INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa • University of Ottawa

Small Area Variations in Surgical Rates:

Simulation as an Aid in Interpretation of Findings

by

© Andrew William Howard

Thesis submitted to the School of Graduate Studies and Research in partial fulfillment of the requirements for the MSc degree in Epidemiology

University of Ottawa

June, 1998

Abstract

Purpose: To explore the uses of Monte Carlo simulation in the statistical summarization and interpretation of small area variations data regarding surgical rates. Currently, four summary statistics are used to describe these rates: the extremal quotient, the coefficient of variation, the systematic component of variation, and chi-square. However, it has been shown that these measures have limitations. The chi-square is the only one that can easily be used as a test statistic. The others are simply descriptive. Changes in population size, regional divisions, or overall surgical rate, can cause significant changes in the value of these descriptive statistics even without variation in the underlying rate. This limits their descriptive value.

Monte Carlo simulation is useful in studying statistics whose distributions are mathematically difficult to calculate. The method gives empirical information regarding the expected distributions of the currently used statistics. It can also be extended to allow description of newly defined statistics with particular policy relevance, and to allow modifications to the traditional testing of the null hypothesis.

Methods: A series of Monte Carlo simulations was used to generate sets of rates of surgery observed in each region under the null hypothesis of no interregional variability in true rate. The statistics of interest were calculated for each set of rates. The expected distributions of each statistic were described over a wide range of parameters.

A new statistic, the case count, was defined. This measure is based on the index of dissimilarity used for analyzing socioeconomic inequalities in health. The expected distribution of this measure was determined using the same scenarios as for the traditional statistics. The power of this statistic to detect true rate variability was estimated and compared with the power of the currently used statistics.

The standard null hypothesis of equal rates in all regions was changed to allow some true intraregional variability. The amount of variability allowed was based on proposed proxies for population morbidity differences that persist after age and sex adjustment. Proxy measures used were the positive primary appendectomy rate and the standardized mortality ratio.

Results: Of the four currently used statistics, the chi-square and the coefficient of variation had stable and interpretable values over the range of population sizes, divisions, and surgical rates tested. The extremal quotient and the systematic component of variation did not.

The case count was stable and interpretable despite changes in population size, divisions, and surgical rate. The case count had adequate power to detect most unequal distributions of interest, but had poor power at detecting a single low rate outlier in a small region.

Modification of the null hypothesis by incorporation of appendectomy or SMR data to allow some true intraregional variability changed the results of hypothesis testing.

Only two of the five operations analyzed showed significantly greater variability than expected, compared with five out of five under traditional testing.

Conclusions: Monte Carlo simulation is a useful tool for small area variations studies. If standard statistics are used, they can be placed in context by using Monte Carlo simulation to estimate their expected values.

The case count, as a new nonstandard statistic, relies on Monte Carlo simulation for estimation of its expected values. The case count adds valuable summary information about equity and estimated expenses to observed small area variations. This should be useful for health policy.

Modification of the null hypothesis allows adjustment of the sensitivity of small area variations studies. Small area variations studies are often described as a screening test, but it is a test with 100% sensitivity and low specificity, since the standard analysis always shows statistically significant variability. Monte Carlo simulation allows one to modify the sensitivity and specificity of the screening test in a flexible manner, which should allow better decisions about which observed variations need further investigation.

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisors, Dr. Rama Nair, Dr. Bob Spasoff, and Dr. Andreas Laupacis, for their expertise, suggestions, encouragement and support. Dr. Nair guided me through the technical and statistical aspects of the simulation, and taught me a lot about statistics along the way. Dr. Spasoff introduced me to the topic of this thesis. The case count statistic was largely his idea. Dr. Laupacis kept the project moving forward, and was very helpful in the interpretation and application portions.

My colleagues in the division of paediatric orthopaedics at CHEO, Drs. Letts, MacIntyre, Lawton, Jarvis, and Lalonde, have all generously helped to protect my academic time.

My wife Lianne and my daughter Emma have provided encouragement, support, quiet, and perspective. I thank them for their forbearance.

Table of Contents:

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 History of Small Area Variations Research

Small area variations studies are not a new field in health services research. The first widely cited study of geographic variation was that by Glover (1), who in 1938 found a ten fold geographic variation in tonsillectomy rates among British schoolchildren. An acceleration in the performance and impact of this type of study occurred in Canada and the United States in the early 1970's. Key investigators were John Wennberg in Vermont (2-5), and Noralou Roos in Manitoba (6-12). Studies from this era investigated overall rates of surgery, as well as rates of specific procedures, and compared the rates by patient origin in contiguous small geographic areas. The surprising finding was that for many procedures, there was a high degree of variability in rates of use across small areas with very similar population characteristics.

Wennberg and his associates performed some survey work aimed at confirming the similarity of the populations involved and excluding differences in morbidity. They were unable to explain the observed variation by differences in population characteristics(3).

Two hypotheses arose from Wennberg's initial work. The first was that physicians were uncertain about the indications for the surgical procedures they performed, and

that different surgeons practised differently based on differing understandings of the purpose of operating. The second hypothesis was that the physician, and not the patient, was in control of deciding whether an operation took place or not. These statements, not always explicitly separated, became known collectively as the practice style hypothesis(2,3).

The practise style hypothesis was extensively investigated by Wennberg. Surgery for benign prostatic hypertrophy was selected as the case study(5,11). By interviewing surgeons he found two distinct theories about indications. One group of surgeons believed in operating early, when obstructive urinary symptoms first appeared, in order to prevent progression and to avoid having to operate on the patient when he became older and sicker. Another group believed in operating only to relieve current symptoms, and therefore would defer surgery in favour of watchful waiting. Finding this clear divide in surgical opinion yielded strong support for the practise style hypothesis. A closer look revealed inadequate scientific information about the outcomes of prostate surgery to allow surgeons to rationally choose between the two strategies. Careful outcome studies were performed, both by following cohorts of operated and nonoperated patients, and also by analysing large administrative databases to estimate clinical outcomes such as mortality and reoperation following prostate surgery(11).

The outcome studies showed clear differences in the primary variable - urinary flow rates improved markedly in operated groups. However, quality of life issues and side effects of surgery, including differences in sexual function, were often more important

to patients than urinary flow rates. When the total picture of surgical versus nonsurgical outcomes and side effects was assembled into an interactive videodisc and presented to patients, only one in five patients with severe obstructive symptoms opted for surgery(13). At the individual patient level the decision hinged on such factors as risk tolerance or aversion, and valuation of side effects, and did not hinge on the main biological effects of the surgery.

Several important advances came out of Wennberg's work - identification of unexplained variation pointed out a need for better understanding of both outcomes and patient preferences in surgery. Based on this success, the Agency for Health Care Policy and Research (AHCPR) funded patient outcome research teams (PORTs) in the 1980s to replicate the Wennberg research agenda in other areas(13,14). One of the stimuli for this entire movement was the relevance of small area variation studies to considerations of both the cost and the quality of medical care—both core concerns of health care policy. The success of this movement has been partial at best, and the criticism from some very respected quarters has been harsh:

> "This conference is supposed to be on the evaluation of interventions and I can't see anything in either this paper or the preceding one that actually does anything towards evaluating the interventions...It isn't really evidence. This isn't a serious way of trying to work out what works and what doesn't." Richard Peto, in the published floor discussion section of (14).

International research efforts have replicated the finding of excessive variation in surgical rates in all countries studied (2). This is important because organizationally, fee for service health care in the United States, without universal coverage, is somewhat of an anomaly. One might imagine that countries more concerned with access to health care for all would have less variability in surgical rates. However, although overall rates of surgery vary greatly from country to country, within every country studied the small area variations in surgical rates are considerable(2,15,16). This lends support to the practise style hypothesis, as well as prompting study of small area variations in all systems.

An extreme interpretation of the practise style hypothesis is that high rates of surgery are explained by operations being performed for inappropriate reasons in some regions. This has been explicitly studied in the US by the RAND corporation (17-22) and in Canada by the Institute for Clinical Evaluative Sciences (ICES) (23). Although inappropriate use occurred, it did not occur to a higher degree in the high use areas. Inappropriate excess surgery does not explain small area variations, according to these studies(17-23).

Historically, small area variations research has focused mainly on surgical procedures. This has been for various reasons. Administrative databases are often used, and the occurrence of a surgical procedure is relatively easier to discern and more reliably coded than is medical treatment. The per capita cost of surgery is a significant proportion of overall health care costs in most jurisdictions(5,24). Recently, small area techniques have been applied to medical diagnoses (24),

psychiatric care (25), and even prescription drug utilization (26) in Canada, although they are not always done in a methodologically correct fashion. In Ontario, however, the major thrust of small area variations research remains the investigation of surgical rates. Rates of many surgical and a few medical procedures are tracked yearly from 1989 to the present, and are published as an atlas which is regularly updated. The intention of the atlas is to stimulate research into reasons for rate variability and appropriate policy responses to rates which are too variable. The organization (ICES) responsible for the publication of the atlas, although independent of the Ontario Ministry of Health, is influential as a source of information for health policy decisions both within and outside the ministry.

This thesis restricts itself to small area variations in surgical rates. The province of Ontario is used as the case study to present the discussion of statistical and methodological issues. Surgical rates only are used because of the traditional reasons discussed in the previous paragraph, and for two additional reasons. In Ontario, some surgical procedures are individually rationed, such as coronary bypass surgery by a formalized provincial priority list, and joint replacement by specific capped funding to hospitals to pay for implants. Such rationing requires monitoring of population based surgical rates. The second reason is that the problem of multiple care episodes (readmissions) attributable to the same person causes serious statistical problems if person level data are not available. Multiple operations for the same diagnosis in the same year are rare(27), and make little differences to the statistical analysis as

detailed below, whereas certain medical conditions are characterized by repeated admissions or care episodes and must be analyzed very differently.

The substance of this thesis is an exploration of the statistical issues surrounding small area analysis of surgical rates. The focus is on producing statistical analyses that are useful in setting health policy. The specific method used is that of Monte Carlo simulation for interpreting statistical analysis. The flexibility of this technique may allow specific new policy-relevant statistics to be designed, and also may allow reformulation of the null hypothesis to match the specific goals of a particular analysis.

1.2 Proposed Reasons For Variation

There is a long list of potential reasons for an observed true variation in rates. The populations may differ in risk factors, morbidity, risk tolerance, or health values. Access to primary care may differ. Primary caregivers may differ in diagnostic ability, willingness to treat, and referral threshold. Access to surgical care may differ, due to geographic barriers, underservicing, or overservicing. Surgeons may differ in their opinion regarding the benefit of operation and its indications. Scientific information may be inadequate to guide surgical opinion, or may be poorly disseminated or appreciated. There may in fact be true differences in the actual results of operations among surgeons. Resource limitations, including diagnostic tests (CT, MRI, angiography), operating time, surgical beds, and required equipment or implants may differ between regions. Patient participation in decision making may

differ, for reasons that may be patient related or surgeon related. Appropriate tradeoffs in the cost and benefit of surgery may differ with how much a society can afford, and what proportion of its resources it wishes to allocate to medical care.

A reasonable approach to small area analysis would be to match the data collection, and the data testing, to the potential reason for variability of interest. For example, if access to health care is the source of variation sought, then the regions should correspond to planning bodies responsible for access to care delivery. If the populations of these regions differ in morbidity rates (after age and sex adjustment), or in other characteristics affecting surgical use rates, then differences in access should only be invoked as a cause of variation if the observed variation exceeds that attributable to differences in population characteristics.

1.3 Statistical Issues

In the foregoing discussion, variation between regions has been discussed without reference to how it is measured and determined. A certain amount of variability in observed surgery rates would be expected by chance, even if true underlying probabilities of surgery were the same in each region. The purpose of statistical analysis is to summarize observations in a usable fashion, and describe how likely it is that a given set of observations arose through chance, without any systematic pattern.

In small area variations studies, the statistics which provide easily understandable summaries of the data are from nonstandard distributions, and it is not usual to assess

the role of chance in producing the observations. The only commonly used statistic that allows for testing of hypotheses is the chi-square, but the value of chi-square itself does not provide a meaningful summary of the data. A brief review of the commonly used small area statistics follows.

Wennberg's original article (3) presented the variations in a table of the two highest and two lowest rates among the 13 hospital service areas. The reported rates were standardized by age. In the text, the highest rate as a percentage of the lowest rate was used to dramatize the size of the range. This ratio, now called the *extremal quotient* (EQ), has pervaded the small area variations literature since. Although the numbers are not provided, the text reports that chi-square tests of significance were also performed on the data.

Because the extremal quotient is range driven and sensitive to outliers, investigators began to include other summary measures in small area studies. Initially the *coefficient of variation* (CV), which is the ratio of the standard deviation to the mean, was popular (16). The overall standard deviation used is a weighted average of the observed standard deviation in each region, with weights proportional to region population. In introducing the coefficient of variation, Chassin makes the important observation that the 'differences summarized ... were not produced simply by one or two atypical sites' and that the 'variations cannot have been due to the behaviour of a few physicians or groups of physicians'. In 1982, McPherson introduced the *systematic component of variation* (SCV). This statistic was designed to subtract random variability from total variability, and give an estimate of how much

variability was nonrandom(2). Both the CV and the SCV became widely published standard summary measures in small area variations research (28). Formulae for calculating these statistics, and others described later, are appended (appendix 1).

None of the above measures lends itself easily to hypothesis testing, since their sampling distributions are not known. The chi-square statistic has become commonly used in small area variations literature to fill this void(28,29). A k by 2 chi-square test can reject the hypothesis that the k proportions in a table are equal. A series of two by two chi-squares (with appropriate adjustment for multiple testing) can flag statistically significant outliers.

The literature review for Monte Carlo simulation is restricted to its application to small area variations. In 1990, an assessment of the expected values of the aforementioned statistics (EQ,CV,SCV, and chi-square) under the null hypothesis was published(29). This study used Monte Carlo computer simulation to generate sets of rates meant to represent surgical or medical procedures. The null hypothesis was that the true rate in each region was equal for each set. The effect of varying the population structure (from uniform counties of 100,000 each to the true population structure of Washington State), the rate (from 50 to 10,000 per hundred thousand), and the probability of readmission or a repeat procedure was studied. The expected values of the EQ, the CV, and the SCV varied markedly with these factors, even when true variability between regions was zero. For example, using Washington state populations the 95th percentile of the extremal quotient was 11.08 for a surgery rate of 50 per 100,000, but it was 2.62 for a surgery rate of 500 per 100,000. The

corresponding values for the CV were 1.58 and 0.49, and for the SCV were 100.34 and 8.09(29). These differences in the presence of no true variability between regions limits the interpretability of these statistics. The behaviour of the chi-square statistic was more predictable provided readmissions were not possible. The article recommends against continued use of the EQ and SCV(29). In later work, Cain and Diehr used a Monte Carlo simulation method to estimate the power of the EQ, CV, SCV, and chi-square to detect various non null distributions based on the back surgery rate in Washington State(20).

The chi-square test may be preferred, but it is not perfect. The test statistic can be compared to expected values, but it has no inherent numerical meaning. Furthermore, the chi-square is highly sensitive to small deviations from equality when the expected number of events is large(24). Consequently, the chi-square test almost always shows a statistically significant deviation from equal proportions when applied to small area variations data, even if the deviations themselves are smaller than would be considered important in policy decisions. A test that is almost always positive in this way lacks discriminatory power, and does not add to simple inspection of the data. If the number that the test is based on also lacks inherent meaning, it is fair to ask if a better test exists.

In seeking better summary statistics for small area variations, I have looked at measures currently applied to describe socio-economic inequalities in health, because the problem seems similar. Kunst and Mackenbach were commissioned by the World Health Organization to review methods of summarizing socioeconomic inequality in

health. They present 12 different measurements of inequality that might be used(30). Six of them are based on regressing health status on at least an ordinal ranking of SES, so they do not apply because regions do not sensibly rank ordinally. The other six, however, can be adapted for use in small area variations(31). The first two are the relative and absolute differences between the health status of high versus low SES groups. This relative measure is analogous to the EQ in small area variations. In conventional epidemiology, this would correspond to relative risk. The absolute difference between the lowest and highest groups would correspond to the attributable risk. The third and fourth measures described by Kunst and Mackenbach are the absolute and relative versions of population attributable risk. The analogue for small area variations is computing either the total excess provision (if the lowest rate is believed correct) or total underprovision (if the highest rate is believed correct). If empirical evidence existed showing the correct surgical rate, then this type of summary would be useful in small area variations. In general the correct rate of surgery is not known, and this limits the applicability and interpretability of the population attributable risk concepts.

The fifth and sixth measures described are absolute and relative versions of the index of dissimilarity. This index sums up how much redistribution would have to occur in order for everybody to achieve equal health status at the level of the current mean. The small area analog would be computing the relative or absolute number of cases one would need to redistribute in order to have the same surgical rate in each region. Although this type of measure has not been used in small area variations to my

knowledge, it has immediate appeal. The single number will summarize the overall variability, as do the CV and SCV. Unlike these statistics, the index of dissimilarity has an immediate interpretation for policymakers. The index reports the number of patients involved in the observed inequality. In areas with rates below the mean, these are patients not operated upon. In areas with rates above the mean these are additional operations performed. The index of dissimilarity can be thought of as a 'Robin Hood' type of index(32). Resources could be shifted to perform more cases in the low use areas and fewer in the high use areas. The minimum number of cases to redistribute to achieve rates equal to the mean in all regions is one half of the case count. As defined by Kunst and Mackenbach, the index of dissimilarity has this coefficient of ½ built into the calculation. A case count translates easily into resources needed, and addresses equality of access concerns important to Canadians. If small area variations are seen as indicating a problem of distribution, this single number summarizes the size of the distribution problem, either in absolute terms, or relative to total provision. Unlike the EQ, it will not be driven by extreme rates that occur in smaller regions.

Although the case count is a new measure in small area variations, its use is supported by methodological work in the area. Park illustrates his arguments regarding inappropriate use with graphs that have shaded areas corresponding to the case count, but does not use them numerically as such(18). Wennberg, the originator of small area variations research, recently wrote:

"..in the 1980's, this country *(U.S.)* abandoned the idea of planning. This neglect is unfortunate. I think it is inevitable that there will be a return to the problems of how many resources there are and how much is enough. I want to propose the development of descriptive statistics of resource allocation and capacity as a priority area for research in the 1990s." (33)

1.4 Alternative Statistical Methods

Two alternative methods of analyzing small area variations data have received attention in the recent methodological literature. One is empirical Bayes estimation of rates, and the other is based on availability of person-level data for the complete population.

Empirical Bayes estimation is a method of converting crude rates or observations into smoothed sets of rates, which are meant to be the best estimates of the true rate in each region (34-37). The focus is on estimation, not on hypothesis testing, and it is assumed that the true rate is different in each region. Usually, a Poisson distribution of cases is assumed within regions, and the set of true region rates is assumed to vary according to a gamma distribution. The gamma distribution describes how the 'person-time' is expected to vary from county to county to produce a count of n operations. Its relationship to the Poisson distribution is similar to that of the negative binomial to the binomial distribution. Parameters for the Poisson and gamma distributions are estimated from the observed data. The best estimate of the

rate for each region is then produced as a blend of the observed value, and the value predicted by the prior distribution(34). Practically speaking the result is to narrow the range of the estimates of true rates, leaving fewer extreme rates or outliers.

The variance in the crude rates is decreased considerably by empirical Bayes estimation, but it is decreased very little either by age and sex adjustment, or by Poisson regression based rate estimation(34). Comparing empirical Bayes to standard rate estimates, based on the criterion of which best predicts future rates, has shown the empirical Bayes method to be superior, and the standard method to overestimate interregional variation by up to 70%(35).

Hierarchical logistic regression models are an application of empirical Bayes estimation techniques to the analysis of small area data (37).

The literature describing empirical Bayes techniques is at the stage of presenting and refining the method, but it has not yet been widely applied in small area variations studies. One significant issue regarding the approach is failure to justify the use of the prior distribution by explaining what it corresponds to in the real world.

Hypothesis testing in small area studies has typically relied on using a chi-square test. Writing in the American Journal of Epidemiology, Carriere and Roos expressed concern that '...any conventional procedure for analyzing rates that requires a parametric assumption may tend to underestimate the underlying variation in the data, leading to a too frequent rejection of the null hypothesis.'(38). They propose an alternative method of testing, wherein no parametric assumptions are made, and the

variance is calculated directly from person-level data within each geographic and age/sex stratum. This requires the availability of person – level data, rather than simply counts of total numbers of admissions. In the single admission, binary event case, the method reduces to the usual chi-square test. If the measure is not binary (eg. cost of care, length of stay), or if multiple admissions per person are possible (eg. medical admissions for chronic heart, lung, or kidney disease), then the variance calculated from the person level data is considerably different. Carriere and Roos propose a general method called the T-squared test for testing small area hypotheses when person level data are available for the whole population. This ideal condition holds, or almost holds, in the province of Manitoba, but in most other jurisdictions person level data are not yet available for the performance of small area analyses(39).

1.5 Policy Issues

Health policy results from decisions and actions of many different people and organizations, working at different levels within the hierarchy. Some are concerned with a single hospital or region, others for overseeing the provincial system. The distributed nature of policymaking requires that the data important for setting policy be available to all, in an interpretable and accessible fashion(40,41).

Although health services research is of interest to policymakers, those who produce research are cautioned that 'in the policy world, simplicity is eloquence ... study results should be summarized in nonstatistical terms ... at all costs, avoid the use of more technical mathematical notations' (40). The results should be presented in a

usable form, interpretable by policymakers. Wennberg provides some examples of how to summarize small area variations data drawn from practical experience with the Iowa plan to reduce hospital costs. In this context, numbers of cases judged excessive were estimated, and their costs were calculated based on unit care costs(5).

Policy makers are interested in absolute numbers and impact on population, if the aim is to achieve equity in access. Policy makers are interested in costs and therefore need to know numbers of cases if marginal costs are known. In Ontario, marginal cost - benefit analysis is potentially important because there is more infrastructure built than is currently being used (wards are empty and hospitals are closing), so it is possible to both add and subtract services according to marginal and not capital costs.

1.6 Objectives

The methods and results sections of this thesis will explore Monte Carlo simulation as a tool for small area variations studies. The focus will be on analyses relevant to health policy. The role of simulation is to improve interpretation of observed values of the summary statistics, by providing an appreciation of their expected distributions. Specific objectives are:

1. To use simulation to assess the currently used small area statistics. The probability sampling distributions of EQ, CV, and SCV will be studied under a wide range of surgical rates, population sizes, and population divisions applicable to Canada's health system.

2. To define and describe a new small area statistic tailored to policy questions. The new statistic will be based on counting the cases involved in an observed maldistribution. This makes it relevant both to population health and economic questions.

3. To incorporate a prior distribution and allow modification of the standard null hypothesis for small area analysis. Instead of a null hypothesis of strict equality of rates in each region, a modified null hypothesis allowing small amounts of true variability between regions will be applied. This is meant to represent variability contributed by differences in the population between regions.

Objective 1 represents an application of the previous simulation work of Cain and Diehr (29) to Canadian circumstances. Objectives 2 and 3 are completely novel to this thesis.

2 Methods

2.1 Overview of Methods

A computer program to perform the required simulations was written. The situations to be simulated, and the parameters for each simulation, were defined based on Ontario populations and surgery rates, with extensions to allow application and interpretation in smaller Canadian provinces. The expected distributions of the EQ, CV, SCV, chi-square, and the newly defined case count were simulated for each situation. Presentation of results focused on describing the value of the 95th percentile of the simulated result, as this corresponds to the usual 5% alpha level for rejection of a finding based on chance.

Simulations were then run to estimate the power of each statistic to detect various patterns of true variability. Finally, a group of simulations were run to estimate the expected distributions of the statistics under a modified null hypothesis, which allowed small true rate variations between regions. This was instead of the usual null hypothesis, which is based on exactly equal true rates in each region.

2.2 Computer Programming

A program to perform Monte Carlo simulations of surgical rates was written in Visual Basic. This program runs as a macro in Microsoft Excel, as spreadsheets are used for the input and output. The program code is attached as an appendix. A brief explanation of the program follows.

The program is modular, with the modules arranged hierarchically. The lowest level procedures are the case number generators. These routines randomly generate simulated numbers of cases. Inputs are the numbers of regions (k), the number of simulations requested (i), and two vectors. The first vector contains the population of each region, and the second contains the true rate (in cases per 100,000 population) of surgery for each region. When testing the null hypothesis (all region rates equal), this second vector will contain k identical numbers. Four case number generators were written (bingennum, Poissongennum, normgennum, and Poissonreadmit). The first three were for the situation of one admission only per patient, using the binomial distribution, the Poisson approximation to the binomial, or the normal approximation to the binomial to generate case numbers. The fourth case number generation routine was written to allow for readmission of the same patient in the same year. This routine generates an initial number of cases using a Poisson approximation to the binomial, and then generates a number of readmissions out of these cases using a binomial random number generator. The details of this fourth routine are further discussed below in the section on multiple admissions.

After the case numbers are generated, a simple routine (numtorates) converts the numbers to rates per 100,000 for each region by dividing by the region populations. Case numbers are preserved for convenience, as they are used in subsequent calculations.

The next procedures calculate the summary statistics from the simulated surgical rates and case numbers. These procedures are named eq, cv, scv, and idchi for the statistics

they calculate. The same routine calculates both the index of dissimilarity (absolute and relative) and the chi-square statistic because these calculations have many steps in common.

