to 20 percent among counties. We considered two extreme-seeming cases: (a) the young have a rate ten times as high as the old and (b) the old have a rate ten times as high as the young. We calculated the required variance multiplier factors—that is, (C.1)/(C.2)—and used these in the simulation program. The multipliers were generally near 1, but some were below and some were above 1. For condition (a) the 95th percentiles of the EQ, chi-square, and CV were all 0 to 3 percentage points higher than those obtained assuming a homogeneous population, varying depending on the underlying rate. For the SCV, percentiles were 0 to 10 percentage points higher. Thus, the tabled percentiles would be substantially correct, except for the SCV. In condition (b) the EQ 95th percentile varied from 81 to 101 percent of the homogeneous value; the chi-square from 93 percent to 96 percent; the CV from 92 percent to 98 percent; and the SCV from 66 percent to 97 percent. Again, the simulated numbers assuming a homogeneous population are substantially correct, except for the SCV. Another example with six age groups gave similar results.

These results suggest that using 95th percentiles obtained from the homogeneous population is adequate for a first look at the data, but that SAA results that are fairly near to the 95th percentile should be looked at more carefully using a simulation that incorporates the age and sex distribution of the counties, and the age and sex distribution of the surgery rate. It is not a difficult proposition to do this for a specific problem.

ACKNOWLEDGMENT

The authors wish to thank Patricia Coley for help in the preparation of this article.

REFERENCES

Armitage, P. *Statistical Methods in Medical Research.* New York: John Wiley & Sons, 1973.

Breslow, N. "Poisson Variation in Log-Linear Models." *Journal of the Royal Statistical Society Series C* 33, no. 1 (1984):38-44.

Chassin, M. R., R. H. Brook, R. E. Park, et al. "Variations in the Use of Medical and Surgical Services by the Medicare Population." *New England Journal of Medicine* 314, (January 30, 1986):285-90.

Connell, F. "Analysis of Diabetes Hospitalizations for the Medicare Population." University of Washington School of Public Health, 1983.

Copenhagen Collaborating Center. *CCC Bibliography on Regional Variations in Health Care.* ISBN 87-7488-202 = 3. Copenhagen: Vedbaek, Tekst og Tryk A/S, 1985.

Diehr, P. "Small Area Statistics: Large Statistical Problems." *American Journal of Public Health* 74, no. 4 (1984):313-14.

Dixon, W. J., and F. J. Massey. *Introduction to Statistical Analysis.* New York: McGraw Hill, 1957.

Health Affairs. "Special Issue on Medical Practice Variations." 3, no. 2 (1984).

Kazandjian, V., P. Durance, and M. Schork. "The Extremal Quotient in Small Area Variation Analysis." *Health Services Research* 24, no. 5 (December 1989): 665-84.

Knickman, J., and A. M. Foltz. "Regional Differences in Hospital Utilization: How Much Can Be Traced to Population Differences?" *Medical Care* 22, no. 11 (1984):971-86.

Lewis, C. E. "Variations in the Incidence of Surgery." *New England Journal of Medicine* 281, no. 16 (1969):880-85.

McPherson, K., J. Wennberg, O. Hovind, and P. Clifford. "Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New England, England, and Norway." *New England Journal of Medicine* 307 (November 18, 1982):1310-14.

Pasley, B., et al. "Geographic Variations in Elderly Hospital and Surgical Discharge Rates, New York State." *American Journal of Public Health* 77, no. 6 (1987):679-84.

Paul-Shaheen, P., J. Clark, and D. Williams. "Small Area Analysis: A Review and Analysis of the North American Literature." *Journal of Health Politics, Policy and Law* 12, no. 4 (1987):741-809.

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

Sarhan, A., and B. Greenberg. *Contributions to Order Statistics*. New York: John Wiley & Sons, 1962.

Tsutakawa, R. "Mixed Model for Analyzing Geographic Variability in Mortality Rates." *Journal of the American Statistical Association* 83, no. 401 (1988):37-42.

U.S. Dept. of Health and Human Services. "Inpatient Hospital Services: Use, Expenditures, and Source of Payment." Report prepared by A. Taylor. Data Preview 15 from the National Health Care Expenditure Study. Publication No. (PHS) 83-3660. Washington, DC: Government Printing Office, 1983.

Volinn, E., J. Mayer, P. Diehr, F. Connell, and J. Loeser. "Small Area Analysis of Surgery for Low Back Pain." Final report for NCHSR grant HS 05497-01A1. University of Washington, Seattle, 1988.

Wennberg, J., and A. Gittelsohn. "Variations in Medical Care among Small Areas." *Scientific American* 246 (April 1982):120-34.