The output procedure (output) summarizes the distributions of each statistic by recording the mean, standard deviation, minimum, maximum, and 95th percentile. For certain applications (such as power testing) the deciles of each statistic are recorded as well (outputpercentiles). This output is saved along with the input scenario so it can be graphed and analyzed. If it is necessary to look at the full distribution of the simulated statistics (for instance to provide pseudo p-values for an observation), then the entire file of observations from a single simulation can be saved. The histogram function in Excel is then used to sort the data by single percentiles used to determine pseudo p-values.

The highest level routines automate some common tasks, such as setting up input scenarios and running multiple simulations sequentially. These are not described in detail, and most of their design is determined by the methods Excel uses to manage data. The highest level routines are often written specifically for each program run, but the lower level procedures remain identical.

2.2.1 Validation of Random Number Generator Routines

The random number generation tool from the Microsoft visual basic analysis toolpack was used. This provides routines for binomial, Poisson, and normal random number generation. Strictly speaking, these numbers are pseudorandom as are all computer

generated 'random' numbers based on algorithms. A different seed was used for each simulation run of the program, but the same seed was used for each run during the testing and debugging phases. The routines were tested by generating sequences of random numbers and performing Kolmogorov - Smirnov tests (in SPSS version 7.5) to confirm that they conformed to the correct distribution(42).

2.2.2 Confirmation of Poisson Approximation to Binomial:

The underlying process is binomial, but generating binomial random numbers by repeated Bernoulli trials is computationally inefficient once n is large(42). Therefore, the first step was to define the role of the Poisson approximation to the binomial, and of the normal approximation to the binomial. Because the statistics calculated, in particular the EQ, reflect the outliers in the distribution it is not enough to know that the approximations to the binomial have the correct mean. They must also have a very similar pattern in the tails of the distribution. Previous simulation work (29) has used a normal approximation to the binomial for expected case numbers more than five, and an exact binomial for fewer expected cases. However, there is nothing in the methods section of that paper to support the arbitrary cutoff of five cases as a reasonable level to begin using the normal approximation to the binomial.

Accordingly, a series of simulations was performed to produce 3000 sets of numbers of cases, with an underlying binomial distribution. A rate of 100 cases per 100,000 population was used, and populations provided were 1000, 2000, 3000,...9000, 10000, 20000,...90000, 100000, 200000,...1000000. This yielded 28 expected values of the

number of cases, ranging from one to 1000. The Poisson approximation and normal approximation to this binomial distribution were also generated. Calculations were stopped for the binomial calculation at the 100000 population, because the computational time is a linear function of the number of trials and becomes unwieldy. Kolmogorov - Smirnov tests were used to compare the data generated to what would be expected from both normal and the Poisson distributions. The significance value returned by the Kolmogorov - Smirnov test varies from zero (very little correspondence between data and distribution) to one (almost perfect correspondence between data and distribution). There is no standard cutoff for rejecting data as coming from a distribution, but I used .2 as a cutoff. This is an easy test to pass. Data that fail this test can be considered quite unlikely to come from the comparison distribution. Histograms of the distributions were compared graphically. Finally, the summary statistics EQ, CV, and SCV were calculated based on Ontario populations and 180 cases per 100,000 using each of the random number generators (binomial, Poisson, and normal), and the results of the simulations were compared. The rate of 180 cases per 100,000 corresponds to 30 expected cases in the smallest region. This is the smallest case number at which Kolmogorov-Smirnoff testing suggested reasonable approximation of the Poisson by the normal distribution.

Based on the results (presented later) of the testing just described, the Poisson approximation to the binomial distribution was used for all further simulation runs.

2.3 Data Requirements for Simulations

The simulation program requires two basic items of input. The first is a vector of k region populations. The second is a vector of k true rates of surgery. Under the standard null hypothesis (no rate variability between regions) this second vector will be k identical rates.

As a reference point, Ontario populations and surgical rates are used as a starting point for simulation. To permit generalization, further simulations were performed with smaller populations down to 500,000. The ICES Atlas reports rates of nineteen different surgical procedures across thirty three regions defined by district health council of residence. I have selected five of them as case studies to carry through the analysis. These are representative of the spectrum of realistic scenarios, as they include the lowest and highest rate procedures as well as some of intermediate rate. The five procedures are abdominal aortic aneurysm repair (AAA), coronary artery bypass grafting (CABG), total hip replacement (THR), uterine dilatation and curettage (D+C), and transurethral resection of the prostate (TURP). The 1992 Ontario mean rates of these procedures are given in table 1, and a graphical summary of the age and sex adjusted rates in each region is given in figure 1. Abdominal aortic aneurysm was selected since it has one of the lowest mean surgical rates. Coronary artery bypass grafting was selected as a procedure with an intermediate rate and high variability, total hip replacement as a procedure with an intermediate rate and low variability. TURP and D+C were selected because of high mean rates, and because each operation applies to a subpopulation.

Table 1 Ontario Age and Sex Adjusted Surgery Rates

Ontario age and sex adjusted surgery rates '89-'92

Procedure	Rate	
Abdominal Aortic Aneurysm Repair	22.9	per 100,000 age 20+
Coronary Artery Bypass Grafting	68.8	per 100,000 age 20+
Total Hip Replacement	76.2	per 100,000 age 20+
Uterine Dilatation and Curettage	538	per 100,000 age 20+ female only
Transurethral Resection of Prostate	1402	per 100,000 age 50+ male only

The ICES atlas is available in both print and electronic forms, but neither gives the actual region populations broken down by age and sex. The electronic version does supply sufficient information to allow them to be calculated. As well as age and sex adjusted rates for each region, the electronic atlas supplies the crude rates, and the numbers of cases performed. These were used to calculate regional populations. For the adult population (20+) the cholecystectomy rates were used, since they are the highest rates, so they give the most accurate population figures. The adult female (20+) population was calculated from the case numbers and crude rates for D+C, and

the older male population (50+) was calculated from the case numbers and crude rates for TURP. 1989 to 1992 data were used in all cases as data from later years were incomplete.

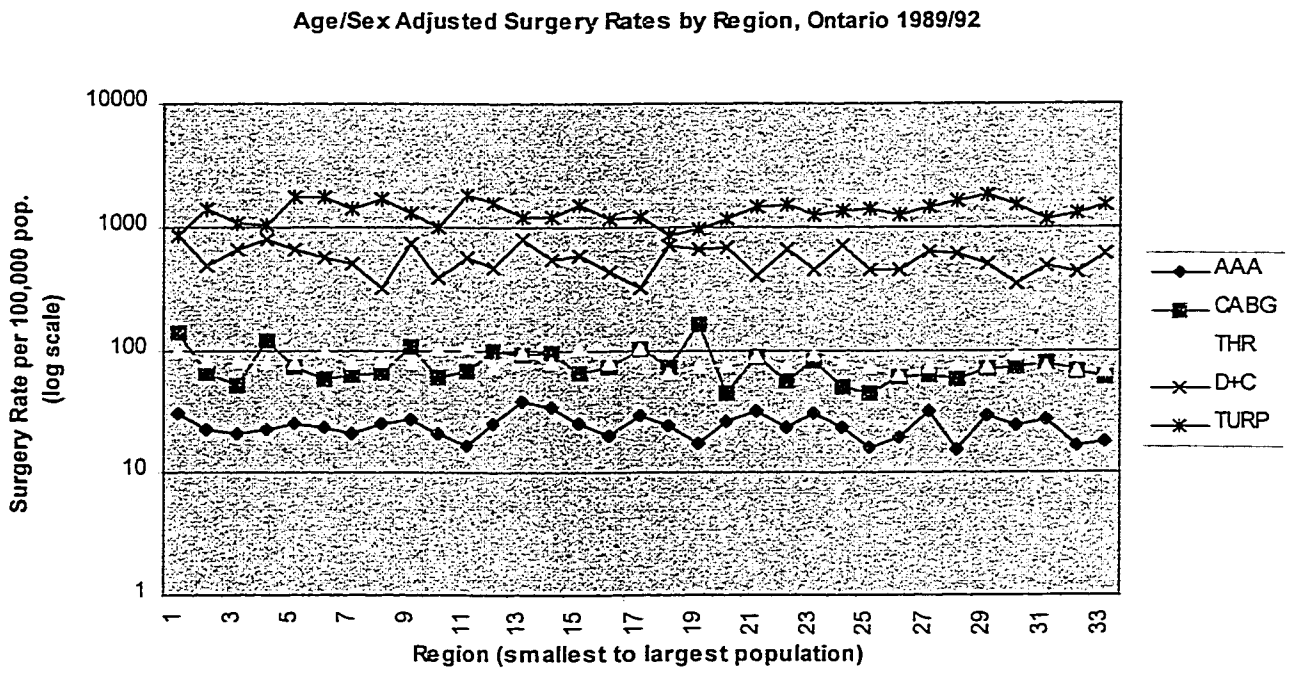


Figure 1 Age and Sex Adjusted Surgery Rates By Region

2.4 Varying Surgical Rate

The first set of simulations studied the effects of varying the mean rate of surgery on the expected values of the statistics EQ, CV, SCV, and chi-square. For this set of simulations, the Ontario adult population (age 20+) was used, divided into 33 regions according to District Health Council of residence. The surgical rates considered were rates of 10, 20, 50, 100, 200, and 500 cases per 100,000. These rates were chosen to span the range of actual surgical rates of interest in the procedures listed in table 1. A total of 3000 iterations were run to generate 3000 simulated sets of case numbers and rates, and the EQ, CV, SCV and chi-square statistics were calculated for each rate. The mean, standard deviation, minimum, and maximum as well as the 95th percentile of each statistic were then calculated from the 3000 values and recorded. In general, the 95th percentile is the value graphed and analyzed since it is the most commonly used cutoff for hypothesis testing. It is expected that the values of statistics will change as their parameters vary. The focus of the description of how the values change is the interplay between clinical and statistical significance.

2.5 Varying Population Size

The population at risk for many of the operations is only a fraction of the full Ontario population. Gynecological surgery, for instance, draws only from the adult female population. Surgery for prostatic urinary obstruction draws only from males, but

predominantly from males over fifty years of age. The population sizes based on these demographic characteristics were used to study the effects of a smaller total population, with the same division into 33 regions. Again, surgical rates of 10, 20, 50, 100, 200, and 500 were used in order to span the range of surgical rates of interest. Surgical rates were set to be equal in all of the 33 regions. Another way of considering these subpopulations is as representative of the entire population of a smaller province. The Ontario population structure was used in order that the results have immediate relevance. Varying population structures was done to generalize the results. In order to study even smaller population units, hypothetical small province rates for populations of 1,000,000 and 500,000 were also studied. Each of these was divided into twenty regions, assuming that smaller provinces have fewer administrative boundaries.

2.6 Varying Population Structure

The region populations vary from 15,944 to 1,851,913 (adults 20 and over). If 33 equally sized regions are formed, each has a population of 240,133. These population structures were used as extremes, and a series of eleven population structures was created by blending them. A scaling factor ranged from 0 to 1 in steps of .1, and for each scaling factor, each region was given a population equal to the mean plus the scaling factor times the difference between the true region population and the mean. The population structures thus created are shown in figure 2. This provides a method of holding the total provincial population constant, and divides it in ways that are

more and more unequal, with the most unequal division being the current actual region sizes. The purpose is to investigate when the presence of small regions affects the expected value of the statistics. Surgical rates of 10 and of 50 per 100,000 were used, because prior tests had shown that the statistics vary more markedly with small regions and low surgical rates.

The other method of varying population structure is to divide the population into a different number of regions. Simulations were performed based on dividing the Ontario population into six, eleven, seventeen, or thirty three equally sized regions. These numbers are somewhat arbitrary. They represent the strategy of combining sets of two, three, or six current regions into larger administrative bodies. A surgical rate of 50 per 100,000 was used. Equally sized regions were used because prior results showed that only the EQ and SCV were sensitive to how the population is divided up. The expected values of the other statistics are unaffected by equal versus unequal population divisions.

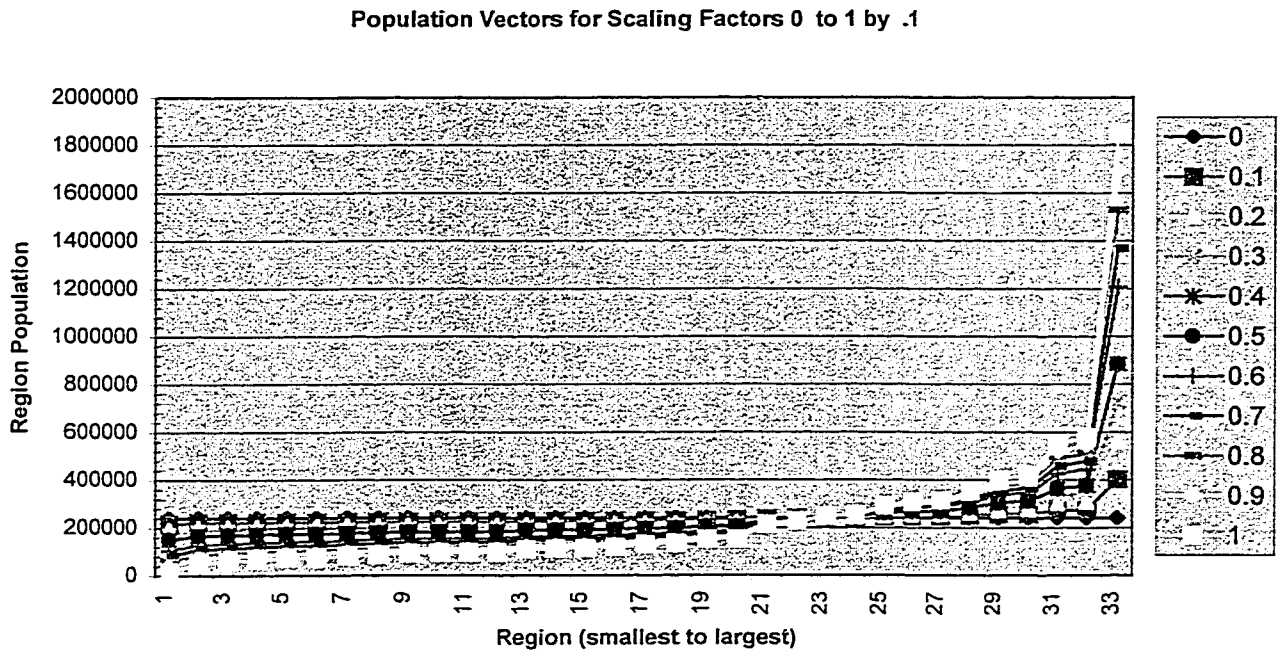


Figure 2 Population Vectors for Scaling Factors 0.1 to 1

2.7 Allowing for Readmissions

One of the largest potential sources of deviation from a binomial assumption is the possibility of multiple admissions per patient for the same procedure. This is not a concern with 'ectomy' type operations where a nonpaired organ is fully removed, since a reoperation on the same patient is not possible. This is not true in the case of prostatectomy, during which the organ is not fully removed, and reoperations are common (11). On the other hand, for medical diagnoses such as congestive heart failure or chronic obstructive pulmonary disease, multiple admissions per year for the same patient are sometimes the norm. Most of the surgical procedures considered are intermediate. That is, although readmission is possible, more than two procedures per patient per year are unlikely. To allow for readmission, a fourth case number generating routine was written. Prior testing had established that the Poisson approximation to the binomial was appropriate for estimation of case numbers in realistically sized regions, so the routine initially generates a number of 'primary' cases based on the Poisson distribution. The 'true' rate is appropriately deflated so that it becomes correct once readmissions are added. This small number of cases is often too small a population for application of the Poisson approximation to estimate readmissions, so a number of readmissions based on the binomial distribution is calculated, using the simulated case number from the Poisson as the input number of trials. The probability of readmission can be varied arbitrarily. For illustration purposes, a set of simulations with a 15% probability of readmission was performed.

This is a realistic upper limit for the number of readmissions in the surgical situation, as argued in the discussion section. Each of the three populations of interest was studied (adults 20+, females 20+, and males 50+), and for each population the effect of 15% readmission rate was studied with surgical rates of 10, 20, 50, 100, 200, and 500 per 100,000 population.

2.8 Case Count as a New Statistic

The case count was investigated as a potentially useful new statistic in summarizing small area variations. The case count is an adaptation of the index of dissimilarity described by Kunst and Mackenbach(30). The intent of the case count is to summarize the overall variability in a distribution in a meaningful and interpretable way. The case count (CC) is the sum of the absolute differences between the expected and observed number of cases in each region. The expected number is based on the supposition that the true rate of surgery is the same in each region, and is equal to the overall mean rate of surgery.

The case count can also be expressed as a proportion of the total number of cases performed. In this form it will be referred to as the case proportion or abbreviated as CP. The CP is dimensionless, as is a coefficient of variation. The advantage of expressing the CC in relative terms like this is that it may be more appropriate for comparing the overall variability of different types of operations with different base rates.

The relative version of the CC, or CP, is the one for which simulation results are provided. The statistical properties of the CC are similar, since it differs only by a constant multiple for any situation.

In the social inequalities in health literature, the index of dissimilarity is added up in the way described, but then divided by two. The reason for this is that the amount to be redistributed has been counted twice in the index of dissimilarity or case count – once as a positive for those who have more than the mean, and once as a negative for those who have less. Therefore dividing by two defines the amount which needs to be redistributed. It is still possible to do this with the CC the way it is used here, simply by dividing the figure by two. However, a choice was made to compute the CC without this division. This is because redistribution of resources by moving cases from high use to low use areas is only one potential response to small area variations. It is only the correct response in the limited (and difficult to establish) circumstance that the mean rate of surgery is the 'correct' rate of surgery. Dividing the case count by two overemphasizes the importance of the current mean rate, which may be far from the 'correct' rate, so a decision was made to present the case count without division by two.

For each of the simulations performed to describe the behaviour of the currently used statistics, the value of the CP was also calculated and summarized.

2.9 Testing Power of Various Statistics

Statistical power describes the probability that a statistic will detect a non null distribution. Before the CP can be promoted as an alternative to EQ, CV, SCV, and chi-square, the power to detect true variability in surgical rates should be assessed for each statistic. Simulation was used to estimate the power of the statistics in selected examples of interest, and to rank the statistics in terms of their power. More general power questions require either analytical solution, or very extensive simulation.

Two scenarios were used to test statistical power. The first was to use the actual distribution of rates observed for Coronary Artery Bypass Grafting. The variability among observed rates for this procedure are statistically significant at the .95 level for each simulated statistic. Sets of rates were created between the extremes defined by equal rates at the mean for all regions, and the observed rates for all regions. Intermediate rate sets were created by defining the rate for each region as:

$$\text{Rate} = \text{mean rate} + \text{scaling factor} * (\text{observed rate} - \text{mean rate})$$

for scaling factors of 0 to 1 in increments of 0.1.

The second scenario was uniform rates for 32 regions, with an outlier in the 33rd region. The outlier was designed to be either low or high, and to vary from 10% to 100% of the true rate in steps of 10%, and then from 100% to 1000% of the true rate in steps of 100%. A variant was run with the outlier in the largest region (population 1,851,913) and one with the outlier in the smallest region (population 15,944).

To compare the power of different statistics, first a simulation was run with equal rates in all areas. The 95th percentile of the simulated values was used to define a threshold for rejection of the null hypothesis. One sided hypothesis testing is appropriate since each of these statistics increases monotonically with increasing variability, and the 'left tail' of each distribution either does not exist or describes a circumstance which is not of interest – that of less variability than expected. The power of a statistic is the proportion of the time that statistic rejects a false null hypothesis. To estimate this by simulation, the output routine was modified to record the deciles for each statistic, as well as the fifth and ninety fifth percentiles. The fifth and ninety fifth percentiles were included simply to improve plotting of the power curves, since all of the graphs are sigmoid curves and more points are needed on the ordinate axis where the abscissa changes quickly. The fifth and ninety fifth percentiles imply nothing about hypothesis testing in this context. A simulation was then run with each non null scenario defined above. The threshold value was compared to these deciles obtained from each simulation of a non null distribution, allowing power to be estimated (to the closest decile) for each statistic. The results of this power analysis were presented graphically to allow comparisons between statistics.

2.10 Analyzing the actual rates of various operations

Monte Carlo simulation allows simulation based hypothesis testing to be performed, even for statistics whose distribution under the null hypothesis is not known. This

means that the EQ, CV, SCV, and CP of observed data can be tested against the simulated distribution. The 95th percentile of the simulated distribution was used as a cutoff point, and any value exceeding this can be judged as statistically significant at the five percent level when compared with the simulated distribution. Simulation of the chi-square was also performed as a check on the method. The value obtained from the simulations should be the same as the cutoff value for chi-square obtained from tables or from standard calculations.

The standard null hypothesis used is that the age and sex adjusted rates of surgery are equal for all regions in the province:

H_o(standard): The true rates of surgery in each region are equal.

Simulation was performed for five different types of surgery - abdominal aortic aneurysm repair (AAA), coronary artery bypass grafting (CABG), total hip replacement (THR), uterine dilation and curettage (D+C), and transurethral resection of the prostate (TURP). The observed values of the statistics were calculated from the age and sex adjusted data, using the same subroutines that calculate the statistics based on simulated cases. These observed values were checked and found to correspond to those published in the atlas for the EQ, CV, and SCV. This simply confirms that the routines written to calculate the values of these statistics agree with ICES calculations given the same raw data. ICES did not use any simulation to describe expected distributions of these statistics. There were slight discrepancies between the calculated and the published chi-square, because the published chi-square

was calculated as a likelihood ratio rather than a simple chi-square. The differences were very slight and had no bearing on the interpretation of results.

The output of the simulation was then used to perform hypothesis testing and to calculate pseudo p-values for all of the statistics, for the observed surgery rates. The phrase pseudo p-value is used in place of p-value to signify that the underlying distribution is estimated from simulation rather than calculated analytically, but the usage and implications of a pseudo p-value are otherwise identical to those of a p-value.

Hypothesis testing is straightforward. The simulation program calculates the 95th percentile of the distribution of each statistic, and if the statistic calculated from the observed data exceeds this, it is declared statistically significant at the 5% level.

Pseudo p-values are more complex than hypothesis testing. The statistic calculated from the observed data must be compared to the simulated distribution, and its percentile determined. Fine cumulative frequency histograms (300 divisions) were made from the simulated distributions of each statistic, and used as lookup tables to determine the percentile of the simulated distribution that the observation corresponds to.

2.11 Modifying the null hypothesis

The standard null hypothesis in small area variations is that the age and sex adjusted rate of surgery is exactly equal in each region. This assumes that age and sex adjustment fully accounts for morbidity differences between populations, an

assumption which is known to be false(43-46). Standard means of hypothesis testing (ie the usual chi-square test) cannot, however, allow modification of the null hypothesis to allow small amounts of true intraregional variability to be present under the null hypothesis.

An advantage of simulation based hypothesis testing is that the null hypothesis can be specified in any way, and simulation can determine the expected values of the statistics under the null hypothesis. This allows a small amount of true intraregional variability to be present under the null hypothesis. One way of thinking of this true intraregional variability built into the null hypothesis is that it accounts for differences in population characteristics (e.g. morbidity) not captured by age and sex adjustment. An analogy can be drawn to a 'random effects' model, where the true mean surgical rate in each region varies somewhat.

Two types of data were used to generate null hypotheses that allow for a small true intraregional rate variability. One is data on appendectomy surgery, and the other is the standardized mortality ratio. Full discussion and justification of the reasons for using these measures is deferred to the discussion. Briefly, appendectomy data represent an operation for which controversy over indications is minimal(47), and for which access barriers should not exist in any region of Ontario. Standardized mortality ratios are the routinely used proxy measure for differences in population morbidity used in other applications such as the Resource Allocation Working Party model for population based funding(44-46).

The general plan for using appendectomy and SMR rates is as follows. The CV and/or CP are used as the measure of variability. The expected values of the CV and CP are generated by simulation. The simulation uses, as its input, an overall surgical rate equal to that of the operation of interest; but also a pattern of true variability based on either appendectomy or SMR data. The observed CV or CP can then be compared to expected values generated under a modified null hypothesis where a small amount of true interregional rate variability exists. Rejection of this modified null hypothesis is likely to be more clinically meaningful than rejection of the standard null hypothesis. Details of the steps in this hypothesis testing follow.

The 'true' rates of appendectomy in each area which produce the observed rate distribution must be computed. The observed rate distribution is a result of 'true' underlying variability, plus variability from sampling. The simulation program adds variability from sampling. Therefore the correct set of 'true' rates is that one which, when provided to the simulation as input, gives output corresponding to the observed data. This set of rates was determined by repeated trials. It was found that using a linear multiplier of .93 on the absolute difference between the mean rate and the observed rate in each region produces a distribution with the same CV, SCV, chisquare, and CP as the observed age and sex adjusted rates. (see table 2) . The EQ differs only slightly. This distribution is hereafter referred to as the distribution of true appendectomy rates. Dividing each rate in this distribution by the mean appendectomy rate gives a relative distribution of true appendectomy rates.

Correspondence between statistics of observed appendectomy rates, and those after true' rates scaled to .93 of observed have variability added by simulation.						
	eq	cv	scv	chisq	cc	cp
appy observed rates -age sex adjusted	*2.0288*	*0.1624*	*41.9678*	*244.90*	*1177.45*	*0.1267*
'true' rates scaled to .93	2.1368	0.1632	44.7110	248.67	1161.56	0.1250

Table 2 Derivation of appendectomy rates without sampling variability.

Once the 'true' rates of appendectomy are known, the relative distribution of true appendectomy rates can be used to generate appropriate null distributions for hypothesis testing for other surgical rates. The null distribution for a given operation is generated by multiplying the mean rate for the specific operation by the relative distribution of true appendectomy rates. This null distribution corresponds to a null hypothesis of intraregional variability no greater than that observed in appendectomy rates, rather than to a null hypothesis of no intraregional variability at all.

> H_o(appy): The intraregional rate variability observed in operation X is no greater than the intraregional variability in appendectomy rates.

It is important to appreciate the difference between this null hypothesis, and that tested by the routine chi-square. The expected values for the chi-square test can certainly be specified to test for correspondence to unequal rates in each area. However, the expected rate in each region must be uniquely specified a priori. This is quite restrictive. It is unlikely, for example, that the population factors affecting the rate of CABG surgery exactly follow those affecting the rate of appendectomy region by region. This, though, is all that a standard chi-square can test. The chi-square can test an observed distribution against only one, uniquely specified, null distribution at

a time. This null distribution may, of course, have equal or unequal rates in each region. The question of greater interest is whether the overall variability in coronary artery bypass grafting, across thirty three regions, is greater than or similar to that seen in appendectomy. This requires simultaneous testing against a whole family of related non null distributions. Simulation can answer this question, provided a measure of overall variability is agreed upon. Using either the CV or the CP as this measure is justifiable, and allows simulation based hypothesis testing to incorporate data from a 'control' procedure.

As an example, consider four equally sized counties with numbers of appendectomy cases of 16,18,22, and 24. The case count is 12, and the case proportion (CP) is 12/80 or .15. Fifteen percent of the cases performed would have to be redistributed to achieve equal rates in each region. Consider two sets of bypass surgery rates. One is 16,18,22, and 24; the other is 24,22,18,16. Both of these have a case proportion (CP) of .15 as well. Many other sets of rates with the same CP of .15 can be constructed. The overall rate variability in any of these sets of bypass rates is the same as that in the appendectomy rates. Using a case proportion as the test statistic, and appendectomy rates as the standard, none of these distributions would be considered statistically significantly different from the appendectomy rates. Using chi-square as the test statistic, the first set would be found not to differ, but the second would be found to differ significantly ($p = .04$). Rather than testing overall variability, the chi-square is testing for correspondence to a particular pattern of variability. Simulation based hypothesis testing can test overall variability, using either CV or CP as a

measure for overall variability. Simulation is required because hypothesis testing requires allowance for the role of chance.

The simulation program was used to generate the expected distributions of the EQ, CV, SCV, chi-square, and CP under the modified null hypothesis. The 95th percentiles of the simulated distributions were then compared to the observed values for each statistic to test for compatibility with the modified null hypothesis.

The Standardized Mortality Ratios were available for each region. These were used as a second source of a modified null hypothesis. The null distribution was generated by multiplying the mean rate by the standardized mortality ratio for each region. (The standardized mortality ratios were not 'scaled' in the manner of appendectomy data. This is both because SMR's are indirectly age standardized, and because I did not have death count or rate data that would have permitted this). This null distribution corresponds to a null hypothesis which states that intraregional variability is no greater than that observed in SMR's, rather than a null hypothesis which says there is no intraregional variability at all.

> H_o(SMR): The intraregional rate variability observed in operation X is no greater than the intraregional variability in standardized mortality ratios.

Again the simulation program was used to generate the expected distributions of the EQ,CV, SCV, chi-square, and CP under the modified null hypothesis. The 95th

percentiles of the simulated distributions were then compared to the observed values for each statistic to test the modified null hypothesis.

3 Results

3.1 Random Number Generator Tests

Kolmogorov Smirnov tests performed in SPSS confirmed that the output from the normal distribution random number generator in Excel conformed closely to a normal distribution, and that the output from the binomial and Poisson random number generators conformed closely to a Poisson distribution. Because these simply confirm accuracy of the routines, the numerical results are not reproduced here.

3.2 Appropriateness of Poisson Approximation

Table 3 gives results showing how closely the output from three different routines for generation of numbers of cases conforms to the Poisson distribution. Kolmogorov – Smirnov test results near 1 indicate a high likelihood that the data come from the reference distribution, results near zero indicate a low likelihood of this. (Ideally, one would be testing how well this output conforms to the binomial distribution but SPSS uses only Poisson or normal as a comparison – however it can be seen that the binomial distribution corresponds closely to the Poisson over the range of interest, making the Poisson the desired proxy test.)

Results of Kolmogorov – Smirnov Tests, see text for explanation				
Comparison Distribution	**poisson**	**poisson**	**poisson**	
Data Generated By:	**binomial**	**poisson**	**normal**	
Expected Number of Cases:				
1	**1.000**	**0.998**	0.002	
2	**1.000**	**1.000**	0.002	
3	**1.000**	**0.979**	0.004	
4	**1.000**	**0.998**	0.005	
5	**0.996**	**0.829**	0.008	
6	**1.000**	**0.982**	0.017	
7	**0.999**	**0.976**	0.026	
8	**0.991**	**0.994**	0.034	
9	**0.998**	**0.448**	0.030	
10	**0.973**	**0.748**	0.043	
20	**0.904**	**0.893**	0.168	
30	**1.000**	**0.800**	**0.333**	
40	**1.000**	**0.782**	**0.521**	
50	**0.983**	**0.858**	**0.607**	
60	**1.000**	**0.777**	**0.617**	
70	**0.986**	**0.882**	**0.646**	
80	**0.966**	**0.885**	**0.752**	
90	**0.985**	**0.877**	**0.771**	
100	**0.701**	**0.974**	**0.868**	
200		**0.991**	**0.906**	
300		**0.936**	**0.862**	
400		**0.764**	**0.897**	
500		**0.749**	**0.880**	
600		**0.965**	**0.943**	
700		**0.929**	**0.905**	
800		**0.864**	**0.895**	
900		**0.932**	**0.907**	
1000		**0.891**	**0.873**	

Table 3 Kolmogorov-Smirnoff tests for correspondence to Poisson distribution.

Data generated by both the Poisson and binomial random number generator routines very closely matches the Poisson distribution (significance close to 1) across the entire range of expected case numbers from 1 to 100. The normal random number generator routine generates data which do not conform well. Normal approximation initially generates negative case numbers, as well as noninteger case numbers. Neither is appropriate, so generated data are rounded to the nearest integer and negative case numbers replaced by zeroes before comparing to the Poisson distribution. Despite this, the normal approximation does not correspond to the Poisson approximation if the expected number of cases is 20 or fewer, and only weakly corresponds to the Poisson distribution if the expected number of cases is 30. Correspondence gets better at expected case numbers of 100 or greater.

To determine whether these discrepancies actually make a difference to the results of the simulation, a simple simulation of 1000 iterations was run based on Ontario populations, and a case number of 180 per 100,000. This case number was selected to make the expected number of cases in the smallest region (16,000 population) equal 29, close to the threshold at which K-S testing suggests substituting the normal approximation for the Poisson. As can be seen in table 4, using the normal approximation instead of the binomial or Poisson results in quite different 95th percentiles for the EQ, CV, and SCV. Both of these statistics are highly sensitive to the rate variability in smaller areas. The normal distribution not only does not correspond to the Poisson distribution according to the Kolmogorov-Smirnov tests

(Table 3), but it also does not produce corresponding results in the simulation (Table 4).

Example: 95th Percentile of Statistics Calculated Using Different Random Number Generators
1000 iterations,33 Regions, Ontario Population, Surgery Rate 180 per 100,000

	Binomial	Poisson	Normal
EQ	1.7000	1.6800	1.4100
CV	0.0576	0.0613	0.0432
SCV	3.6900	4.4400	0.5406

Table 4 Agreement between binomial and poisson results.

Based on these results, the Poisson approximation to the binomial process is used for all of the following simulations, with the exception that exact binomial calculations are used to generate numbers of readmissions.

3.3 Varying Surgical Rate

Figure 3 shows the results of varying the surgical rate from 10 per 100,000 to 500 per 100,000 using the Ontario adult population (20 years old and over).

The extremal quotient (EQ) varies markedly with the underlying surgical rate, even in the presence of no variability. An observed EQ of 5, for instance, would be outside that expected by chance for a surgical rate of 200 per 100,000; but it would be well within the bounds of chance for a surgical rate of 20 per 100,000. As the surgical rate becomes very low, the expected value of the EQ becomes higher. Small regions, with intrinsically more variable observed rates, are usually responsible for both the low and high rate outliers.

The systematic component of variation (SCV) behaves in a similar fashion to the EQ. As the surgical rate becomes lower, the threshold value of the statistic becomes higher, even in the presence of no variability. Again, an observed SCV of 5, 10, or even 100 does not have a straightforward interpretation. Whether this is more variability than chance alone depends on the underlying rate of surgery.

The coefficient of variation (CV) also has some dependence on the underlying surgery rate, but less so - the slope of the line is smaller, and the slope appears constant over the range of interest. Theoretically, the coefficient of variation should approach infinity as p approaches 0, and should approach 0 as p approaches 1. This asymptotic behaviour is never observed in realistic simulations. Rather, a nearly linear behaviour with a small slope is observed. The CV varies from 0.03 to 0.24, but for surgical rates of 50 and above it is expected to be less than 0.10.

The critical value for chi-square, as determined by simulation, remains constant across the surgical rates at 46. This corresponds to the value obtained from tables for $X^2_{(32.0.95)}$ (48) and from computation {Excel}. The fact that the simulation returns the expected threshold value for chi-square is important in confirming both the accuracy of the simulation process, and the adequacy of 3000 trials allowing the simulated estimates to be robust.

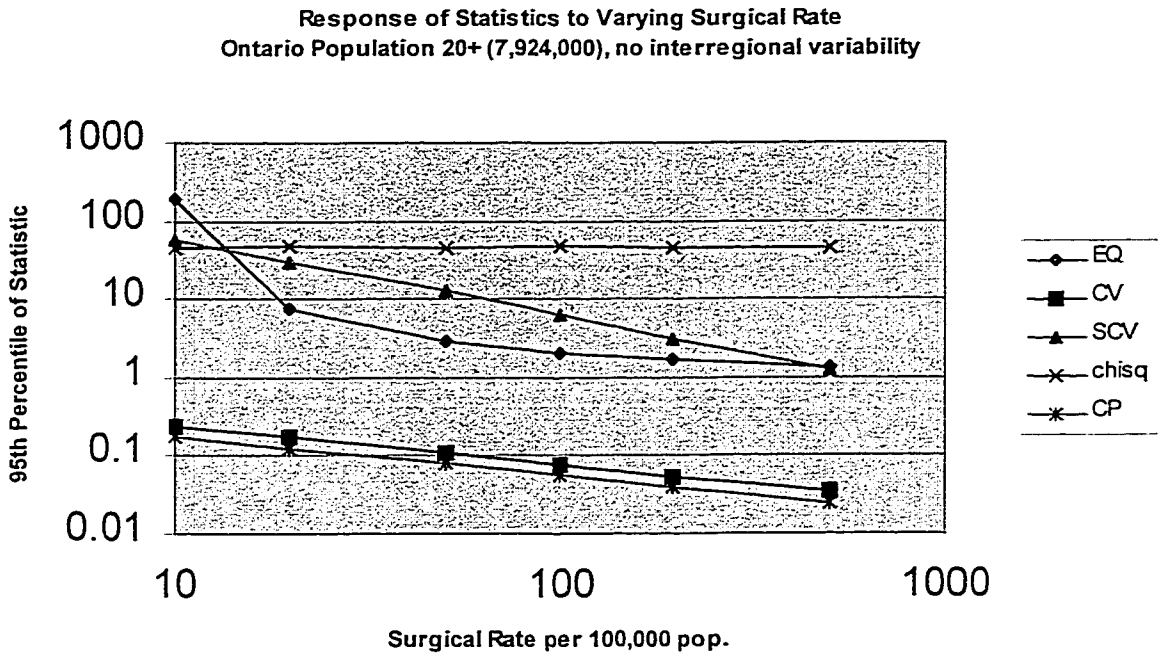


Figure 3 Response of Statistics to Varying Surgical Rate, Ontario

3.4 Varying Population Size

Figures 4 and 5 show the results of varying the population size, in this case by restricting the population of interest to females 20 and over (Figure 4) or to only males 50 and over (Figure 5). This is done because some operations (most urological and all gynecological procedures) apply only or largely to one sex or sex/age subpopulation. Another way of thinking of results in a smaller population is that they apply to the entire population of a smaller province. This allows some generalizability of the results. To extend this a population of 500,000 was also simulated (figure 6). This was divided into twenty equal regions.

The adult female population (figure 4) totals 3,861,438 people, divided into regions ranging from 7,471 to 959,667. The 95th percentiles of the statistics are graphed. These values were determined by 3000 iterations of simulation. The EQ again shows a strong sensitivity to underlying surgical rates, although with a smaller overall population the problem is worse as for the lower surgical rates the EQ is now infinite (ie lowest number of cases zero) more than five percent of the time. (The plotted point for the EQ is suppressed when the value is infinity more than 5% of the time. In these cases, the last point plotted may be numerically inaccurate for the EQ. This is marked on the graphs where appropriate. The error trapping routine substitutes a rate of 0.1 per 100,000 when the surgical rate is zero.). The SCV again shows a strong sensitivity to underlying surgical rates, and again as the rate becomes small the

threshold value of the SCV becomes very large (140) with no variability between regions, whereas for a surgical rate of 500 the threshold value of the SCV is 2.62. The CV is sensitive to underlying surgical rates, and now the CV is only expected to be 0.10 or less if the surgical rate is 200 per 100,000 or higher. Again the threshold value of the chi-square statistic is 46, confirming accuracy of the simulated values and adequacy of the number of trials.

In the male only population 50 and above (figure 5), the number of people is now 1,183,000 overall, in regions varying from 3,400 to 285,000 in population. The figure again shows 95th percentiles for each statistic, determined by 3000 iterations of simulation. With this smaller population (although not small in comparison to most provinces!) the instability of the EQ and SCV becomes more dramatic, and they are very unstable for surgical rates of 200 or below. The CV has a threshold of 0.10 only for a surgical rate of 500, and the threshold for the CV goes as high as 0.41. The chi-square statistic again has the appropriate and constant threshold value of 46.

The results for a population size of 500,000 (figure 6) show EQ and SCV having very large threshold values for smaller surgical rates. The CV approaches 1 with the lowest surgical rates, and is only below 0.1 for surgical rates larger than 200 per 100,000.

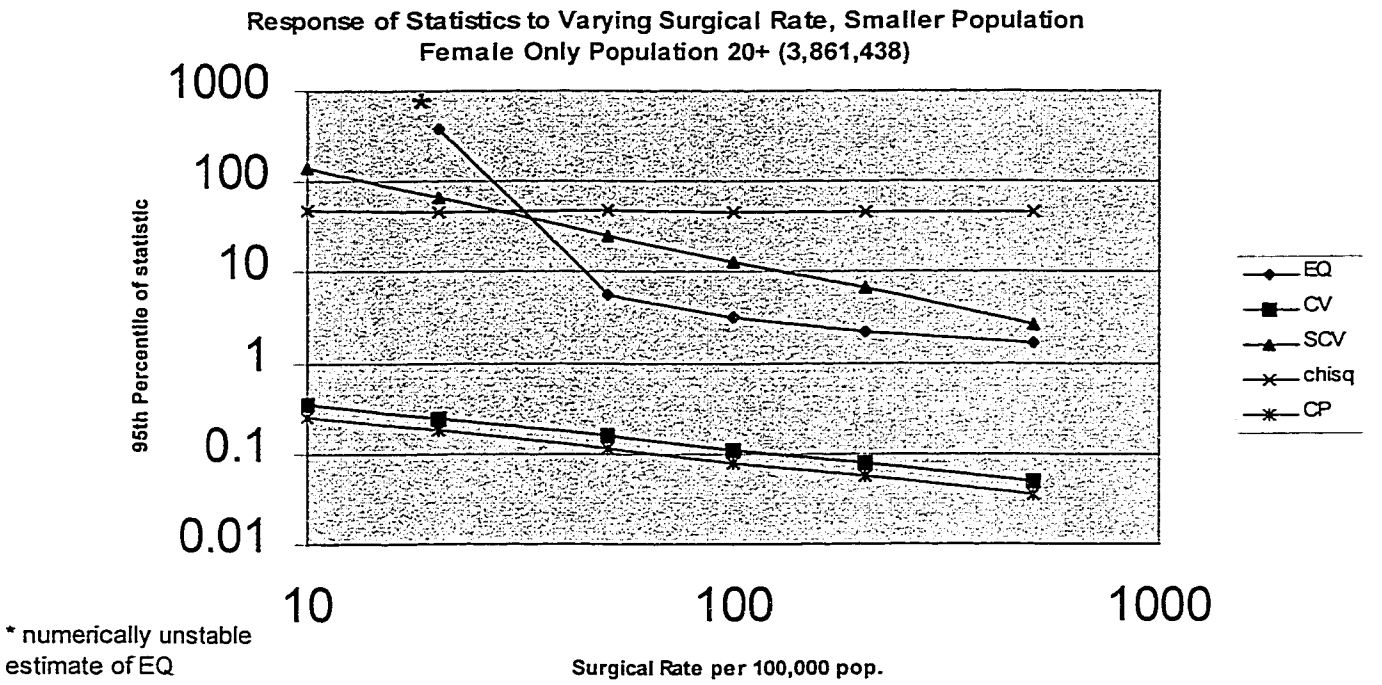


Figure 4 Response of Statistics to varying rate, female only population

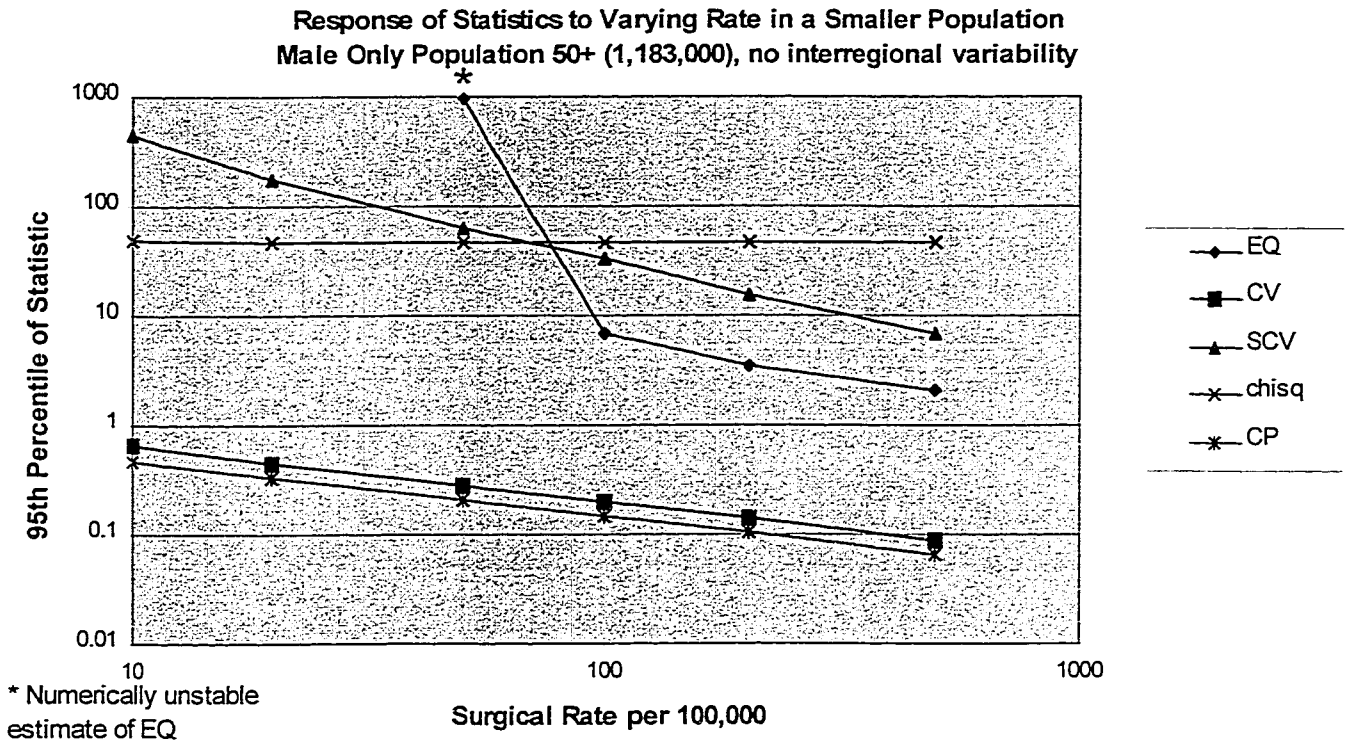


Figure 5 Response of Statistics to Varying Rates, Ontario Male Age 50+

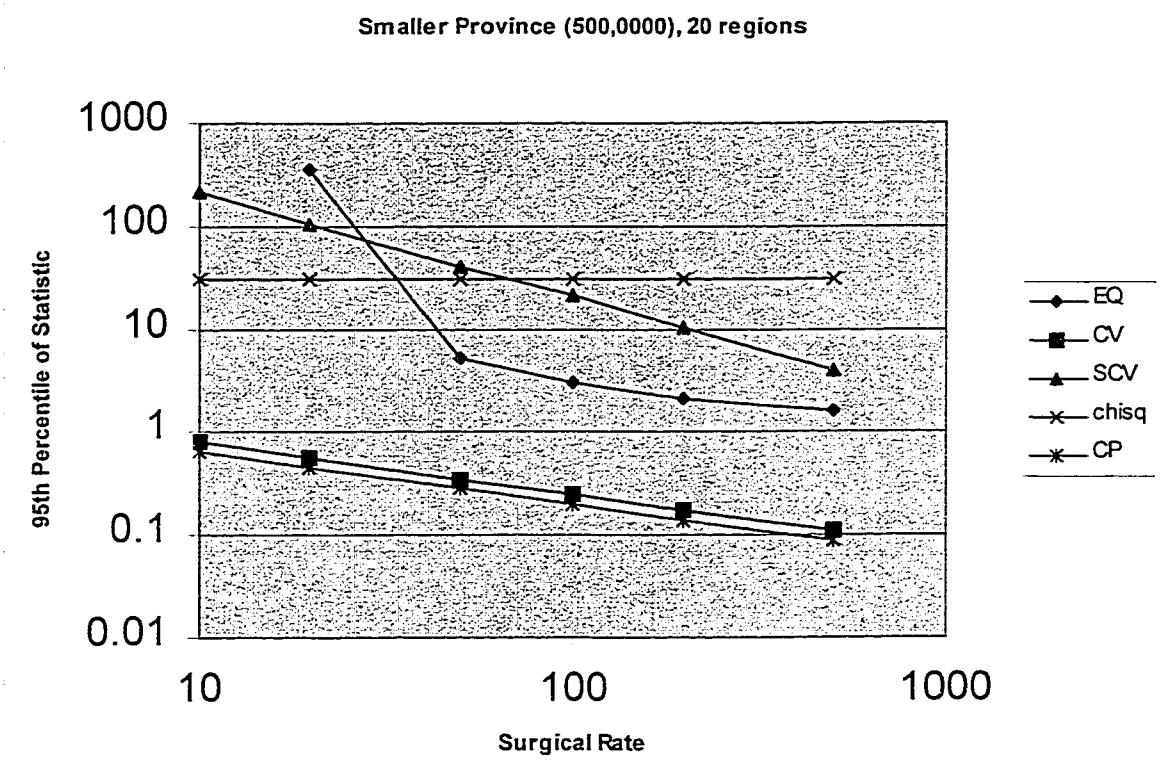


Figure 6 Response to Varying Surgical rates in a smaller province (500,000) with 20 regions and no interregional variability

3.5 *Varying Population Structure*

Two ways of varying the population structure were explored. The first was to keep 33 regions, but to divide the population more and more equally until all regions were equal in size with 1/33 of the province's population. The second way was to divide the province into smaller numbers of regions - six, eleven, or seventeen regions instead of thirty three.

The results of dividing into 33 regions of more and more equal size are shown in figures 7 and 8. The 'scaling factor' is a linear multiplier of the difference between each region's population and the mean region population. When the scaling factor is zero the populations are all equal, at 240,133 per region. When the scaling factor is one, the populations vary from 15944 to 1,851,913. The points plotted are the 95th percentiles of the statistics, determined by simulation. Surgical rates of 10 and of 50 per 100,000 are shown.

Both the EQ and the SCV are sensitive to whether the population is divided into equally or unequally sized regions. Both of these statistics are much more stable at the left end of the graph, when the population is equally or roughly equally divided. There is quite a change in the value of EQ and SCV when the population is divided into the unequally sized (but administratively sensible) regions defined by the district health councils. By contrast, the CV and the chi-square have exactly the same

threshold values whether the population is divided into equally sized regions, or very unequally sized ones.

The other way of changing the population structure is to divide the population into different numbers of regions. Figure 9 shows the results if six, eleven, or seventeen regions are used instead of thirty three. In this example, all of the regions are taken to be of equal size. It is known from the preceding that unequally sized regions would not change the behaviour of CV or chi-square, but would change the behaviour of EQ and SCV.

Dividing the population into different numbers of equal regions changes the 95th percentiles of all of the statistics. As expected, it is not possible to directly compare variability across different numbers of divisions of the same population. This analysis is done only to show that dividing the population into different numbers of regions changes the expected values of all statistics summarizing rate variability. At its extremes (e.g. going from one region to two regions or three regions) this result is intuitively obvious.

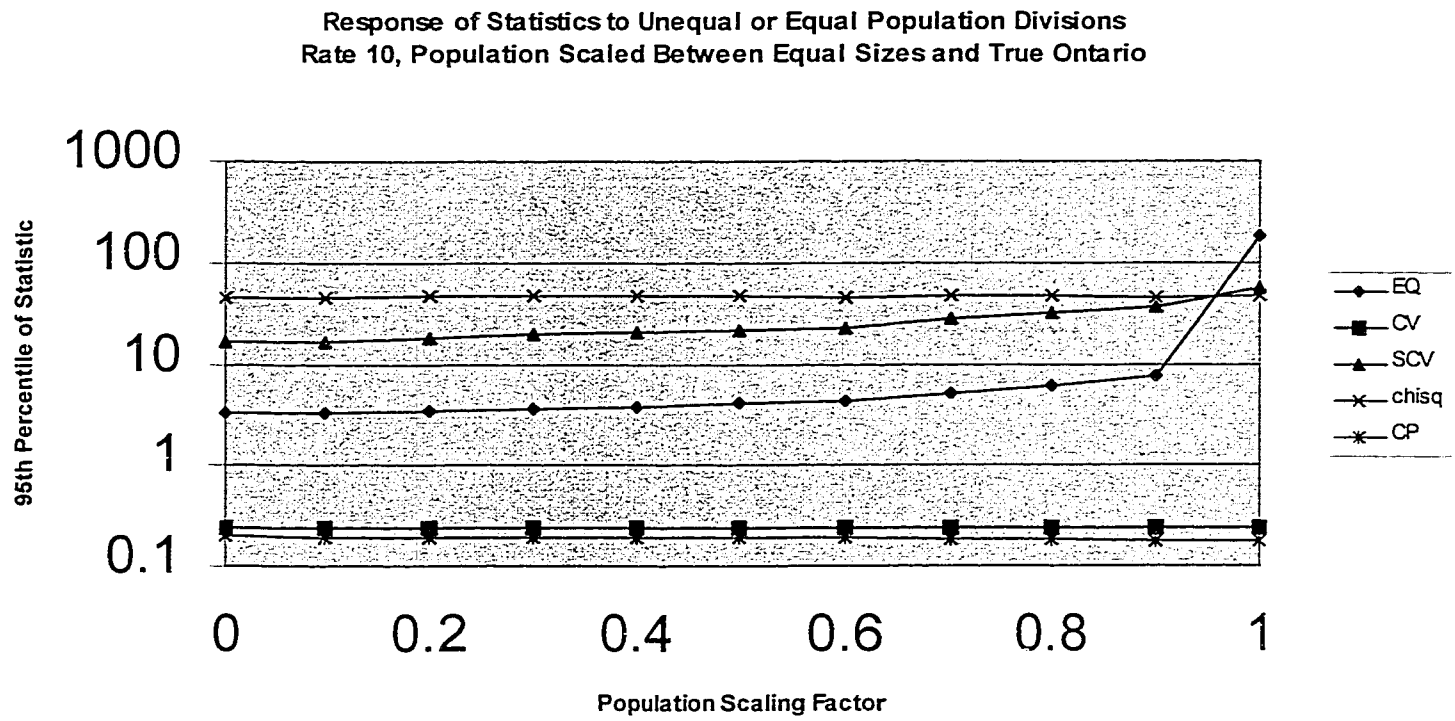


Figure 7 Response to Unequal Population Divisions, Rate 10

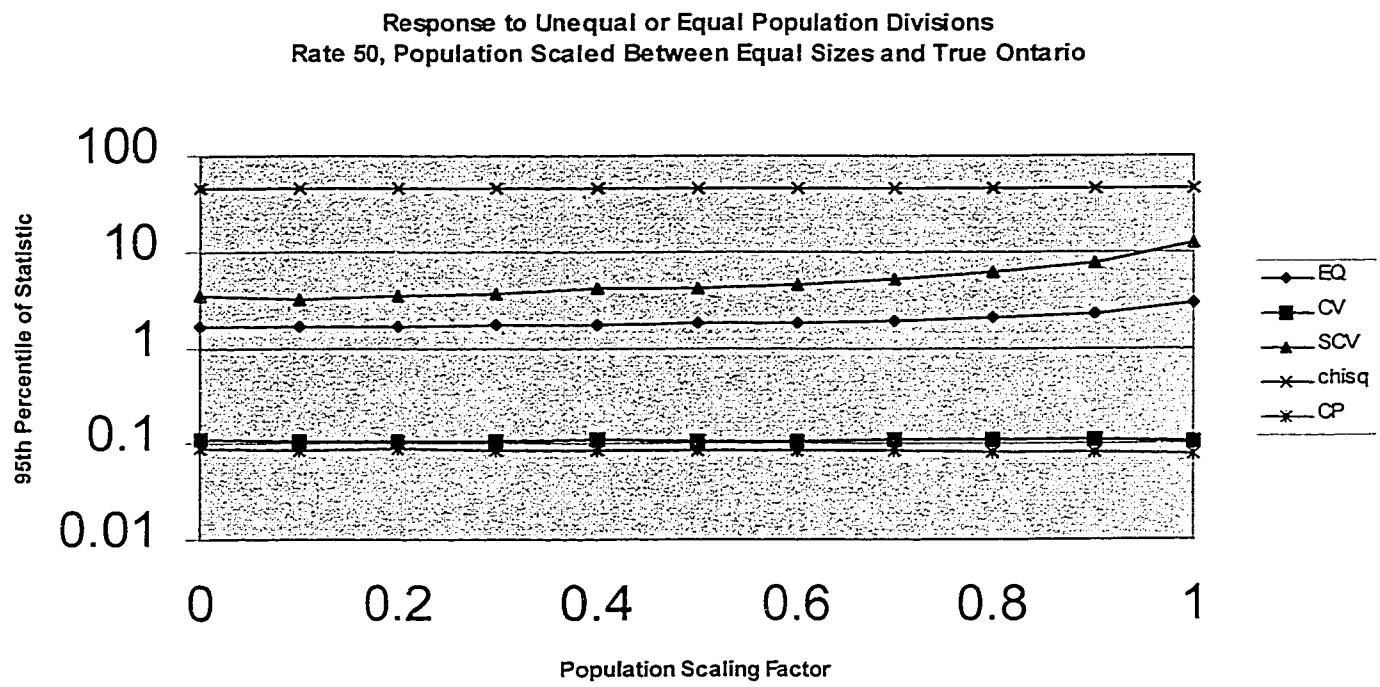


Figure 8 Response to Unequal Population Divisions, Rate 50

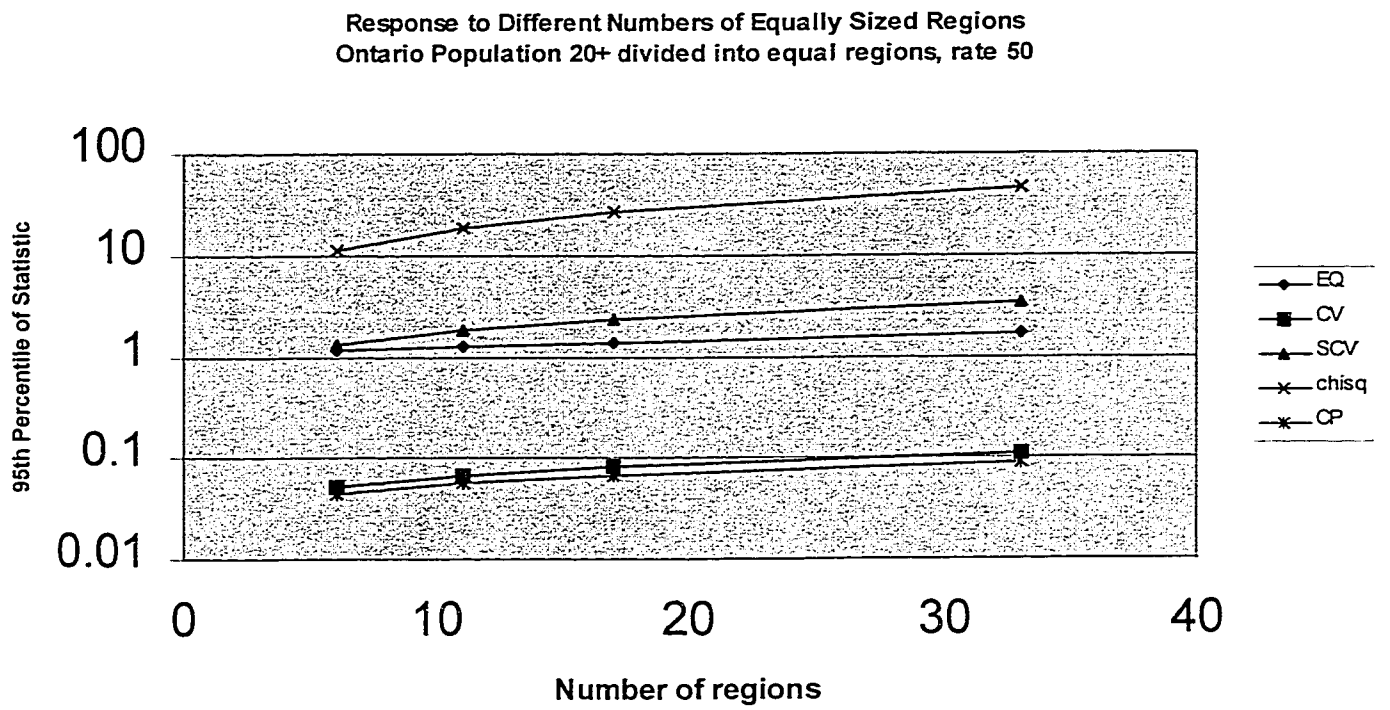


Figure 9 Response to Different Numbers of Regions, Rate 50

3.6 *Influence of Readmissions*

The readmission algorithm allows for zero, one or two admissions per year per patient for the same surgical procedure. The example provided assumes a 15% rate of readmission for those with a primary admission. This is meant to be a high estimate of the potential effect of readmissions on the rates of operations of interest, but is an inappropriate analysis for medical diagnoses which can be associated with a much higher readmission rate. Details of the reasons and assumptions are discussed in section 4.9.

To see how readmsisions change the expected distributions of the statistics, it is useful to compare readmission figures 10 and 11 to corresponding no readmission figures 3 and 5.

If readmissions are possible, the EQ and SCV are even more responsive to underlying surgical rate, and both quickly become large at even surgical rates in the middle of the range of interest. In fact, the 95^{th} percentile of the EQ becomes infinity at surgical rates below 50 if the entire population is used, and at rates below 200 if the male over 50 population is used.

The coefficient of variation is relatively insensitive to the introduction of readmissions in this fashion. The 95th percentile value increases only slightly at each surgical rate.

The 95th percentile for the chi-square changes if readmissions are possible - from 46 to approximately 62. This is somewhat different from the multiple admission factor adjustment suggested by Diehr et al. (39) which would suggest a critical value of approximately 52 for the chi-square.

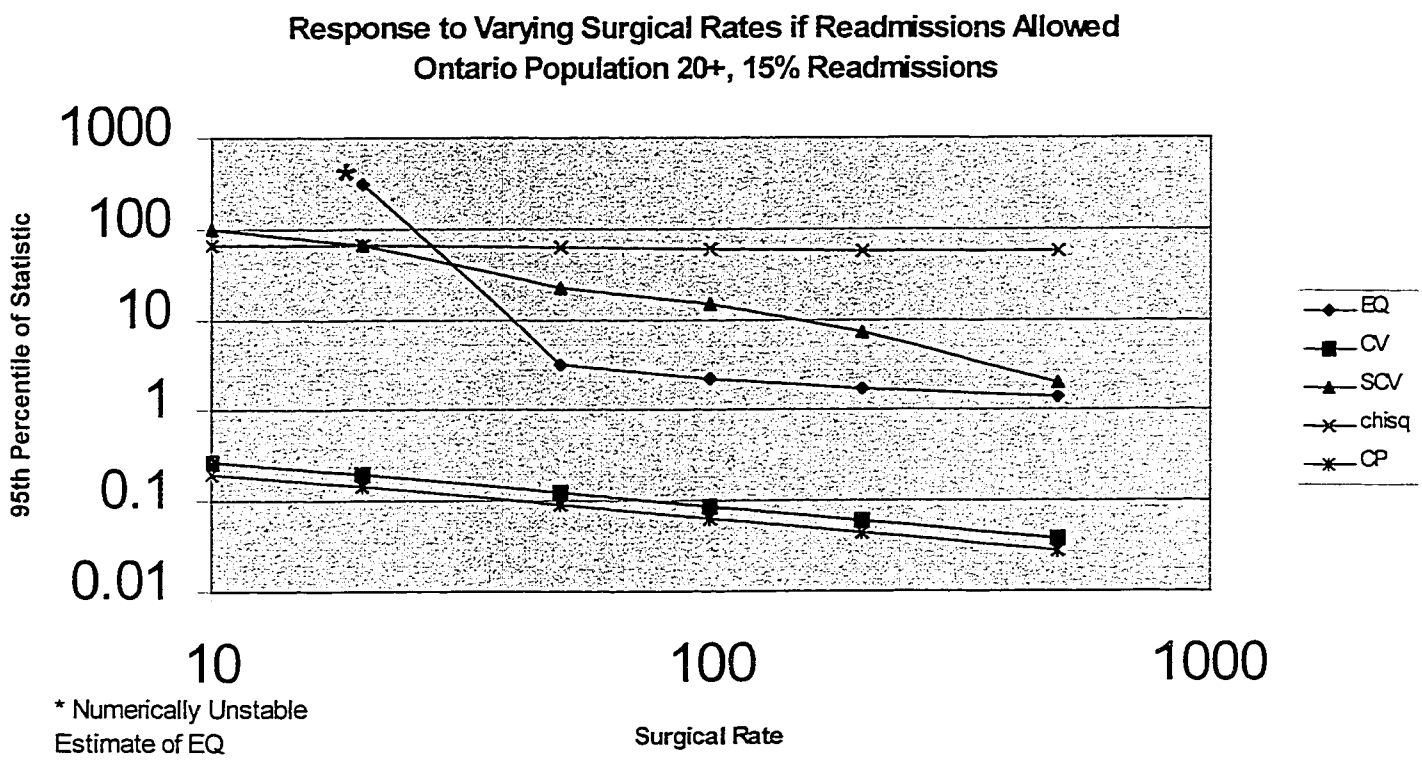


Figure 10 Response to Varying Surgical Rates with 15% Readmissions, Ontario

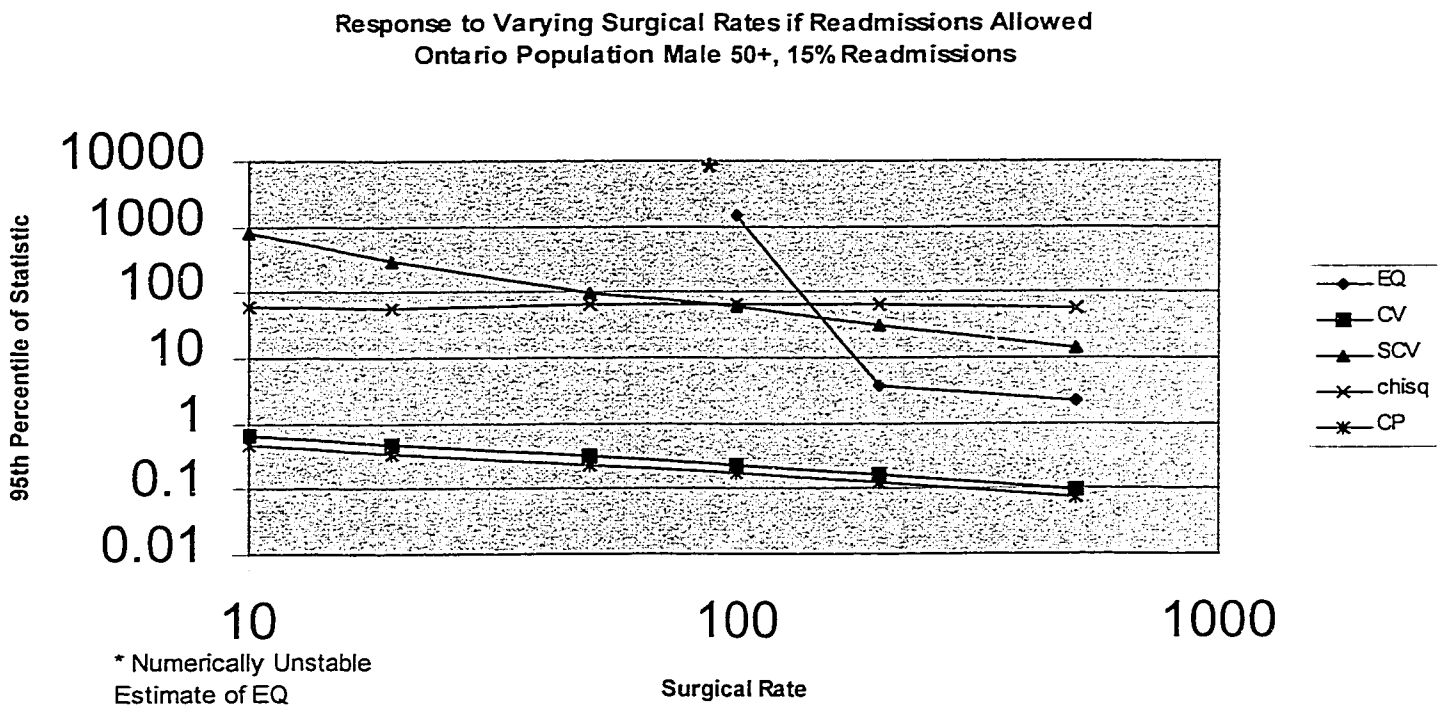


Figure 11 Response to Varying Surgical Rates with 15% Readmissions, Ontario Male 50+

3.7 The Case Count

Definition: The case count is the number of cases by which the observed distribution differs from that expected if true rates are equal to the overall mean in each region. It is the sum of the absolute differences between the observed and expected numbers of cases in each region. It can be expressed either in an absolute form, i.e. n cases, or as a proportion n/t of the t total cases observed. It is easy to convert one to the other, and in the following the relative version is used. I use the abbreviation CP to signify case proportion.

Figures 3 through 11 correspond to the scenarios used to test the response of the other statistics to changes in surgical rates, population size, population structure, and readmission rate. They include the CP as well as the other statistics.

The CP closely parallels the CV in distribution under the situations tested, and has a somewhat smaller numerical value. Like the CV, the CP is relatively less sensitive to variations in the surgical rate than are the EQ and SCV (figures 3 to 6). The CP remains constant whether the population is divided very equally or very unequally (figures 7 and 8). As with all the other statistics, the CP is expected to have lower values if smaller numbers of regions are used (figure 9). A readmission rate of 15% does not make an important difference to the distribution of the CP, even with the smallest populations and smallest surgical rates tested (figures 10 and 11).

3.8 Testing Power

Three power testing scenarios were used. The first assesses the power of the various statistics to detect realistic variability based on actual data, in this case the Ontario coronary bypass surgery rates. The second assesses the power of the statistics to detect a single outlying rate in a region with a large population, and the third assesses the power of the statistics to detect a single outlying rate in a region with a small population.

The power curve for the Ontario coronary bypass surgery rates (figure 12) uses a scaling factor which is again a linear multiplier applied to the absolute difference between the equal rate and the observed rate in each region. When the scaling factor is zero the true rates are equal in all regions, and when the scaling factor is one the true rates in each region equal the observed rate.

As expected, all of the statistics have a power of .05 when the true rates are equal, since using a 95^{th} percentile for rejection of the null hypothesis will result in rejection of null hypothesis 5% of the time when it is true. All of the statistics have a sigmoid shape to the power curve. The chi-square, CV, and CP all have high power at detecting the variability observed in the real data, such that they all have power of 80% or better when the variability is scaled to only 0.2 to 0.3 of that observed in Ontario. The SCV has less power to detect the observed difference in rates, achieving 80% power only when the rate is scaled to 0.4 or greater. The EQ has the worst power of all, achieving a power of 80% only when the rate is scaled to 0.7 or greater.

The statistics differ markedly in their ability to detect outliers in large and small regions. Chi-square, CV, and CP all have power of 80% or better to detect a low outlier in a large region if the rate is 0.8 of the mean rate or lower (figure 13) . SCV has 80% power once the rate drops to 0.6 of the mean rate, and EQ cannot detect the outlier with 80% power until the rate drops to 0.4 of the mean rate. High outliers are also better detected by Chi-square, CV, and CP than they are by SCV or EQ when they occur in large regions (figure 14).

If the region with the outliers is small the situation is quite different. A low outlier in a small region is eventually detected by EQ and SCV, but only when the rate is 0.3 times the mean rate or lower. Chi-square, CV, and CP do not detect a low outlier in a small region at all right down to a rate of 0.1 times the mean rate. EQ, CV, SCV, and chi-square all detect a high outlier in a small region with 80% power or better if the rate is 3 times the mean rate or more. CP never detects a high outlier in a small region even when the rate is 10 times the mean rate. All of these results relate to the fact that in small regions, a small difference in case numbers makes a large difference to the observed rate (figures 15 and 16).

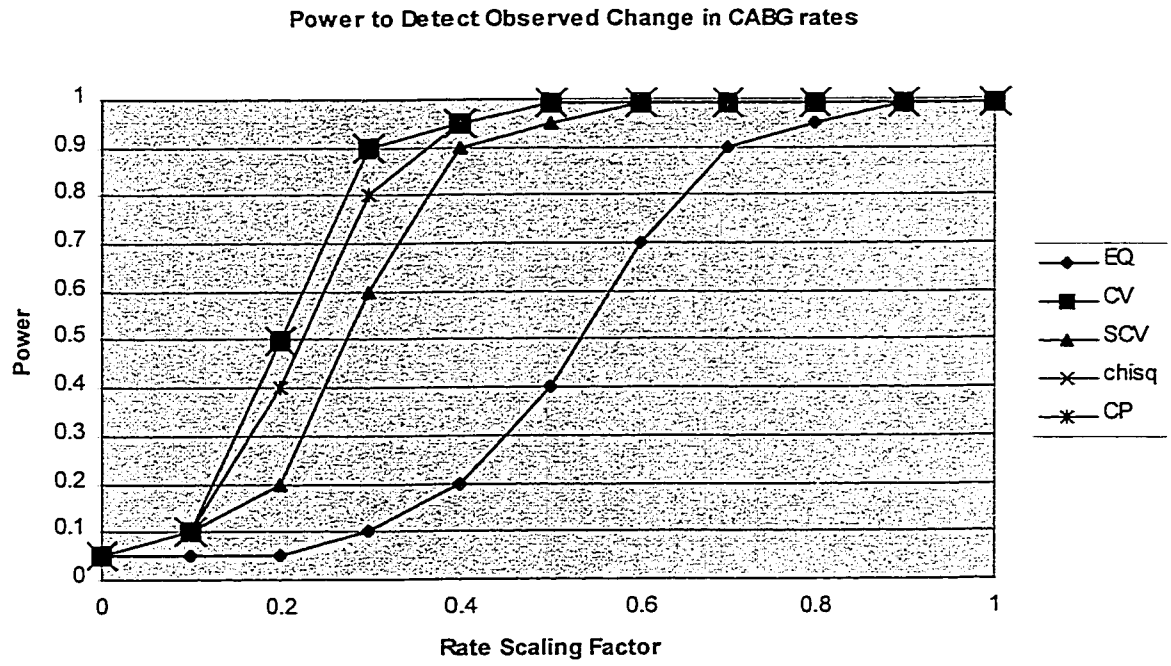


Figure 12 Power to detect Observed Change in CABG Rates

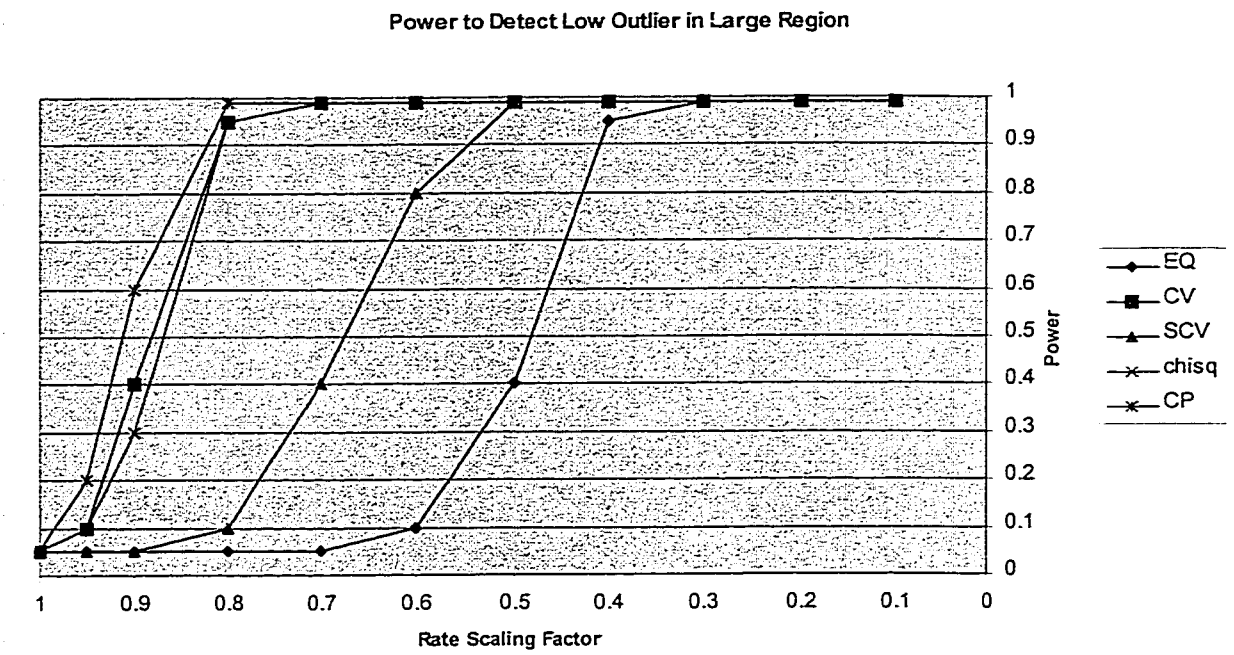


Figure 13 Power to detect Low Outlier in Large Region

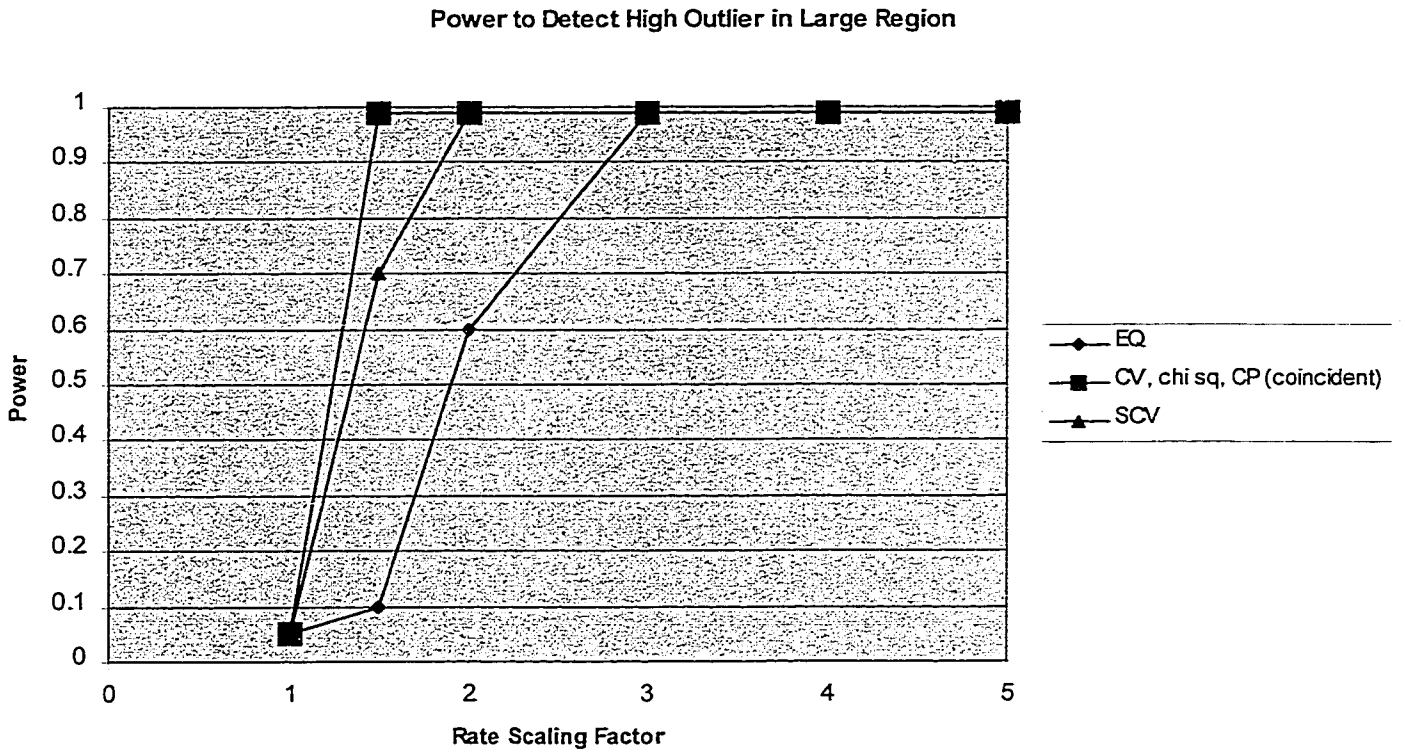


Figure 14 Power to detect High Outlier in Large Region

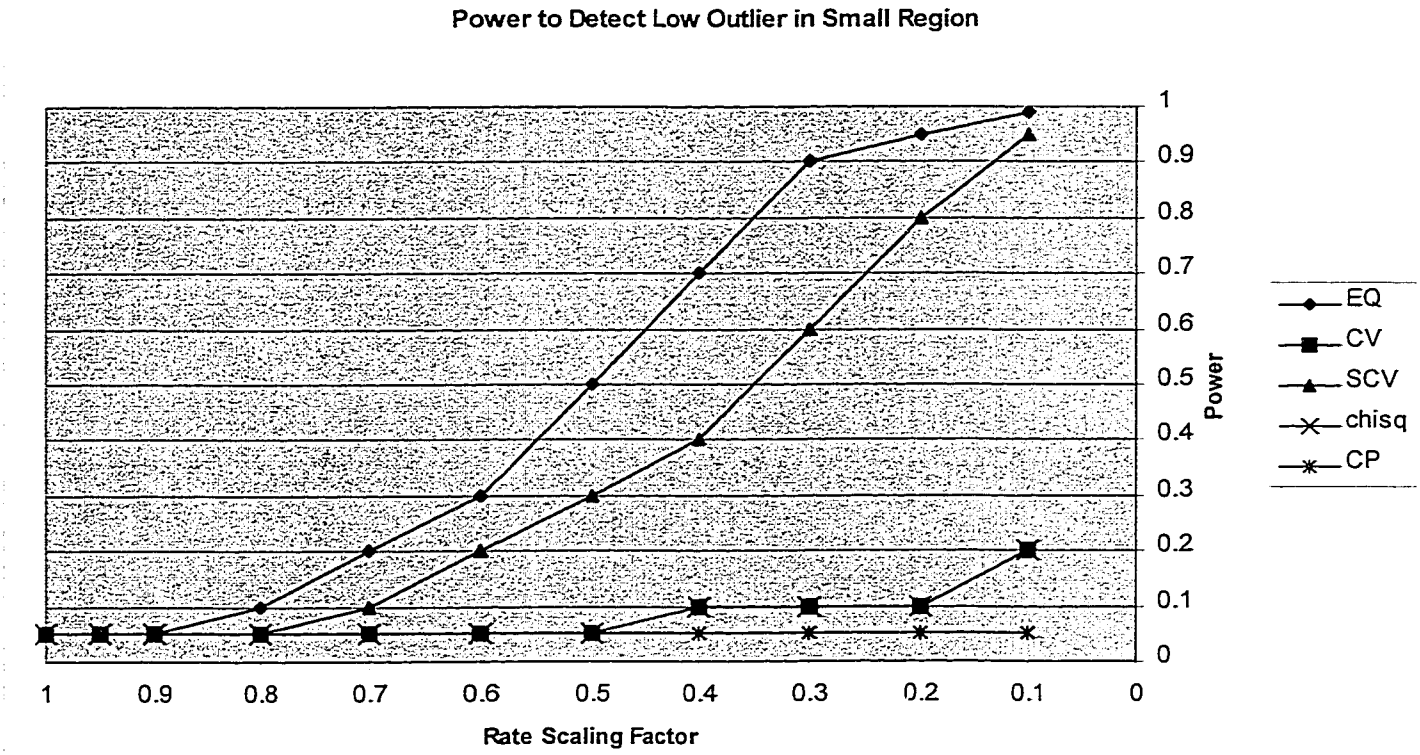


Figure 15 Power to detect Low Outlier in Small Region

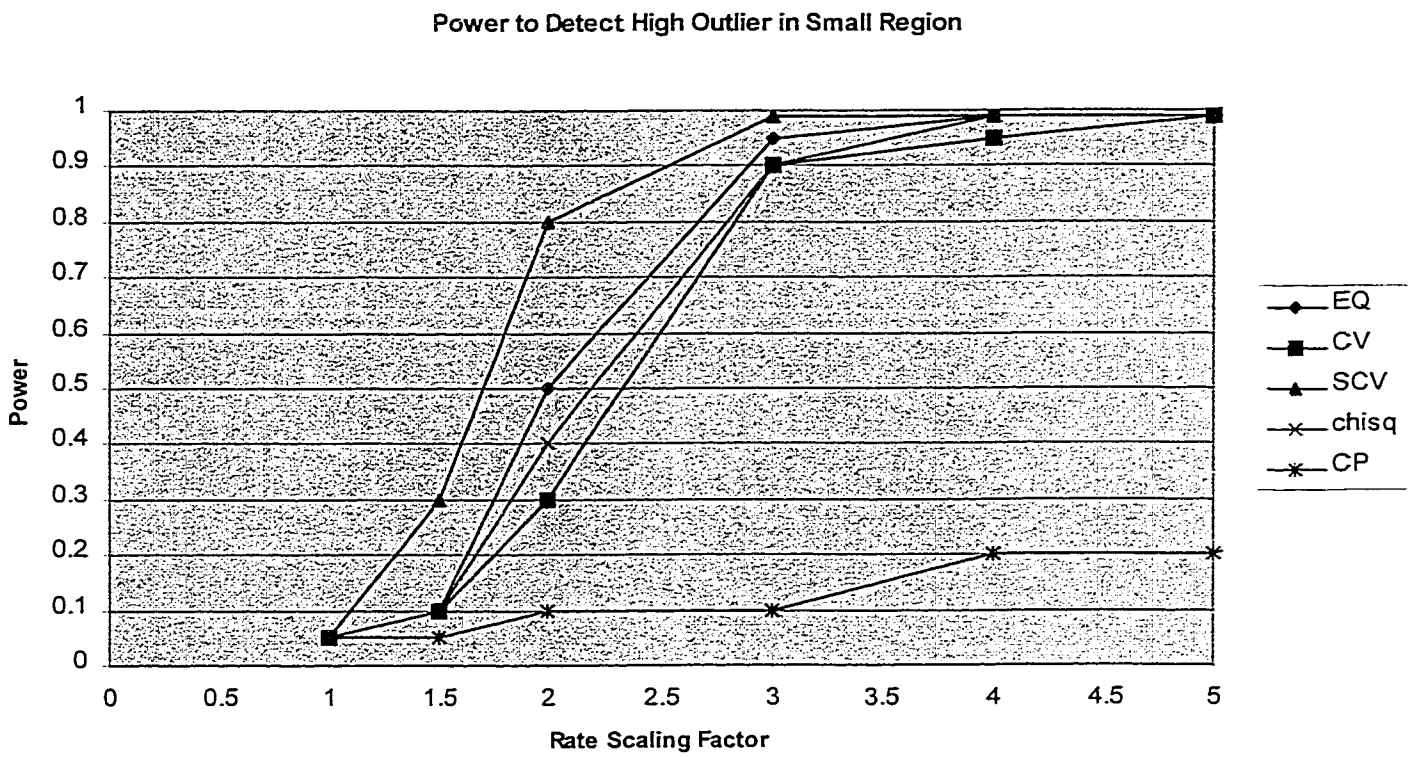


Figure 16 Power to detect High Outlier in Small Region

3.9 Analyzing Real Data

Table 5 presents the statistics calculated from observed rates for six different types of surgery, as obtained from the electronic version of the ICES practice atlas for 1989 to 1992. Both crude and age and sex adjusted rates are presented. Statistics calculated from observed data are italicized. Following the observed data, the results of simulations based on an equal rate of surgery in each region (equal to the provincial mean rate for that operation) are presented. The mean and 95th percentile of the simulated distribution are presented. Comparison of values for the age-sex adjusted rates to the 95th percentile (bolded for CV and CP, for ease) provides a significance test at the .05 level. The percentile of the (simulated) expected distribution that the age/sex adjusted observed value corresponds to is also given. The pseudo p-value is one minus this, and is expressed as a probability (out of one) rather than a percentile (out of 100).

Comparing Observed Values to those Expected Based on Simulations

1989-92 ICES Data. Italics indicate observations, nonitalics are simulated distributions.

CV and IDR are bolded and used to test whether observed rates exceed 95%ile expected from simulation.

	eq	cv	Scv	chisq	cp	significant variability? random
Appendectomy:						
appy observed rates – crude	*2.033*	*0.161*	*40.661*	*240.988*	*0.121*	
appy observed rates -age sex adjusted	*2.029*	***0.162***	*41.968*	*244.903*	***0.127***	
Simulated mean	1.561	0.058	-0.012	32.241	0.042	
Simulated 95% ci	2.033	**0.071**	5.814	46.708	**0.052**	yes
Percentile observed vs sim	0.950	1.000	1.000	1.000	1.000	
Pseudo p-value	0.050	0.000	0.000	0.000	0.000	
Abdominal Aortic Aneurysm Repair:						
aaa crude	*4.318*	*0.363*	*138.168*	*224.449*	*0.291*	
aaa age and sex adustedj	*2.529*	***0.244***	*20.751*	*100.760*	***0.212***	
simulated aaa mean(adj)	14.820	0.136	0.169	32.247	0.096	
simulated aaa 95%ile (adj)	5.706	**0.164**	28.849	46.299	**0.121**	yes
percentile aaa actual (adj) vs sim	0.394	1.000	0.897	1.000	1.000	
pseudo p-value	0.606	0.000	0.103	0.000	0.000	
Coronary Artery Bypass Grafting:						
cabg crude	*3.983*	*0.306*	*191.467*	*485.269*	*0.205*	
cabg age and sex adjusted	*3.592*	***0.271***	*131.470*	*381.820*	***0.178***	
simulated CABG mean (ADJ)	1.829	0.078	-0.268	32.187	0.055	
simulated CABG 95%ile (ADJ)	2.509	**0.094**	9.695	46.318	**0.070**	yes
percentile CABG actual (adj) vs sim	0.992	1.000	1.000	1.000	1.000	
pseudo p-value	0.008	0.000	0.000	0.000	0.000	
Total Hip Replacement:						
thr crude	*2.771*	*0.241*	*83.455*	*328.948*	*0.185*	
thr age and sex adjusted	*1.672*	***0.134***	*15.266*	*102.036*	***0.103***	
simulated thr mean (adj)	1.778	0.074	-0.163	32.014	0.052	
simulated thr 95%ile (adj)	2.392	**0.089**	9.021	45.594	**0.066**	yes
percentile thr actual (adj) vs sim	0.486	1.000	0.987	1.000	1.000	
pseudo p-value	0.514	0.000	0.013	0.000	0.000	

Comparing Observed Values to those Expected Based on Simulations : continuation of table						
Dilatation and Curettage:						
d+c crude	*2.535*	*0.203*	*58.542*	*859.633*	*0.174*	
d+c age and sex adjusted	*2.603*	***0.207***	*61.937*	*895.772*	***0.181***	
simulated d+c mean (adj)	1.340	0.039	-0.041	31.934	0.027	
simulated d+c 95%ile (adj)	1.578	**0.047**	2.570	46.339	**0.034**	yes
Percentile d+c actual (adj) vs sim	1.000	1.000	1.000	1.000	1.000	
pseudo p-value	0.000	0.000	0.000	0.000	0.000	
Transurethral Resection Prostate:						
turp crude	*2.067*	*0.167*	*34.551*	*457.617*	*0.145*	
Turp age and sex adjusted	*2.180*	***0.152***	*32.968*	*382.811*	***0.123***	
simulated turp mean (adj)	1.347	0.044	0.170	33.313	0.031	
simulated turp 95%ile (adj)	1.547	**0.053**	2.400	45.957	**0.039**	yes
percentile turp actual (adj) vs sim	0.999	0.999	0.999	0.999	1.000	
pseudo p-value	0.001	0.001	0.001	0.001	0.000	

Table 5 Observed versus expected values for all statistics and operations.

For example, the observed data for hip replacement show an extremal quotient (EQ) of 1.672 and a coefficient of variation (CV) of 0.134. The simulated data, based on equal rates in all regions, show an expected EQ of 1.778 and an expected CV of 0.074. The 95th percentile of the expected distribution of the EQ is 2.39, and of the CV is 0.089. The observed CV is greater than the 95th percentile of the expected distribution based on equal rates, so the hypothesis of equal true rates in each region can be rejected. The observed EQ of 1.67, however, is less than the 95th percentile of 2.39 which is expected under equal true rates. In fact, it is less than the 50th percentile (mean) of the expected distribution, which is 1.77. The overall range in observed rates is not greater than would be expected if true rates were equal. The EQ lacks the power to reject the null hypothesis of equal true rates.

In most cases, the statistics calculated from the observed data far exceed those from the simulated distributions, meaning that the observed results are highly statistically significant when compared with the null hypothesis. The observed value is not statistically significant in the case of the EQ for abdominal aortic aneurysm surgery, appendectomy surgery, and total hip replacement; and in the case of the SCV for abdominal aortic aneurysm surgery. This is because of the low power of the EQ and SCV to detect the type of rate variability which is actually present. This result is consistent with that found in the power section above.

3.10 Appendectomy test

The appendectomy test constitutes a test of the modified null hypothesis H_o(appy).

> H_o(appy): The intraregional rate variability observed in operation X is no greater than the intraregional variability in appendectomy rates.

Testing this hypothesis involves comparing the CV or CP from the observed rates to a different set of simulation results. The simulation is performed with true rates in each region varying according to the amount that appendectomy rates vary. Table 6 contains the results of these simulations. Coronary artery bypass grafting and dilation and curettage still have significant enough observed variability to reject H_o(appy), but total hip replacement and transurethral resection of the prostate do not. Abdominal aortic aneurysm repair has marginally significant results. The table is presented with the values of CV and CP bolded, as these are used as the test statistics. In all of these cases the results are identical using any of EQ,CV,SCV, chi-square, or CP as a test statistic.

Comparing Observed Data to Expected Values from Simulations Based on Appendectomy Rates or SMR's.

1989-92 ICES Data. Italics indicate observations, nonitalics are simulated distributions.

The simulations are based on three scenarios: first, equal 'true rates' for all regions, all variability from sampling second, true rates vary in magnitude in the way true appendectomy rate do, and third, true rates vary in magnitude the way SMR's do.
CV and IDR are bolded and used to test whether observed rates exceed 95%ile expected from simulation.

						significant variability?		
	eq	**cv**	**scv**	**chisq**	**cp**	random	Appy	smr
Appendectomy:								
appy observed rates - crude	*2.033*	***0.161***	*40.661*	*240.988*	***0.121***			
appy observed rates -age sex adjusted	*2.029*	*0.162*	*41.968*	*244.903*	*0.127*			
simulated mean	1.561	0.058	-0.012	32.241	0.042			
simulated 95% ci	2.033	**0.071**	5.814	46.708	**0.052**	yes		
percentile observed vs sim	0.950	1.000	1.000	1.000	1.000			
true' rates scaled to .92	2.131	0.161	43.500	242.936	0.123			
true' rates scaled to .93	2.137	0.163	44.711	248.666	0.125			
Abdominal Aortic Aneurysm Repair:								
aaa crude	*4.318*	*0.363*	*138.168*	*224.449*	*0.291*			
aaa adjusted	*2.529*	***0.244***	*20.751*	*100.760*	***0.212***			
simulated aaa mean(adj)	14.820	0.136	0.169	32.247	0.096			
simulated aaa 95%ile (adj)	5.706	**0.164**	28.849	46.299	**0.121**	yes		
percentile aaa actual (adj) vs sim	0.394	1.000	0.897	1.000	1.000			
aaa simulated .93 appy adj mean	10.441	0.206	50.339	74.376	0.151			
aaa simulated .93 appy adj 95%ile	6.396	**0.247**	103.924	105.033	**0.185**			marginal
aaa simulated smr adj mean	58.302	0.259	34.489	128.616	0.219			
aaa simulated smr adj 95%ile	400.384	**0.290**	63.802	160.951	**0.248**			no
Coronary Artery Bypass Grafting:								
cabg crude	*3.983*	*0.306*	*191.467*	*485.269*	*0.205*			
cabg adj	*3.592*	***0.271***	*131.470*	*381.820*	***0.178***			
simulated CABG mean (ADJ)	1.829	0.078	-0.268	32.187	0.055			
simulated CABG 95%ile (ADJ)	2.509	**0.094**	9.695	46.318	**0.070**	yes		
percentile CABG actual (adj) vs sim	0.992	1.000	1.000	1.000	1.000			
cabg sim .93 appy adj mean	2.329	0.171	46.270	152.370	0.129			
cabg sim .93 appy adj 95%ile	2.971	**0.196**	70.117	197.531	**0.148**		Yes	
cabg sim smr adj mean	5.039	0.237	36.240	322.883	0.200			
cabg sim simr adj 95%ile	7.560	**0.256**	50.320	375.693	**0.218**			yes

Comparing Observed Data to Expected Values from Simulations Based on Appendectomy Rates or SMR's. continuation of table

Total Hip Replacement:								
thr crude	*2.771*	*0.241*	*83.455*	*328.948*	*0.185*			
thr adj	*1.672*	***0.134***	*15.266*	*102.036*	***0.103***			
simulated thr mean (adj)	1.778	0.074	-0.163	32.014	0.052			
simulated thr 95%ile (adj)	2.392	**0.089**	9.021	45.594	**0.066**	yes		
percentile thr actual (adj) vs sim	0.486	1.000	0.987	1.000	1.000			
thr sim .93 appy adj mean	2.279	0.169	45.955	164.399	0.128			
thr sim .93 appy adj 95%ile	2.880	**0.191**	68.064	210.469	**0.146**		No	
thr sim smr adj mean	5.414	0.236	36.254	354.285	0.198			
thr sim smr adj 95%ile	5.996	**0.253**	50.758	406.684	**0.215**			no
Dilatation and Curettage:								
d+c crude	*2.535*	*0.203*	*58.542*	*859.633*	*0.174*			
d+c adj	*2.603*	***0.207***	*61.937*	*895.772*	***0.181***			
simulated d+c mean (adj)	1.340	0.039	-0.041	31.934	0.027			
simulated d+c 95%ile (adj)	1.578	**0.047**	2.570	46.339	**0.034**	yes		
Percentile d+c actual (adj) vs sim	1.000	1.000	1.000	1.000	1.000			
d+c sim .93 appy adj mean	2.016	0.155	44.601	500.177	0.119			
d+c sim .93 appy adj 95%ile	2.270	**0.168**	54.837	585.852	**0.130**		Yes	
d+c sim smr adj mean	2.815	0.227	36.916	1190.074	0.189			
d+c sim smr adj 95%ile	3.566	**0.236**	43.598	1288.076	**0.198**			no
Transurethral Resection Prostate:								
turp crude	*2.067*	*0.167*	*34.551*	*457.617*	*0.145*			
turp adj	*2.180*	***0.152***	*32.968*	*382.811*	***0.123***			
simulated turp mean (adj)	1.347	0.044	0.170	33.313	0.031			
simulated turp 95%ile (adj)	1.547	**0.053**	2.400	45.957	**0.039**	yes		
percentile turp actual (adj) vs sim	0.999	0.999	0.999	0.999	1.000			
turp sim .93 appy adj mean	2.026	0.158	42.672	416.952	0.123			
turp sim .93 appy adj 95%ile	2.297	**0.172**	53.428	493.068	**0.135**		No	
turp sim smr adj mean	2.787	0.216	36.279	872.868	0.178			
turp sim smr adj 95%ile	3.416	**0.227**	42.907	957.904	**0.189**			no

Table 6 Observed versus expected values, expectations based on appendectomy or SMR data.

3.11 SMR Test

The SMR test is for rejection of the modified null hypothesis H_o(SMR).

> H_o(SMR): The intraregional rate variability observed in operation X is no greater than the intraregional variability in standardized mortality ratios.

Table 6 also contains the results of the SMR test as well as the appendectomy test. Only Coronary Artery Bypass Grafting has sufficient variability to allow rejection of the SMR modified null hypothesis. Commentary upon these tests, including rationale, assumptions, and implications, will be deferred to the discussion section.

4 Discussion

4.1 Signal and Noise Problem

The purpose of small area variation studies, presumably, is to identify rate variations which merit a response. However, the list of potential reasons for variation is extensive, and not all of these reasons would merit a response. Suppose we are interested in whether Ontario's residents have equal access to a certain type of operation. Variations due to true differences in population morbidity, for instance, or to differences in health values between urban and rural populations, do not imply unequal access to resources. True interregional variation in rates caused by unequal access to resources is the 'signal' we want to pick up, but differences in population morbidity or other characteristics add noise to the signal.

This noise is a form of confounding. If the populations differ both in DHC of residence and in morbidity, we will not be able to separate the effects of one from the effects of the other. Furthermore, a statistically significant result may be due entirely to the effects of the confounder. This would result in a false positive test for the main effect.

In most research designs, adding noise to a signal or a system makes the signal more difficult to detect, and this is a reason for controlled experiments. In the standard method for analyzing small area variations, both noise and signal have similar effects

on the statistics. The only commonly used test statistic is a standard chi-square done in a very large population, and this test has high power to pick up small true interregional differences. The statistic cannot help us decide whether the observed 'statistically significant' result is due to the signal we are interested in, or to noise from some real, but unsought, cause of variability.

In most small area analyses, an attempt is made to control for confounding by age and sex adjustment of the rates. This step is taken largely because of the strong association between age and sex and disease incidence at the individual level. Populations, however, have markedly different levels of morbidity and disease incidence after age and sex adjustment(43,44). This is further discussed in section 4.4.

Two further approaches can be taken to deal with the almost certain presence of confounding in small area observational studies.

The traditional multivariate approach attempts to measure all potential confounders and include them in a mathematical model developed to explain the variation(28). The difficulty with this approach is that many of these population characteristics are difficult to define and expensive to measure, or unknown.

An alternative to this traditional approach is proposed in this thesis. The idea is to identify a control procedure which is believed on theoretical grounds to be free or largely free of the main effect sought. It is only likely that the main effect is significantly present if the overall variability in the procedure of interest exceeds that

observed in the control procedure. The advantage of this approach is that it does not require extensive collection of data on all potentially confounding independent variables. It provides a quick, empirical way to assess whether the main effect of interest is likely to be present or not.

4.2 What is the Process Modeled?

The process of surgical care needs to be considered because it gives a framework for understanding potential sources of rate variability. Patients develop symptoms of a condition for which surgical treatment is an option. They may or may not consult a general practitioner. A physician must diagnose the condition, and then may either treat the patient or refer to a surgeon. If a referral occurs, the patients will wait a variable amount of time to see a surgeon. Waiting may result in symptom resolution, adaptation to the symptoms, treatment with another modality, or worsening of the symptoms. At consultation the surgeon may recommend operative or nonoperative management. Patients will decide upon management, taking into account expected outcomes and probabilities, risk tolerance, degree of symptoms, valuation of current versus future health, and the recommendations of the surgeon. If operative management is chosen, the patient and the surgeon must wait for the required tests, operating room time, and required resources (e.g. implants, postoperative ICU time, transplant organs or tissues) to become available. Once this has happened, the operation will occur. It is at this point that the patient's operation is counted in the numerator of a small area statistic.

Given the complexity of the process, there are many points at which variations in rate could arise. Some of these sources of variation fall correctly into the domain of health system problems which policy may be able to correct - unequal access to resources across the province, for example. Some of these sources of variation suggest need for better information on outcomes. Some, however, simply reflect the complexity of the process but do not demand any effort to fix them.

Not all procedures have as complex a process leading to their occurrence. Consider the special case of appendectomy surgery. The symptoms of appendicitis are severe and bring patients to medical attention quite quickly. The diagnosis of appendicitis is noncontroversial. Urgent operative management is the standard treatment in the developed world(47,49,50). Published experience with nonoperative management was until recently confined to special military circumstances such as ships at sea, although a recent randomized trial of nonoperative management in 40 civilian patients has been published. In this trial, at least 40 percent of those assigned to nonoperative management ultimately underwent appendectomy(51). There are many fewer sources of rate variability for this procedure than in the generic circumstance above. In fact, in Canada the only source of variability in appendectomy rates should be the incidence rate of appendicitis, since nobody in the population should have such poor access to medical care that appendicitis goes undiagnosed or untreated. Furthermore, uncertainty regarding indications for surgery is minimal(50), and routine pathological confirmation of the condition is performed following surgery. Positive primary appendectomy rates, then, might be used as a type of 'control group' to assess how

much interregional variability in procedure rates is seen in the absence of health system factors requiring correction. Other similar operations for which the operation rate should equal the incidence rate are hip fracture, and perhaps normal obstetrical care. Myocardial infarction (carefully defined) may be a medical diagnosis where admission rate is determined by incidence rate. Good data regarding myocardial infarction would, if available, make a suitable control procedure for CABG rate. Study of procedures of this type offers the opportunity to empirically assess the 'baseline' variability which should exist in the absence of important clinical or administrative discrepancies between regions.

By contrast, procedures such as prostatectomy, hip replacement, and heart bypass surgery are all elective, and all have medical alternatives to treatment (although absolute surgical indications such as relief of acute urinary retention, treatment of pathological fracture, or symptomatic left main coronary disease, may exist in a minority of cases performed). Elective procedures like this account for a significant portion of surgical spending within Ontario(24). These are the operations for which rate variability between regions may reflect either differences in clinical practice, or differences in availability of health services - both of which are important to determine in order to ensure equity and efficiency in health care delivery. Small area variations studies have emphasized investigating operations of this sort.

Not all emergency surgery is done for noncontroversial indications. Some types of emergency surgery (e.g. internal fixation of fractures) have nonoperative alternatives. For instance, a fracture of the humerus may often be treated either by closed

manipulation and plaster immobilization, or by open reduction and plating. The absolute indications for open reduction include open fractures, and postreduction nerve palsies. These absolute indications account for a minority of cases. In the vast majority of cases there is considerable latitude to choose between operative or closed treatment. It was erroneous for Chassin to state that the operation rate for humeral fracture repair was determined by the incidence rate(16).

Fractures of the neck of the femur (commonly called 'fractured hip'), on the other hand, would be a condition for which the incidence rate largely determines surgical rate. These fractures occur largely in the elderly. The period of bedrest required for healing without operative stabilization is so dangerous in the elderly that nonoperative management is essentially not practiced on this continent.

Some types of elective surgery have low variability because the condition is easy to recognize and universally treated surgically. Inguinal hernia repair has been proposed as an example of this type of procedure in North America, but not in Britain where use of a truss is still popular(2).

In general, then, the process of surgical care is complex, with many levels at which intraregional and interregional variability in case rates might occur. For most operations, it is likely that many of these potential sources contribute simultaneously to the variability observed. However, it is possible to identify some surgical or medical conditions which are much less prone to specific sources of variability. These conditions might be used as comparison groups, when studying other

conditions for these specific sources of variability. Results of this approach are discussed below, under 'Modifying the null hypothesis'.

4.3 Relationship Between Region Sizes and Sources of Variation

It is likely that multiple factors contribute to an observed rate variation. However, by defining the 'small' areas in different ways, it should be possible to highlight variability from different causes.

At the finest level of division, each individual forms his or her own region. Differences between regions will be maximal, since each will have a surgical rate of either zero or 100%. The biggest contributor to the difference should be patient characteristics with respect to morbidity - those who have had the surgery will have had a high incidence of whatever condition is treated, those who have not had the surgery will have a lower incidence (although not necessarily zero). A different fine grouping might be families, and again morbidity should explain much of the variability. Related people living together share many genetic and environmental risk factors for disease, and disease incidence would be expected to explain much of the variability between families in surgical rates.

If physician practice style is the source of variability sought, then the most logical analysis is to divide the population into regions roughly corresponding to the practice of a single surgeon or a small group of surgeons. Because only a minority of the population sees a surgeon in a given year, assigning the rest of the population to the correct practice as part of the denominator is difficult. This is what the geographic

techniques of small area variation analysis attempt to do. It is the contention of Roos et al (7) that geographic small area analysis does not work very well within cities, because people cross geographic boundaries. They demonstrate a new approach – linking patients to their outpatient physicians and then to hospitals – which allows one to study how practice differs in different hospitals, which may be where much of the small area variation within cities resides. Both of these approaches are attempts to cluster patients into groupings corresponding to the practices of one or a few surgeons, or of one or a few hospitals. This technique requires very detailed data on the entire population.

The 'practise style' hypotheses have been quite controversial, both because of perceived threat to physician autonomy(52), and because practise style is difficult to operationalize and test for. The practise style and physician uncertainty hypotheses were tested in an economic model and found not to be significant contributors to interregional variability. (The strategy was to divide demand into first occurrences demand and intensity demand)(53). The specific instance of cataract surgery has been investigated in a similar way, separating patient demand factors from physician driven demand, and it was found that patient demand factors were most important in rate variation across US metropolitan areas. (54)

If the question of interest is which factors in the health system affect either allocation of resources or access to care, then the regions must be defined along administratively sensible lines. In other words, some organizational structure must be able to do something about the variability. At a health services policy level there are many

things one cannot change including population health risk behaviours, morbidity, and health values (in the short term) all of which then may go into the 'noise' component of the observation. The signal component would be variability attributable to characteristics of the health system which might need changing. These would be things like poor access to health care in some regions, or inappropriate allocation of resources across regions. The appropriate level of aggregation for studying this type of variability seems to be the District Health Council level, which is how small area variations are analyzed in Ontario. Note that at this level of aggregation, the larger regions will include hundreds of surgeons, and any variability due to 'practise style' will have a tendency to be averaged out and missed at this level of aggregation.

An example of a small area study which has used county level aggregation to study resource access in Ontario is the study by Ferris et al. on abortion services. The counties with age standardized rates below the 25th percentile had the highest proportion of women who sought abortion services outside their county of residence. In some of these counties no abortions were performed(55).

If the question is about the impact of spending differing proportions of GDP on health, then the correct level of analysis is to compare rates in countries with different spending on or organization of their health care systems. Comparing countries is no longer comparing small areas. However, studies of this type can often be revealing about macro resource allocation questions. For example, a recent study of treatment of coronary disease found a bypass surgery rate eight times higher in the United States than in Ontario, with no difference in one year survival(56).

An important issue in the Canadian health care system is equal access to care for all residents. Another issue of increasing importance is cost effectiveness, which is necessary to preserve the system for the future. Health services policy questions of this type suggest performing analyses not at the level of the individual practice, but in larger regions corresponding to the districts which either plan or fund health care. The federal government must watch the provinces to see that people are being treated fairly in each, and the provinces must ensure that care levels are appropriately distributed to the population within each province.

The choice of region size is directly related to the types of variability one is able to identify, since the statistical answers will depend upon the relationship of variance within to variance between regions. The regions used in Ontario small area variation studies are District Health Councils, which correspond to health policy questions about access to care and distribution of resources.

4.4 Age and Sex Adjustment

The principal statistical technique applied to the analysis of small area data is age and sex adjustment. This is done to adjust for differences in the characteristics of the population which may account for some of the differences in surgical rates between regions. Sometimes the summary statistics (EQ and CV) are calculated directly from crude rates (16) and sometimes they are calculated from age and sex adjusted rates (24).

Does age and sex adjustment make a big difference to the variation in crude rates? There is empirical evidence that it does not, as published by Stevenson(34). The plots of crude versus age and sex adjusted rates from the ICES data, however, are contradictory. For hip replacement, age and sex adjustment compresses the rate variability (figure 17). For coronary artery bypass, age and sex adjustment makes very little difference to the crude rates (figure 18).

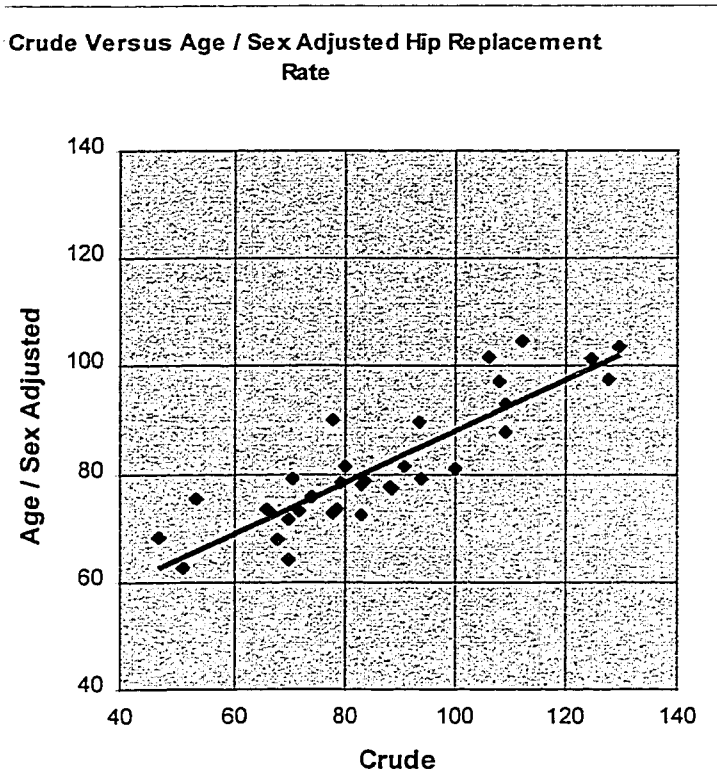


Figure 17 Crude versus Age and Sex Adjusted Hip Replacement Rates

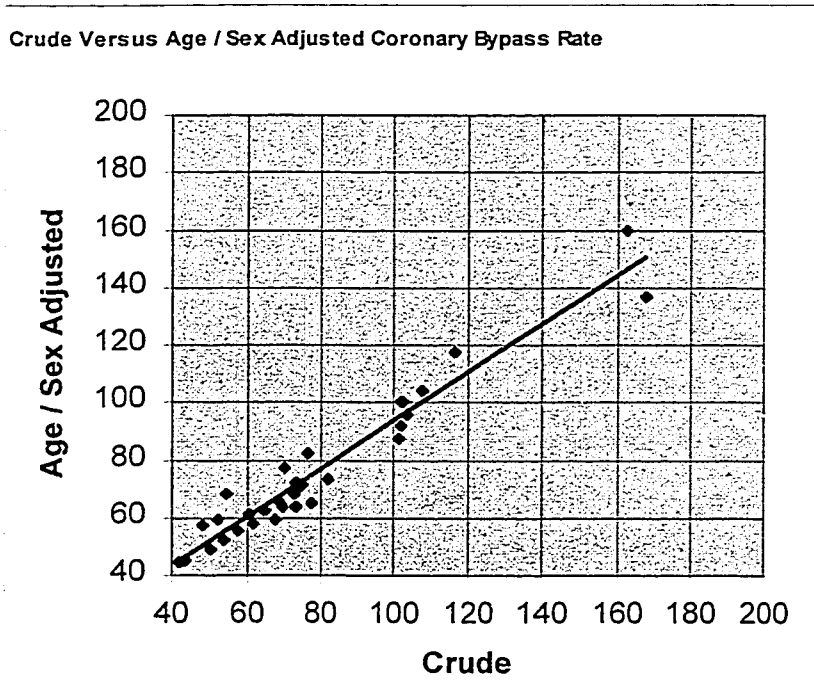


Figure 18 Crude versus Age and Sex Adjusted Coronary Bypass Rates

In table 5 there are two conditions where the summary statistics are quite different depending whether they are calculated from crude or age/sex adjusted observations. The EQ for abdominal aortic aneurysm repair is 4.3 based on crude data and 2.5 based on age sex adjusted data. For total hip replacement the EQ is 2.8 based on crude data and 1.7 based on age/sex adjusted date. EQ is calculated based on the two extreme rates and so is most likely to show significant shrinkage after age/sex adjustment. However, for both of these conditions the CV also changes considerably. For abdominal aortic aneurysm repair it is .36 crude and .24 adjusted; for hip replacement it is .24 crude and .13 adjusted. These differences support using age and sex adjusted data when calculating summary statistics.

The more important question about age and sex adjustment is whether it is sufficient to make the null hypothesis plausible. The standard null hypothesis in small area analysis is:

> H_0 (standard): Surgical rates are identical in each region, after age and sex adjustment.

In order for this hypothesis to be true, age and sex adjustment must adjust for all important population characteristics which might contribute to differences in the surgical rate. Specifically, age and sex adjustment is being asked to adjust for differences in morbidity between populations. Is this reasonable?

It turns out not to be. There is abundant evidence that morbidity differences persist between populations after age and sex adjustment is performed. Using data from the

(United States) National Health Interview Survey, Blumberg found considerable differences in morbidity among large regions in the US, and among metropolitan areas in the US, after age and sex adjustment(43). This holds whether morbidity is judged by bed disability days per person per year, by percent with limitation of activity resulting from a chronic condition, or as persons injured per 100 persons per year, or as percent with fair or poor perceived health status. Blumberg questions the assumptions behind standard small area analyses, and points out that 'epidemiologists generally use age and sex adjustment to remove confounding effects of age and not as a proxy for health status' (43).

Further evidence of difference in population morbidity after controlling for age and sex can be found in survey work from the United Kingdom (57), and in persistent differences in hip fracture rates after controlling for age, sex, and race in the United States(58). U.K. data, either from the SMR or from specific census questions about health status, always show considerable regional variability after age and sex adjustment (59).

Another field of work is also relevant. The capitation models developed for funding health regions in the UK and in Ontario based on the work of the Resource Allocation Working Party (RAWP) require an adjustment for population morbidity after age and sex are taken into account. This again implies that age and sex adjustment are not considered adequate adjustment for morbidity. Furthermore, after extensive work on how best to adjust for morbidity, these investigators have proposed the use of the Standardized Mortality Ratio (SMR)(44-46). This gives a precedent for using

empirical observations (estimates) to account for population morbidity in DHC's or similar areas, and also establishes that it is important to incorporate effects of morbidity after age and sex adjustment, for policy purposes. The RAWP formula needs a proxy for morbidity whose exact distribution by DHC is correct enough to determine resource allocation. Small area research needs a proxy for morbidity which has a correct overall variation to adjust for morbidity after age and sex adjustment. Adoption of the SMR as a morbidity adjustment in small area work is therefore justified.

The simulation work of Cain and Diehr acknowledges the frequently performed age and sex standardization of small area rates, but does not include age and sex standardization in the simulations due to complexity. Carrying age and sex standardization through the calculations creates a large parameter space (how many age and sex strata, how defined) that makes only a minimal difference to the result, so as a first approximation this can be conveniently ignored(29).

4.5 *Advantages of Monte Carlo Simulation*

Monte Carlo simulation is useful when the probability distribution of a statistic is mathematically difficult to calculate. This makes it useful in small area variations analysis, where three of the four currently used summary statistics do not have standard, known distributions.

Monte Carlo simulation is a relatively quick way to gain an empirical appreciation of the sampling distribution of any nonstandard statistic. In small area variations

analysis, this allows for design and testing of summary statistics specifically intended for policy use – for example the case count and case proportion.

Gathering sufficient empirical evidence on the distributions of statistics under the null hypothesis and non null scenarios allows one to select the better statistics – the CV and CP – as standards for judging overall variability. CV and CP are better that EQ and SCV for two reasons. First, they show much less change in value related to different populations, choices of regions, and overall surgical rates when the null hypothesis is true. Second, they are able to detect meaningful change from the null hypothesis with good power. There is sufficient evidence to choose either or both of these statistics as an agreed summary measure of overall variability.

Given an agreed summary measure for overall variability, it is possible to perform simulation based hypothesis testing. Simulation allows wide latitude in how the null hypothesis is specified. This allows incorporation of a control procedure into small area variations studies and permits adjustment of the sensitivity of the test.

There are precedents for the use of simulation in analyzing postoperative mortality rates among hospitals(60), including estimation of p-values for observations based on simulated distributions(61). Simulation has also been used to provide P-values in geographic distribution of childhood leukaemias near nuclear installations (62).

4.6 Testing Nonstandard Statistics

4.6.1 EQ:

The EQ is simple to understand, and gives some idea of the range of the rates. Simulation results, however, show the EQ to be a very misleading statistic. Intuition suggests that the EQ should be approximately one when rates are equal across regions. However, simulation shows that the expected value is considerably higher than one, given equal rates, and realistic parameters for population size and division into regions. Furthermore, the expected value varies markedly depending upon the total population, the region size, and the mean rate of surgery. The EQ is highly sensitive to the presence of regions with small populations. This limits its applicability in real situations where existing administrative boundaries define unequal region sizes. The practice of publishing an EQ without its expected value can therefore be quite misleading, and this practice should be discouraged.

The EQ also has limited power to detect real rate variability, when compared with the CV, CP, or chi-square. This means the EQ cannot be recommended as a test statistic.

4.6.2 CV:

The CV is a much more useful statistic than the EQ. It summarizes the variability in all of the data rather than concentrating on outliers. Although the expected value varies somewhat with the mean surgical rate, it does not vary markedly. This makes the CV appropriate for comparing variability between different procedures. This has

also been suggested by Cain and Diehr(63). A simple rule of thumb is that the CV should be below .1 any time the case number in the average sized region is greater than or equal to 120. A nice property of the statistic is that it is not sensitive to whether the regions are divided unequally or equally. This makes it useful when regions are defined by existing administrative boundaries. The CV also has good power to detect real variability in rates. It can be recommended as both a descriptive and a test statistic in small area analysis. Its use as a test statistic relies on performance of simulations to determine its expected value and distribution, but the rule of thumb above provides a convenient first glance assessment of the data without formal simulation.

4.6.3 SCV:

The results of this investigation do not support the continued use of SCV as a small area statistic for either testing or descriptive purposes. The number has no inherent meaning, and the expected value varies widely depending upon the underlying rate, the population size, and whether or not the regions are divided equally. These effects are so marked that the threshold values (expected 95^{th} percentiles) of the SCV vary anywhere between 1 and 400 for realistic surgical rates in the Ontario population. The statistic lacks the power to detect true rates variability. Cain and Diehr drew similar conclusions about the usefulness of the SCV(29).

4.6.4 Chi-Square:

The chi-square is the only test statistic of the four statistics commonly applied in small area variations studies. It tests for equality of proportions. Carrying chi-square through the simulations is useful in that it provides a check that the distributions obtained are as expected when the simulated value of chi-square corresponds to its calculated value.

The other reason that it is useful to simulate chi-square is to show how its expected value is changed by violating the assumptions in certain ways. For example, when readmissions are allowed, the case rates are no longer simple proportions and the standard chi-square test does not apply. The adjustment for multiple admissions suggested by Cain and Diehr approximately corresponds to the simulated chi-square results in this circumstance(39).

The application of a prior distribution to the rates (for example by using appendectomy or SMR data) changes the role of the chi-square calculation. The chi-square statistic can test for correspondence to expected values based on unequal rates in each region, but the particular distribution of unequal rates must be specified a priori. Simulation based hypothesis testing is more versatile, in that it can allow for any distribution of unequal rates that gives the same overall variability, as measured by some agreed upon statistic.

The chi-square statistic is the correct statistic to test the standard null hypothesis of equality of surgical rates in each region after age and sex adjustment. It has high

power to reject this hypothesis if it is not true. Given that age and sex adjustment is known not to account for all differences in population morbidity, it is highly unlikely that the standard null hypothesis will be strictly true. Therefore the chi-square test will almost always show a 'statistically significant' result. In fact, I have not seen a simple chi-square test in small area analysis which was either not significant or even borderline – they always show a highly statistically significant rejection of the standard null hypothesis.

The chi-square value itself is not useful as a descriptive statistic; it has no intuitive meaning in the way the EQ, CV, or CP do.

4.7 Designing New Statistics

4.7.1 Case Count and Case Proportion.

The statistical properties of the CP are as good as those of the CV, and better than those of the EQ or SCV. The CP summarizes the variability in all of the data, not just in the outliers. As with the CV, it varies relatively little across mean rates of surgery, thus making it useful for comparing variability between different procedures. The CP is not sensitive to whether the regions are divided equally or unequally, so it is useful when regions are defined by existing administrative boundaries. The CP also has good power for identifying true variability in rates. The exception is that it does not identify single outliers in low population counties. This is inherent in the design of the statistic, since this type of outlier involves an insignificant proportion of the cases performed provincewide. The CP has appropriate statistical properties to recommend

its use as a test statistic, when testing is performed by comparing results to expectations generated by simulation.

As a descriptive statistic the case count can be thought of in several different ways. The absolute number describes the number of cases (done and not done) responsible for the observed variability in the distribution. Because it is expressed as a number of cases, it has immediate meaning in terms of the number of people involved. Thinking in terms of numbers of people brings equity and access issues to the fore, and shows their magnitude. If the response to variation is an attempt to redistribute resources (or cases), the minimum number of cases that must be redistributed from high rate areas to low rate areas is equal to one half of the case count. Because this is expressed in the natural units of numbers of cases, the redistribution required can easily be translated into an estimate of the financial resources to be shifted if the marginal cost per case is known. The relative version of the case count, or CP, can be used to compare overall variability between different operations with different base rates. Although simulation is necessary to assess the exact significance of the CP, a rule of thumb is that the CP should be 0.1 or below for an expected number of cases in the average sized region of 80 or above.

4.8 Region Size

Selection of region size for small area analysis is important for two reasons. The first is that it has an important bearing on the types of variability that can be detected, and the types that will be averaged out. This is discussed in detail in section 4.3.

The second reason that region size is important is that unequally sized regions, particularly small regions where rates will be unstable, can profoundly affect the expected values of small area summary statistics. However, the results of the simulations show that unlike the EG and the SCV, both of which are highly sensitive to division of the province into unequally sized regions, neither the CV nor the CP is at all affected.

If small area analysis is being used as a tool to identify variability caused by inequalities in the health system, then the appropriate regions for analysis are preexisting administrative regions. There is no problem with using these regions if CV and CP are used for analysis, but EQ and SCV should not be used if the regions are of unequal size.

4.9 Readmissions

Readmissions are not a problem for small area analyses if person level data are available. In this case, a count of the number of primary cases can be obtained by subtracting further admissions of the same patient in the same year. The distribution of rates then can be analyzed as a proportion in the standard fashion.

If person level data are not available, the effect of readmissions is to increase the variability between regions, because those with more primary cases will have more variability in the readmission rate than will those with fewer. This is because readmissions represent a second binomial process, with mean np and variance npq.

The n for readmissions is the number of primary cases. When this goes up, variance for the readmissions and hence overall variance for total admissions rises. The expected value of the summary statistics will increase with this increase in variability.

A set of simulations was conducted to see how sensitive the statistics are to readmissions at a readmission rate of 15%. This was designed to be an extreme scenario, being far more than a reasonable estimate of the readmission rate of these operations within one year. For hip replacement, a realistic rate would be 0.9% in the first year(64). This counts revision of the index hip replacement, but does not count primary surgery on the opposite hip. For TURP, a realistic readmission rate is 4.0%(11). The figures for coronary bypass surgery are not available, but a 10 year duration of the operative result suggests a 10% readmission rate as a reasonable upper limit. One of the most notorious operations for revision is lumbar spine surgery. The readmission rate for this has been estimated at 10% in the first year, based on counting a patient from the same county and with the same birthdate as the same patient(39). This method obviously results in an overestimate.

Some reputable sources state that the problem of readmission does not apply to surgical procedures because of low or no readmission rate (27). A 15% readmission rate makes large, important differences to the EQ and SCV, but does not significantly change the CV or the CP. It is therefore possible to agree with Gittelsohn(27) and state that any realistic readmission rate for surgical interventions will not significantly affect the results of analysis if CV and CP are the statistics chosen.

4.10 Modifying the Null Hypothesis

Simulation allows the null hypothesis for small area variations to be modified. One possible use of this is to allow for a certain variability in the true rates of procedures from region to region. This variability can be based, for instance, on an operation where the main effects of interest are unlikely to be present.

Traditional methods of small area analysis, involve 'testing statistical null hypotheses which are known to be false'(36). The standard null hypothesis is:

H_o(standard): The true rates of surgery in each region are equal.

This null hypothesis is known to be false because age and sex adjustment of surgical rates does not account for all the differences in population morbidity. Alternative methods of analysing small area rates (e.g. Bayesian analysis) consider prior distributions to describe the data, without justifying the prior distributions. The prior distributions are derived from the data being tested, and not from a second set of real world data used as a control.

Modifying the null hypothesis to permit differences in the true interregional rates that are not important to policy is an attractive alternative. There may be many sources of this true interregional variability. Based on preceding arguments, morbidity differences between populations which are not accounted for by age and sex adjustment may play a large role. Although the ideal data for testing these modified

null hypotheses do not exist, appendectomy and SMR data can be used to explore the technique.

Several arguments support the use of appendectomy rates as a comparison group. The first is the argument about the nature of the surgical process presented in section 4.2. The second is the high quality of appendectomy data available in Ontario. Positive primary appendectomy rates are coded only if the clinical diagnosis is appendicitis, and the diagnosis is confirmed both clinically and pathologically. Appendectomy incidental to other intraabdominal procedures is specifically excluded(24). Appendectomy rates are among the least variable in Ontario, and this is a reflection of previous experience in other jurisdictions. In 1975 Vayda found appendectomy to be the least variable of all operative procedures studied, and found that appendectomy rates were the same for both males and females in Canada, England, and Wales(15). Later, appendectomy rates were found to be equal across countries and of low variability within countries in the USA, Norway, and Britain (2). Further evidence from medicare benificiaries across the United States showed that 'the rates for appendectomy and inguinal hernia repair – procedures about which consensus exists—have usually exhibited little variation in other studies, and they followed a similar pattern in our study.' (16). There is certainly sufficient evidence to entertain appendectomy as a procedure whose variability does not indicate a need to change the health system. In this way it might be used as a benchmark to which other procedures could be compared, before drawing the conclusion of excessive variability in these other procedures.

Femur fracture admissions are of low variability and were used as a benchmark for 'obligatory' paediatric admission in a study of the role of nonclinical factors in pediatric hospitalizations (65). This supports the principle of using appendectomy in the same way, although the technique clearly varies.

Similar use might be made of data regarding hip fractures. Compared with appendectomy, there is even less controversy regarding the diagnosis. More importantly, the issue of mild cases that may recover spontaneously is less bothersome. Again the condition is universally treated with an operation. A disadvantage of using hip fracture rates to define a modified null hypothesis is that hip fractures occur primarily in the elderly population, whereas appendicitis occurs across all age ranges. This may be an advantage if the comparison procedure is one which also occurs primarily in the elderly. There is measurable regional variation in the incidence of hip fracture amongst U.S. white women aged 65 years and older (58). The mean rate is 8.23 fractures per 1000, with an interquartile range of 2.25. Hip fracture is almost universally treated surgically, so surgical rate equals incidence rate and the observed variability is likely a result of disease incidence. The correct response of health policy would be to treat the incidence of disease as observed, not to try to decrease the interregional variability with measures aimed to control access to or use of surgery.

Unfortunately, the hip fracture repair rates for Ontario DHC's are not available so practically speaking this cannot be used as a comparator.

The other comparator available is SMR's within the regions. The advantage of using the SMR is that it is a reflection of morbidity differences in the population after age and sex adjustment. It is also a standard proxy measure, with good documentation and justification of its use for morbidity adjustment for resource allocation(44-46). Two reservations about using the SMR as a comparator are: first, that the SMR cannot be 'scaled down' in the same way appendectomy rates were; and second, SMR's are calculated by indirect rather than direct age standardization and thus are not (strictly speaking) comparable to one another.

The findings of the modified null hypotheses are interesting. Compared with appendectomy surgery, only coronary bypass surgery and uterine dilatation and curettage were shown to have statistically significant variability in the Ontario data. The variability of abdominal aortic aneurysm repair was marginal. The rates for total hip replacement and transurethral resection of the prostate varied no more than the appendectomy rates. Adjustment by SMR's made a slightly bigger difference to the expected values, meaning only coronary artery bypass grafting rates showed significant variability.

4.11 Interpretation for Health Policy

How does Monte Carlo simulation add to the interpretability of small area variations data? Total hip replacement will be used as an example. Standard presentation would be to state that the rate of hip replacement varies by a factor of 1.7 from the lowest to the highest region. The coefficient of variation is 0.13; this summarizes

variability in all regions and relates the standard deviation to the mean. The systematic component of variation is 15.3. This figure attempts to summarize the systematic portion of the variation while subtracting the portion due to chance. Overall, the observed rates differ from equal rates in each region with high statistical significance, with a chi-square of 102, and a p-value of less than .0001.

Adding simulation based hypothesis testing allows the statistics to be presented within context. The observed extremal quotient of 1.7 has an expected mean value of 1.8 and 95th percentile of 2.4, so the overall range does not differ from that expected if true rates are equal in each region (p=0.52). The observed coefficient of variation is 0.13; this is statistically significantly different (p<.05) from the CV expected if true rates are equal. The expected mean CV is 0.074 and the expected 95th percentile is 0.089. Therefore the overall variability in intraregional rates is statistically significant, even though the extreme rates are not further apart than expected. The observed systematic component of variation is 15.3, which exceeds the 95th percentile of 9.02.

The case count is 586 and the case proportion is 0.103. Case count is the number of cases responsible for the observed intraregional variability and case proportion is this number expressed as a proportion of the total number of cases perfromed. Both of these are statistically significant (p<.05) when compared with the 95th percentiles of the expected distribution, which are 375 and 0.065 respectively. 586 cases account for the intraregional rate variability observed. This is 10.3 percent of the cases performed. Rates could be made equal across the province by moving 293 cases from

higher rate areas to lower rate areas. If this were desirable then known marginal costs could be applied to determine what funding to shift from region to region. Alternatively, the lower rate regions could be brought up to the mean rate (which would then increase), by funding an additional 293 cases without subtracting resources. Or, it could be accepted that 90% of cases performed across the province are equally distributed and that is satisfactory. Two caveats are necessary when interpreting these numbers. The first is that the case counts are based on age and sex adjusted data, so will not correspond to crude case counts. The second is that the mean rate is not necessarily the right rate. It may be that a higher mean rate of surgery provincewide is beneficial to the population. If other data existed on the value of different overall surgery rates, the case count could be easily modified to describe deviation from the desired rate rather than the observed mean. A case count based on a desired rate far from the mean, although possibly a useful construct in some cases, is no longer a statistic which describes variability in the data.

An alternative to testing for strict equality of rates in each region is to test whether total hip replacement rates vary more than appendectomy rates. Basing expectation on total hip replacement true rates varying proportionally to appendectomy true rates, the 95th percentile for the EQ is 2.9, for the CV is .19, and for the CP is 0.15. None of the observed rates (EQ 1.7, CV 0.133, CP 0.103) is then statistically significant. This suggests that the observed variation is not large enough that differences in physician practice patterns or health system attributes will be found. A study specifically

looking for differences in physician practice patterns related to hip and knee replacement in Ontario did not find these differences(23).

Coronary artery bypass is a condition for which some evidence exists regarding the optimal overall surgical rate. Although the rate of bypass surgery following myocardial infarction is eight times higher in the United States than in Ontario, the survival rates are not different(56). This suggests that there are not survival gains to be made by increasing Ontario's overall bypass rate (this says nothing about controlling morbidity from coronary disease). However, 923 cases, or 18% of those performed, are distributed unequally across regions in Ontario. This is statistically significant ($p<0.05$) when compared with the 95th percentile of the expected distribution, which is 7%. Access to care may differ significantly across the province, and may be an issue for one in five people requiring or receiving cardiac bypass surgery. The statistically significant difference persists even when the variability in coronary bypass is compared to the variability in appendectomy or in SMR, further suggesting that it is aspects of the health system and not population characteristics that are responsible for the observed variability.

5 Recommendations

Monte Carlo simulation has much to offer for small area analysis. The advantages include ability to interpret the standard descriptive statistics against an expected value generated by simulation.

The adaptation of the case count for use as a small area statistic is worthwhile. This measure is interpretable either as a number of cases (people) when considering equity and access issues, or as an estimate of resources required to correct inappropriate distribution of surgical rates. The statistical properties of this measure are as good as those of the CV, and superior to those of the EQ and SCV, so it may be used as a test statistic as well.

Simulation allows for the testing of modified null hypotheses which are more specific to the variability sought. This is important if small area variations studies are thought of as a 'screening test', a construct proposed by ICES. "Variations in service profiles among small areas are like screening tests... They tell us there may be a problem."(24). A screening test is a useful way of thinking of small area studies. They are relatively cheap to do if they use existing administrative data. Like a screening test, the sensitivity and specificity of small area studies needs to be balanced. The current approach to statistical testing, the chi-square, is highly sensitive but not very specific. Modifying the statistical testing by using simulations for analysis, and collecting appropriate prior distributions that allow some true interregional variability, but do not include the policy relevant reasons sought, will increase the specificity of small area analysis as a screening tool. This will allow the resources for more extensive investigation to be focused where they are most likely to be useful.

Reference List

1. Glover JA. The incidence of tonsillectomy in schoolchildren. Proceedings of the Royal Society of Medicine 1938;31:1219-36.

2. McPherson K, Wennberg JE, Hovind OB, Clifford P. Small-Area Variations in the Use of Common Surgical Procedures: An International Comparison of New England, England, and Norway. New England Journal of Medicine 1982;307(21):1310-4.

3. Wennberg J, Gittelsohn A. Small Area Variations in Health Care Delivery. Science 1973;182:1102-8.

4. Wennberg J, Gittelsohn A. Variations in Health Care among Small Areas. Scientific American 1982;246(4):120-34.

5. Wennberg JE. Dealing with medical practice variations: a proposal for action. Health Affairs 1984;4:6-32.

6. Roos LL, Mustard CA, Nicol JP, McLerran DF, Malenka DJ, Young TK, Cohen MM. Registries and administrative data: organization and accuracy. Medical Care 1993;31(3):201-12.

7. Roos NP. Linking patients to hospitals. Defining urban hospital service populations. Medical Care 1993;31(5 Suppl):YS6-15.

8. Roos NP, Black CD, Roos LL, Tate RB, Carriere KC. A population-based approach to monitoring adverse outcomes of medical care. Medical Care 1995;33(2):127-38.

9. Roos NP, Roos LL. High and low surgical rates: risk factors for area residents. American Journal of Public Health 1981;71:591-600.

10. Roos NP, Roos LL. Surgical rate variations: do they reflect the health or socioeconomic characteristics of the population? Medical Care 1982;20:945-58.

11. Roos NP, Wennberg JE, Malenka DJ, Fisher ES, McPherson K, Anderson TF, Cohen MM, Ramsey E. Mortality and Reoperation after open transurethral resection of the prostate for benign prostatic hyperplasia. New England Journal of Medicine 1989;320:1120-4.

12. Wennberg JE, Roos LL, Sola L, Schori A, Jaffe R. Use of claims data systems to evaluate health care outcomes: Mortality and re-operation following prostatectomy. JAMA 1987;257(7):933-6.

13. Wennberg JE, Barry MJ, Fowler FJ, Mulley A. Outcomes research, PORTs, and health care reform. Annals of the New York Academy of Sciences 1993;703:52-62.

14. Freund DA, Katz BP, Callahan CM. Patient outcomes research teams: examples from a study on knee replacement. Annals of the New York Academy of Sciences 1993;703:86-94; discussion 94-5.

15. Vayda E. A comparison of surgical rates in Canada an in England and Wales. New England Journal of Medicine 1973;289:1224-9.

16. Chassin MR, Brook RH, Park RE, Keesey J, Fink A, Kosecoff J, Kahn K, Merrick N, Solomon DH. Variations in the use of medical and surgical services by the medicare populaiton. New England Journal of Medicine 1986;314(5):285-90.

17. Chassin MR, Kosecoff J, Park RE, et.al. Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. JAMA 1987;258:2533-7.

18. Park RE. Does Inappropriate Use Explain Small Area Variations in the Use of Health Services? A Reply. Health Services Research 1993;28(4):401-10.

19. Leape LL, Park RE, Solomon DH, et.al. Does inappropriate use explain small-area variations in the use of health care services? JAMA 1990;263:669-72.

20. Cain K, Diehr P. The relationship betwen small-area variations in the use of health care services and inappropriate use: a commentary. Health Services Research 1993;28(4):411-8.

21. Leape LL, Park RE, Solomon DA, et.al. Relation between surgeons' practice volumes and geographic variation in the rate of carotid endarterectomy. New England Journal of Medicine 1989;321:653-7.

22. Davidson G. "Does inappropriate use explain small-area variations in the use of health care services?" A critique. Health Services Research 1993;28(4):389-400; discussion 401-.

23. van Walraven C, Paterson JM, Kapral M, Chan B, Bell M, Hawker G, Gollish J, Schatzker J, Williams JI, Naylor CD. Appropriateness of primary total hip and knee replacements in regions of Ontario with high and low utilization rates. Canadian Medical Association Journal 1996;155(6):697-706.

24. Goel V; Williams JI; Anderson GM, et al. Goel V, Williams JI, Anderson GM et al. editors. editors.Patterns Of Health Care In Ontario. The ICES Practice Atlas. 2 ed. Ottawa: Canadian Medical Association; 1996.

25. Kelly A, Jones W. Small area variation in the utilization of mental health services: implications for health planning and allocation of resources. Canadian Journal of Psychiatry - Revue Canadienne de Psychiatrie 1995;40(9):527-32.

26. Anis AH, Carruthers SG, Carter AO, Kierulf J. Variability in prescription drug utilization: issues for research. Canadian Medical Association Journal 1996;154(5):635-40.

27. Gittelsohn A, Powe NR. Small area variations in health care delivery in Maryland. Health Services Research 1995;30(2):295-317.

28. Folland S, Stano M. Small Area Variations : A Critical Review of Propositions, Methods, and Evidence. Medical Care Review 1990;47:419-65.

29. Diehr P, Cain K, Connell F, Volinn E. What is Too Much Variation? The Null Hypothesis in Small Area Analysis. Health Scervices Research 1990;24:741-71.

30. Kunst, A. E. and Mackenbach, J. P. Measuring Socio-Economic Inequalities in Health. 1994. Copenhagen, World Health Organization.

31. Mackenbach JP, Kunst AE. Measuring the Magnitude of Socio-Economic Inequalities in Health: An Overview of Available Measures Illustrated with Two Examples From Europe. [In Press] Social Sciences in Medicine 1997;

32. Kennedy BP, Kawachi I, Prothrow-Stith D. Income distribution and mortality: cross sectional ecological study of the Robin Hood index in the United States. BMJ 1996;312:1004-7.

33. Wennberg JE. Future directions for small area variations. Medical Care 1993;31(5 Suppl):YS75-80.

34. Stevenson JM, Olson DR. Methods for analysing county-level mortality rates. Statistics in Medicine 1993;12(3-4):393-401.

35. Shwartz M, Ash AS, Anderson J, Iezzoni LI, Payne SM, Restuccia JD. Small area variations in hospitalization rates: how much you see depends on how you look. Medical Care 1994;32(3):189-201.

36. Cressie N. Regional mapping of incidence rates using spatial Bayesian models. Medical Care 1993;31(5 Suppl):YS60-5.

37. Gatsonis C, Normand SL, Liu C, Morris C. Geographic Variation of Procedure Utilization: A hierarchical model approach. Medical Care 1993;31(5):YS54-YS59

38. Carriere KC, Roos LL. Comparing standardized rates of events. American Journal of Epidemiology 1994;140(5):472-82.

39. Cain K, Diehr P. Testing the Null Hypothesis in Small Area Analysis. Health Services Research 1992;27(3):267-94.

40. Paul-Shaheen P, Clark J, William D. Small area analysis: a review and analysis of the North American Literature. Journal of Health Politics, Policy, and Law 1987;12:741-809.

41. Dunn W.N. Public Policy. An Introduction. Englewood Cliffs: Prentice-Hall; 1981.

42. Bratley P; Fox BL; Schrage LE. A Guide to Simulation. New York: Springer-Verlag; 1983.

43. Blumberg MS. Inter-area variations in age-adjusted health status. Medical Care 1987;25:340-53.

44. Eyles J, Birch S, Chambers S, Hurley J, Hutchison B. A Needs - Based Methodology for Allocating Health Care Resources in Ontario, Canada: Development and an Application. Social Sciences in Medicine 1991;33(4):489-500.

45. Carr-Hill RA, Sheldon TA, Smith P, Martin S, Peacock S, Hardman G. Allocating resources to health authorities: development of method for small area analysis of use of inpatient services BMJ 1994;309(6961):1046-9.

46. Smith P, Sheldon TA, Carr-Hill RA, Martin S, Peacock S, Hardman G. Allocating resources to health authorities: results and policy implications of small area analysis of use of inpatient services BMJ 1994;309(6961):1050-4.

47. Hale DA, Molloy M, Pearl RH, Schutt DC, Jaques DP. Appendectomy: a contemporary appraisal. Annals of Surgery 1997;225(3):252-61.

48. DeGroot MH. Probability and Statistics. 2 ed. Reading, Massachussets: Addison-Wesley; 1986. 691p.

49. Wen SW, Naylor CD. Diagnostic accuracy and short-term surgical outcomes in cases of suspected acute appendicitis. Canadian Medical Association Journal 1995;152(10):1617-26.

50. Wade DS, Marrow SE, Balsara ZN, Burkhard TK, Goff WB. Accuracy of ultrasound in the diagnosis of acute appendicitis compared with the surgeon's clinical impression. Archives of Surgery 1993;128(9):1039-44.

51. Eriksson S, Granstrom L. Randomized controlled trial of appendicectomy versus antibiotic therapy for acute appendicitis. British Journal of Surgery 1995;82(2):166-9.

52. O'Connor GT, Plume SK, Wennberg JE. Regional organization for outcomes research. Annals of the New York Academy of Sciences 1993;703:44-51.

53. Folland S, Stano M. Sources of small area variations in the use of medical care. Journal of Health Economics 1989;8:85-107.

54. Escarce JJ. Would eliminating differences in physician practice style reduce geographic variations in cataract surgery rates? Medical Care 1993;31(12):1106-18.

55. Ferris LE, McMain-Klein M. Small-area variations in utilization of abortion services in Ontario from 1985 to 1992. Canadian Medical Association Journal 1995;152(11):1801-7.

56. Tu JV, Pashos CL, Naylor CD, Chen E, Normand SL, Newhouse JP, McNeil BJ. Use of cardiac procedures and outcomes in elderly patients with myocardial infarction in the United States and Canada [published erratum appears in N Engl J Med 1997 Jul 10;337(2):139]. New England Journal of Medicine 1997;336(21):1500-5.

57. Payne JN, Coy J, Patterson S, Milner PC. Is use of hospital services a proxy for morbidity? A small area comparison of the prevalence of arthritis, depression, dyspepsia, obesity, and respiratory disease with inpatient admission rates for these disorders in England. Journal of Epidemiology & Community Health 1994;48(1):74-8.

58. Jacobsen SJ, Goldberg J, Miles TP, Brody JA, Stiers W, Rimm AA. Regional Variation in the Incidence of Hip Fracture. US White Women Aged 65 Years and Older. JAMA 1990;264(4):500-2.

59. Martin S, Sheldon TA, Smith P. Interpreting the new illness question in the UK census for health research on small areas. Journal of Epidemiology & Community Health 1995;49(6):634-41.

60. Luft HS, Brown BW, Jr. Calculating the probability of rare events: why settle for an approximation? Health Services Research 1993;28(4):419-39.

61. Flanders WD, Shipp CC, FitzGerald DM, Lin LS. Analysis of variations in mortality rates with small numbers. Health Services Research 1994;29(4):461-71.

62. Bithell JF, Dutton SJ, Draper GJ, Neary NM. Distribution of childhood leukaemias and non-Hodgkin's lymphomas near nuclear installations in England and Wales. BMJ 1994;309(6953):501-5.

63. Diehr P, Cain K, Ye Z, Abdul-Salam F. Small area variation analysis. Methods for comparing several diagnosis-related groups. Medical Care 1993;31(5 Suppl):YS45-53.

64. Kavanagh BF, Dewitz MA, Ilstrup DM. Charnley total hip arthroplasty with cement: Fifteen year results. Journal of Bone and Joint Surgery 1989;71-A:1496-503.

65. Goodman DC, Fisher ES, Gittelsohn A, Chang CH, Fleming C. Why are children hospitalized? The role of non-clinical factors in pediatric hospitalizations. Pediatrics 1994;93(6 Pt 1):896-902.

Appendix 1: Equations

Notation: X = rate of surgery, P = population, t= total, i = in ith region, O = observed number of cases, E = expected number of cases, k = number of regions.

$$E.Q. = \text{maximum rate} / \text{minimum rate}$$

$$C.V. = \frac{1}{\overline{X} * P_t} * \sqrt{\sum_{regions} (X_i - \overline{X})^2 * P_i}$$

$$SCV = (1/k)\{\sum_{regions} ((O - E/E)^2 - \sum_{regions} (1/E)\} * 1000$$

$$\chi^2 = \sum_{regions} \frac{(observedcases - expectedcases)^2}{expectedcases}$$

$$C.C. = \sum_{regions} |observedcases - expectedcases|$$

Appendix 2: Program Code

```
Option Base 1

'Monte Carlo Simulation Routines for Small Area Variations
'Andrew Howard MD FRCSC
'Student, Division of Epidemiology
'University of Ottawa, CANADA

'see text of thesis section 2.2 for description

Public Sub multisim()

'sheet 10 is set up with number of simulations and data for first'
'remaining sheets have data for subsequent simulations
'input on even numbered sheets, output is place on odd numbered sheets
'sheet 1 is the input sheet for singlesimulation subroutine
'which actually does the work.


Sheet10.Activate

numsims = Cells(1, 7)

For simul = 10 To (10 + (2 * (numsims - 1))) Step 2

   Worksheets(simul).Activate

   Cells.Select
   Selection.Copy
   Sheets("Sheet1").Select
   Range("A1").Select
   ActiveSheet.Paste


   k = Sheet1.Cells(3, 3)
   n = Sheet1.Cells(4, 3)


   Sheet2.Activate
   Cells.Select
   Selection.Clear

   singlesimulation

   Sheets("Sheet2").Activate
   Cells.Select
   Selection.Copy
   Worksheets(simul + 1).Activate
   Worksheets(simul + 1).Select
   Range("A1").Select
   ActiveSheet.Paste
```

```
Next simul

End Sub

Public Sub singlesimulation()

'performs a single simulation using sheet1
'as the input and sheet2 as the output.

' calculates much faster if sheet4 clear to begin with
'as otherwise calculates formulas linked to 3 during numbergen

Sheet3.Activate
Cells.Select
Selection.Clear
Sheet4.Activate
Cells.Select
Selection.Clear


poissongennum
'eventually will allow choice of rng's
'poissonreadmit

numtorates
eq
cv
scv
idchi
iqrmax
outputpercentiles

End Sub

Sub bingennum()
'generate a sheet with numbers of cases (not rates) using binomial
'k regions (found sheet1 cell3,3
'n trials (found sheet 1 cell4,3
'populations in sheet1 cells(6,1..k)
'prates in sheet1 cells (8,1..k)

'output to sheet3 k rows by n columns
k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

For i = 1 To k

   population = Sheet1.Cells(6, i)
   prate = Sheet1.Cells(8, i) / 100000
   Application.Run "ATPVBAEN.XLA!Random", Sheet3.Range(Cells(1, i), Cells(n, i)), 1 _
      , n, 4, 12, prate, population

Next i

End Sub

Sub poissongennum()
'generate a sheet with numbers of cases (not rates) using poisson
'k regions (found sheet1 cell3,3
'n trials (fournd sheet 1 cell4,3
```

```
'populations in sheet1 cells(6,1..k)
'prates in sheet1 cells (8,1..k)

'output to sheet3 k rows by n columns
k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)
Sheet3.Activate

For i = 1 To k

   population = Sheet1.Cells(6, i)
   prate = Sheet1.Cells(8, i) / 100000
   lambda = population * prate

   Application.Run "ATPVBAEN.XLA!Random", Sheet3.Range(Cells(1, i), Cells(n, i)), 1 _
      , n, 5, , lambda



Next i

End Sub

Sub poissonreadmit()
'generate a sheet with numbers of cases (not rates) using poisson
'k regions (found sheet1 cell3,3
'n trials (fournd sheet 1 cell4,3
'populations in sheet1 cells(6,1..k)
'prates in sheet1 cells (8,1..k)
'probability of readmission in same year (0-1) is in sheet1.cells(12,1)
'output to sheet3 k rows by n columns

'nb. the readmission process generates a binomial based on the number
'of primary cases and the readmit rate.  this means a given patient
'can have only 0,1, or 2 admissions for a procedure.  this is
'sensible for total hip, cabg, aaa etc but is not so for medical
'admissions such as chf or asthma

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)
preadmit = Sheet1.Cells(12, 1)
Sheet3.Activate

For i = 1 To k

   population = Sheet1.Cells(6, i)
   prate = Sheet1.Cells(8, i) / 100000
   tlambda = population * prate
   lambda = tlambda / (1 + preadmit)

   Application.Run "ATPVBAEN.XLA!Random", Sheet3.Range(Cells(1, i), Cells(n, i)), 1 _
      , n, 5, , lambda

   For j = 1 To n

      flag = 0
      ncases = Sheet3.Cells(j, i)
      If ncases = 0 Then flag = 1   ' avoids error message from rng
      If ncases = 0 Then ncases = 1
      Sheet3.Cells(1, k + 1).Clear
```

```
        Application.Run "ATPVBAEN.XLA!Random", Sheet3.Range(Cells(1, k + 1), Cells(1, k + 1)), 1, 1, 4, 12, preadmit, ncases
        readmissions = Sheet3.Cells(1, k + 1)
        Sheet3.Cells(j, i) = ncases + readmissions
        If flag = 1 Then Sheet3.Cells(j, i) = 0

    Next j

Next i

End Sub

Sub normgennum()
'generate a sheet with numbers of cases (not rates) using normal
'k regions (found sheet1 cell3,3
'n trials (fournd sheet 1 cell4,3
'populations in sheet1 cells(6,1..k)
'prates in sheet1 cells (8,1..k)

'output to sheet3 k rows by n columns
k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

For i = 1 To k

    population = Sheet1.Cells(6, i)
    prate = Sheet1.Cells(8, i) / 100000
    mean = population * prate
    sdev = population * Sqr((((prate) * (1 - prate)) / population))
    Application.Run "ATPVBAEN.XLA!Random", Sheet3.Range(Cells(1, i), Cells(n, i)), 1 _
        , n, 2, 12, mean, sdev


Next i

End Sub

Sub numtorates()

'converts numbers of cases on sheet 3 columns 1 to k, n rows
'into rates on sheet 4 coumns 1 to k
'by dividing by population vector in sheet 1 row 6

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)


Sheet4.Activate
Sheet4.Range(Cells(1, 1), Cells(n, k)).FormulaR1C1 = "=Sheet3!RC/Sheet1!R6C*100000"
Sheet4.Calculate

'For i = 1 To k
'   Cells(1, i).Formula = 100000 * (Cells(1, i)) / (Sheet1.Cells(6, i))
'Next i
'Sheet4.Range(Cells(1, 1), Cells(1, k)).Copy
'Sheet4.Range(Cells(2, 1), Cells(n, k)).PasteSpecial (xlPasteFormulas)
'Sheet4.Calculate


End Sub
```

```
Sub eq()
'calculates the eq of k rates in cols 1 thru k , places in col k+1
'performs this operation for n rows of the simulation

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

Sheet4.Activate

   For i = 1 To n
   minrate = WorksheetFunction.Min(Range(Cells(i, 1), Cells(i, k)))
   maxrate = WorksheetFunction.Max(Range(Cells(i, 1), Cells(i, k)))
   If minrate = 0 Then minrate = 0.1
   teq = maxrate / minrate
   Cells(i, k + 1) = teq
   Next i


'Sheet4.Range(Cells(1, k + 1), Cells(n, k + 1)).FormulaR1C1 = "=max(RC1:RC)/min(RC1:RC)"



End Sub

Sub cv()
'calculates the cv of k rates in cols 1 thru k , places in col k+2
'performs this operation for n rows of the simulation

'wmrate is weighted mean rate
'wsdrate is weighted standard deviation of rates
'p is a vector of populations

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)
Dim p(35) As Variant
pt = 0
For j = 1 To k
p(j) = Sheet1.Cells(6, j)
pt = pt + p(j)
Next j


Sheet4.Activate

   For i = 1 To n
   wmrate = 0
      For j = 1 To k
      wmrate = wmrate + Cells(i, j) * p(j)
      Next j
   wmrate = wmrate * (1 / pt)

   wsdrate = 0
      For j = 1 To k
      wsdrate = wsdrate + (Cells(i, j) - wmrate) ^ 2 * p(j)
      Next j
   wsdrate = Sqr(wsdrate * (1 / pt))

   tcv = wsdrate / wmrate

   Cells(i, k + 2) = tcv
```

```
    Next i

End Sub

Sub scv()

'calculates the SCV (systematic component of variation) of k rates,
'places the result in row k+3 of sheet4

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)
Dim p(35) As Variant
pt = 0
For j = 1 To k
p(j) = Sheet1.Cells(6, j)
pt = pt + p(j)
Next j

For i = 1 To n

   total = 0 'the total case number is needed to calculate expected
         'values for this row

   For j = 1 To k
      total = total + Sheet3.Cells(i, j)
   Next j

   baserate = total / pt          'base rate for this row of cases
   tscv = 0: sum1 = 0: sum2 = 0

   For j = 1 To k
      o = Sheet3.Cells(i, j)
      e = baserate * p(j)
      sum1 = sum1 + ((o - e) / e) ^ 2
      sum2 = sum2 + (1 / e)
   Next j

   tscv = (1000 / k) * (sum1 - sum2)   'scv per diehr 1990

   Sheet4.Cells(i, k + 3) = tscv

Next i

End Sub

Sub idchi()
'calculates chisquare and places in row k+4, because calculation
'is so similar to that for id

'calculates the index of dissimilarity, ida (a number) and
'idr (a proportion), of k numbers of cases,
'places the result in row k+5 and k+6 of sheet4

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)
Dim p(35) As Variant
pt = 0
For j = 1 To k
p(j) = Sheet1.Cells(6, j)
pt = pt + p(j)
```

```
Next j

For i = 1 To n      'for each simulated set of rates

   total = 0 'the total case number is needed to calculate expected
         'values for this row

   For j = 1 To k
      total = total + Sheet3.Cells(i, j)
   Next j

   baserate = total / pt          'base rate for this row of cases
   tid = 0        'accumulates index of dissimilarity
   tchi = 0       'accumulates chi square

   For j = 1 To k                 'for each rate in the set
      o = Sheet3.Cells(i, j)        'obs and exp values
      e = baserate * p(j)
      tid = tid + Abs(o - e)        'used to calculte id, chisq etc.
      tchi = tchi + ((o - e) ^ 2) / e
   Next j

   tidr = tid / total

   Sheet4.Cells(i, k + 4) = tchi      'chisquare
   Sheet4.Cells(i, k + 5) = tid       'id or index of dissimilarity
   Sheet4.Cells(i, k + 6) = tidr      'relative version id

Next i

End Sub

Sub iqrmax()
'calculates interquartile range, alone and relative to median,
'and minimum and maximum rate, alone and relative to median, results go
' into rows k+7 thru k+12 of sheet 4

k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

Sheet4.Activate

For i = 1 To n

   minrate = WorksheetFunction.Min(Range(Cells(i, 1), Cells(i, k)))
   maxrate = WorksheetFunction.Max(Range(Cells(i, 1), Cells(i, k)))
   medrate = WorksheetFunction.Median(Range(Cells(i, 1), Cells(i, k)))
   lq = WorksheetFunction.Percentile(Range(Cells(i, 1), Cells(i, k)), 0.25)
   uq = WorksheetFunction.Percentile(Range(Cells(i, 1), Cells(i, k)), 0.75)

   tiqr = (uq - lq)              'interquartile range
   tiqrr = (uq - lq) / medrate    'interquartile range over median
   tminr = minrate / medrate      'minimum and
   tmaxr = maxrate / medrate      'max rate over median


   Cells(i, k + 7) = minrate
   Cells(i, k + 8) = tminr
   Cells(i, k + 9) = tiqr
   Cells(i, k + 10) = tiqrr
```

```
    Cells(i, k + 11) = maxrate
    Cells(i, k + 12) = tmaxr

Next i

End Sub


Sub chi()

'this does not calculate chisquare
'because this is now calculated as part of sub idchi for efficiency

End Sub



Sub varypopulations()

' sheet10 contains the real populations
'sheets 12 thru 30 will be created, varying the
'populations smoothly until they are all equal

'then multisim is run

k = Sheet10.Cells(3, 3)        'k regions
n = Sheet10.Cells(3, 4)        'n cases
Dim p(35) As Variant           'vector of populations
pt = 0

For j = 1 To k                 'read population vector from sheet 10
    p(j) = Sheet10.Cells(6, j)
    pt = pt + p(j)
Next j


Sheet10.Cells(1, 7) = 11       'set for 11 simulations
Sheet10.Activate
Cells.Select
Selection.Copy

For i = 12 To 30 Step 2        'copy sheet10 to sheets 12 thru 30

    Worksheets(i).Select
    Range("A1").Select
    ActiveSheet.Paste

Next i

meanpop = pt / k

For scaling = 0 To 1 Step 0.1  'scale between actual and equal pops

    Sheets(10 + (20 * scaling)).Activate
    Cells(2, 3).Formula = "Population scaling factor  " & scaling

    For region = 1 To k

        If p(region) <= meanpop Then Cells(6, region) = meanpop - (scaling * (meanpop - p(region)))
```

```
    If p(region) > meanpop Then Cells(6, region) = meanpop + (scaling * (p(region) - meanpop))

  Next region

Next scaling

End Sub


Public Sub output()

' generates an output page on sheet 2, based on the statistics
'calculated on sheet 4. currently set for 12 statistics, needs
'modifying if more or fewer.
'output page includes the mean,stdev,min,max,and 95%ile of statistics


k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

  Sheet2.Activate
  Sheet2.Select
  Range("A1").Select
  Selection.Font.Bold = True
  ActiveCell.FormulaR1C1 = "Monte Carlo Simulation SAV Output"
  Range("A2").Select
  ActiveCell.FormulaR1C1 = "Simulation Title:"
  Sheet1.Activate
  Range("C2").Select
  Selection.Copy
  Sheet2.Activate
  Range("C2").Select
  ActiveCell.PasteSpecial (xlPasteValues)
  Range("B4").Select
  ActiveCell.FormulaR1C1 = "EQ"
  Range("C4").Select
  ActiveCell.FormulaR1C1 = "CV"
  Range("D4").Select
  ActiveCell.FormulaR1C1 = "SCV"
  Range("E4").Select
  ActiveCell.FormulaR1C1 = "X2"
  Range("F4").Select
  ActiveCell.FormulaR1C1 = "IDA"
  Range("G4").Select
  ActiveCell.FormulaR1C1 = "IDR"
  Range("H4").Select
  ActiveCell.FormulaR1C1 = "MIN"
  Range("I4").Select
  ActiveCell.FormulaR1C1 = "MIN/MED"
  Range("J4").Select
  ActiveCell.FormulaR1C1 = "IQR"
  Range("K4").Select
  ActiveCell.FormulaR1C1 = "IQR/MED"
  Range("L4").Select
  ActiveCell.FormulaR1C1 = "MAX"
  Range("M4").Select
  ActiveCell.FormulaR1C1 = "MAX/MED"


  Range("A5").Select
```

```
    ActiveCell.FormulaR1C1 = "mean"
    Range("A6").Select
    ActiveCell.FormulaR1C1 = "stdev"
    Range("A7").Select
    ActiveCell.FormulaR1C1 = "min"
    Range("A8").Select
    ActiveCell.FormulaR1C1 = "max"
    Range("A9").Select
    ActiveCell.FormulaR1C1 = "95%ile"

    For statistic = 1 To 12


        Sheet4.Activate
        Sheet4.Select
        tavg = WorksheetFunction.Average(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tsdev = WorksheetFunction.StDev(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tmin = WorksheetFunction.Min(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tmax = WorksheetFunction.Max(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        t95 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.95)
        If ((statistic = 7) Or (statistic = 8)) Then t95 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.05)


        Sheet2.Activate
        Sheet2.Select
        Range(Cells(5, statistic + 1), Cells(5, statistic + 1)).Select
        ActiveCell.Value = tavg
        Range(Cells(6, statistic + 1), Cells(6, statistic + 1)).Select
        ActiveCell.Value = tsdev
        Range(Cells(7, statistic + 1), Cells(7, statistic + 1)).Select
        ActiveCell.Value = tmin
        Range(Cells(8, statistic + 1), Cells(8, statistic + 1)).Select
        ActiveCell.Value = tmax
        Range(Cells(9, statistic + 1), Cells(9, statistic + 1)).Select
        ActiveCell.Value = t95


    Next statistic



End Sub

Public Sub outputpercentiles()

' generates an output page on sheet 2, based on the statistics
'calculated on sheet 4. currently set for 12 statistics, needs
'modifying if more or fewer.
'output page includes the mean,stdev,min,max,and 95%ile of statistics


k = Sheet1.Cells(3, 3)
n = Sheet1.Cells(4, 3)

    Sheet2.Activate
    Sheet2.Select
    Range("A1").Select
    Selection.Font.Bold = True
```

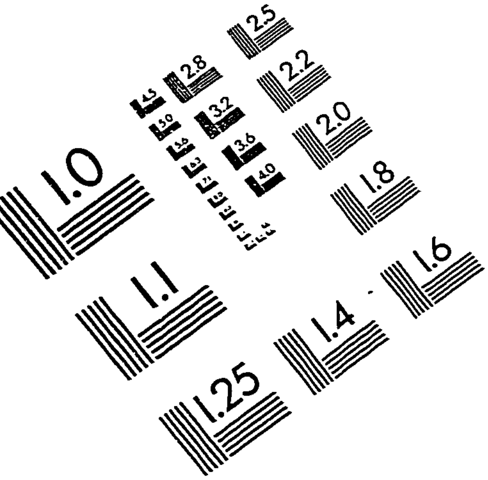
```
ActiveCell.FormulaR1C1 = "Monte Carlo Simulation SAV Output"
Range("A2").Select
ActiveCell.FormulaR1C1 = "Simulation Title:"
Sheet1.Activate
Range("C2").Select
Selection.Copy
Sheet2.Activate
Range("C2").Select
ActiveCell.PasteSpecial (xlPasteValues)
Range("B4").Select
ActiveCell.FormulaR1C1 = "EQ"
Range("C4").Select
ActiveCell.FormulaR1C1 = "CV"
Range("D4").Select
ActiveCell.FormulaR1C1 = "SCV"
Range("E4").Select
ActiveCell.FormulaR1C1 = "X2"
Range("F4").Select
ActiveCell.FormulaR1C1 = "IDA"
Range("G4").Select
ActiveCell.FormulaR1C1 = "IDR"
Range("H4").Select
ActiveCell.FormulaR1C1 = "MIN"
Range("I4").Select
ActiveCell.FormulaR1C1 = "MIN/MED"
Range("J4").Select
ActiveCell.FormulaR1C1 = "IQR"
Range("K4").Select
ActiveCell.FormulaR1C1 = "IQR/MED"
Range("L4").Select
ActiveCell.FormulaR1C1 = "MAX"
Range("M4").Select
ActiveCell.FormulaR1C1 = "MAX/MED"


Range("A5").Select
ActiveCell.FormulaR1C1 = "mean"
Range("A6").Select
ActiveCell.FormulaR1C1 = "stdev"
Range("A7").Select
ActiveCell.FormulaR1C1 = "min"
Range("A8").Select
ActiveCell.FormulaR1C1 = "max"
Range("A9").Select
ActiveCell.FormulaR1C1 = "95%ile"
Range("A10").Select
ActiveCell.FormulaR1C1 = "90%ile"
Range("A11").Select
ActiveCell.FormulaR1C1 = "80%ile"
Range("A12").Select
ActiveCell.FormulaR1C1 = "70%ile"
Range("A13").Select
ActiveCell.FormulaR1C1 = "60%ile"
Range("A14").Select
ActiveCell.FormulaR1C1 = "50%ile"
Range("A15").Select
ActiveCell.FormulaR1C1 = "40%ile"
Range("A16").Select
ActiveCell.FormulaR1C1 = "30%ile"
Range("A17").Select
```

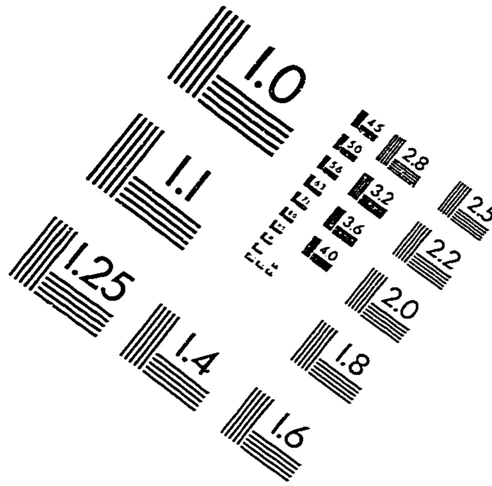
```
    ActiveCell.FormulaR1C1 = "20%ile"
    Range("A18").Select
    ActiveCell.FormulaR1C1 = "10%ile"
    Range("A19").Select
    ActiveCell.FormulaR1C1 = "05%ile"

    For statistic = 1 To 12


        Sheet4.Activate
        Sheet4.Select
        tavg = WorksheetFunction.Average(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tsdev = WorksheetFunction.StDev(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tmin = WorksheetFunction.Min(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        tmax = WorksheetFunction.Max(Range(Cells(1, k + statistic), Cells(n, k + statistic)))
        t95 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.95)
        If ((statistic = 7) Or (statistic = 8)) Then t95 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.05)

        t90 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.9)
        If ((statistic = 7) Or (statistic = 8)) Then t90 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.1)

            t80 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.8)
        If ((statistic = 7) Or (statistic = 8)) Then t80 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.2)

            t70 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.7)
        If ((statistic = 7) Or (statistic = 8)) Then t70 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.3)

            t60 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.6)
        If ((statistic = 7) Or (statistic = 8)) Then t60 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.4)

            t50 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.5)
        If ((statistic = 7) Or (statistic = 8)) Then t50 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.5)

            t40 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.4)
        If ((statistic = 7) Or (statistic = 8)) Then t40 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.6)

            t30 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.3)
        If ((statistic = 7) Or (statistic = 8)) Then t30 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.7)

            t20 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.2)
        If ((statistic = 7) Or (statistic = 8)) Then t20 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.8)

            t10 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.1)
        If ((statistic = 7) Or (statistic = 8)) Then t10 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.9)

            t05 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k + statistic)), 0.05)
        If ((statistic = 7) Or (statistic = 8)) Then t05 = WorksheetFunction.Percentile(Range(Cells(1, k + statistic), Cells(n, k +
statistic)), 0.95)
```

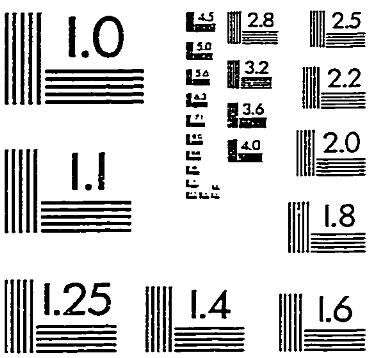
```
    Sheet2.Activate
    Sheet2.Select
    Range(Cells(5, statistic + 1), Cells(5, statistic + 1)).Select
    ActiveCell.Value = tavg
    Range(Cells(6, statistic + 1), Cells(6, statistic + 1)).Select
    ActiveCell.Value = tsdev
    Range(Cells(7, statistic + 1), Cells(7, statistic + 1)).Select
    ActiveCell.Value = tmin
    Range(Cells(8, statistic + 1), Cells(8, statistic + 1)).Select
    ActiveCell.Value = tmax
    Range(Cells(9, statistic + 1), Cells(9, statistic + 1)).Select
    ActiveCell.Value = t95
    Range(Cells(10, statistic + 1), Cells(10, statistic + 1)).Select
    ActiveCell.Value = t90
    Range(Cells(11, statistic + 1), Cells(11, statistic + 1)).Select
    ActiveCell.Value = t80
    Range(Cells(12, statistic + 1), Cells(12, statistic + 1)).Select
    ActiveCell.Value = t70
    Range(Cells(13, statistic + 1), Cells(13, statistic + 1)).Select
    ActiveCell.Value = t60
    Range(Cells(14, statistic + 1), Cells(14, statistic + 1)).Select
    ActiveCell.Value = t50
    Range(Cells(15, statistic + 1), Cells(15, statistic + 1)).Select
    ActiveCell.Value = t40
    Range(Cells(16, statistic + 1), Cells(16, statistic + 1)).Select
    ActiveCell.Value = t30
    Range(Cells(17, statistic + 1), Cells(17, statistic + 1)).Select
    ActiveCell.Value = t20
    Range(Cells(18, statistic + 1), Cells(18, statistic + 1)).Select
    ActiveCell.Value = t10
    Range(Cells(19, statistic + 1), Cells(19, statistic + 1)).Select
    ActiveCell.Value = t05


  Next statistic


End Sub

Sub printresults()

'prints some results pages based on sheets numbered 10 onwards
'even numbered sheets are setup and odd following are output sheets
'the first page of the setup and all output is printed

nsims = Sheet10.Cells(1, 7)

For i = 1 To nsims

  Sheets(8 + (2 * i)).Activate
  With ActiveSheet.PageSetup
    .Orientation = xlLandscape
    .FitToPagesTall = 2
  End With
  Sheets(8 + (2 * i)).Select
  ActiveWindow.SelectedSheets.PrintOut To:=1

  Sheets(9 + (2 * i)).Activate
```

```
    With ActiveSheet.PageSetup
      .Orientation = xlLandscape
      .FitToPagesWide = 1
    End With


    Sheets(9 + (2 * i)).Select

    ActiveWindow.SelectedSheets.PrintOut


Next i



End Sub

Public Sub Varyrates()

'Makes sheets 10 thru 30 with same population, but rates varied
'from all equal (on sheet 10) to all unequal as real (sheet 30)
'sheet 9 row 1 is equal rates
'sheet 9 row 2 is unequal rates
'sheet 9 row 3 is row 2 minus row 1

k = Sheet10.Cells(3, 3)

Sheet10.Cells(1, 7) = 11 'set 11 simulations

Sheet10.Activate
Cells.Select
Selection.Copy

For i = 12 To 30 Step 2

    Worksheets(i).Select
    Range("A1").Select
    ActiveSheet.Paste

Next i

For scaling = 0 To 1 Step 0.1

    Sheets(10 + (20 * scaling)).Activate
    Cells(2, 3).Formula = "Rate Scaling Factor " & scaling

    For region = 1 To k

    Cells(8, region) = Sheet9.Cells(1, region) + (scaling * (Sheet9.Cells(3, region)))

    Next region
Next scaling


End Sub
```

IMAGE EVALUATION TEST TARGET (QA-3)

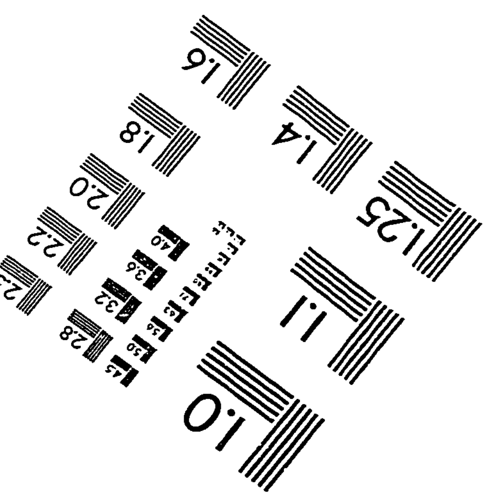


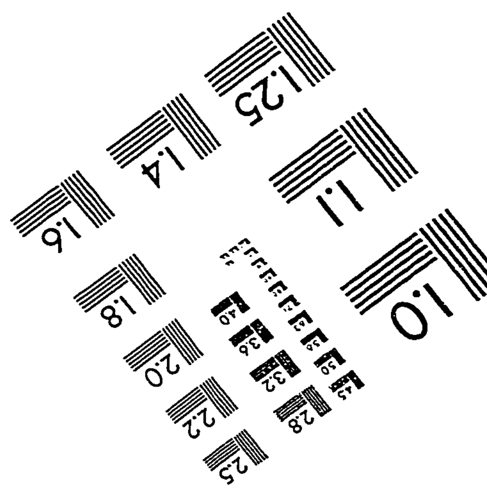




150mm

6″





APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989