# MANAGING THE STATISTICAL CENTER FOR A LARGE NATIONAL STUDY IN THE DEVELOPMENTAL STAGE: THE SENIC PROJECT PART B: STATISTICAL CONSIDERATIONS

Dana Quade[2], Peter A. Lachenbruch[1], Richard H. Shachtman[2], and Robert W. Haley[3]

[1]Department of Preventive Medicine and Environmental Health
University of Iowa

[2]Department of Biostatistics
Univeristy of North Carolina

[3]Center for Disease Control,
U.S. Department of Health, Education, and Welfare,
Atlanta, Georgia

Institute of Statistics Mimeo Series 1131
August 1977

# MANAGING THE STATISTICAL CENTER FOR A LARGE NATIONAL STUDY IN THE DEVELOPMENTAL STAGE: THE SENIC PROJECT*

## PART B: STATISTICAL CONSIDERATIONS

By Dana Quade[2], Peter A. Lachenbruch[1], Richard H. Shachtman[2], Robert W. Haley[3]

July 1977

[1]Department of Preventive Medicine and Environmental Health, College of Medicine, University of Iowa (formerly University of North Carolina)

[2]Department of Biostatistics, School of Public Health, University of North Carolina, Chapel Hill

[3]Center for Disease Control, U.S. Department of Health, Education, and Welfare, Atlanta, Georgia

*This research was supported primarily by Contract No. 200-75-0552, Center for Disease Control, Public Health Service, Department of Health, Education and Welfare.

TABLE OF CONTENTS

# FOREWORD

In a previous paper [2] we were concerned with problems which will arise in the administration of the statistics and data management center of any large research project. Such concern is warranted, since most statistical training is technical rather than administrative, and thus a statistician may need guidance (as we did) when placed in a managerial position. Although this paper was largely based on our experiences with a single project, the Study on the Efficacy of Nosocomial Infection Control (SENIC), we nevertheless believe our recommendations are general in application. Our present paper is more closely tied to specific statistical problems encountered in SENIC. The individual problems considered here, with one exception, involve applications of well-known techniques. But whereas the typical consulting statistician deals with relatively small problems, often one at a time, the statistician manager of a study such as SENIC finds every problem connected with every other so as to constitute essentially one vast problem, and must continually keep in mind the further relationship of statistical solutions to administrative realities. All this makes such studies interesting for the statistician, but also more difficult to manage successfully.

## 1. Introduction

This paper will discuss some of the major statistical decisions made in designing SENIC, and illustrate how these were influenced by various nonstatistical considerations.

The objectives of SENIC are basically three in number:

(I) To determine whether (and if so, to what degree) infection surveillance and control programs (ISCPs) have lowered nosocomial infection rates (NIRs) in major categories of US hospitals[4].

(II) To describe the current status of ISCPs and of NIRs in major categories of US hospitals.

(III) To study in detail the relationships among: a) characteristics of hospitals, b) components of ISCPs, and c) components of changes in NIRs.

In summary, Objective I implies a confirmatory study, with a test of hypothesis; Objective II implies a descriptive study; and Objective III an exploratory study.

The basic design proposed in preliminary planning (i.e., before our Center was formed) was outlined as follows. Select a suitable stratified sample of hospitals, classified into the major categories of interest, including hospitals currently representing various types of ISCPs. Within each hospital determine the NIR for some time $T_1$ before the widespread adoption of modern ISCPs, and also the NIR at the present time ($T_2$), using the technique

---

[4]We measure NIRs in terms of infections per discharge. The question has been raised whether an infection rate per patient-day would not be more appropriate than per discharge. The data are available, so this may be studied in later analyses, but the per discharge rate was adopted because it is the one most commonly used.

of retrospective chart review (RCR). Compare the change in NIR between $T_1$ and $T_2$ for hospitals with active ISCPs and little or no ISCP, in order to satisfy Objective I; describe the ISCPs and NIRs found in the sample at $T_2$, to satisfy Objective II; and explore the data in detail, to satisfy Objective III.

The basic unit of analysis in the SENIC study is a hospital. It was decided that the sampling frame for the main study would consist of the 4194 hospitals in the US excluding Alaska and Hawaii which: provide general medical and surgical care; have median length of stay less than 30 days; are not under administrative control by the federal government, but by state and local governments, religious and charitable organizations, or private for-profit companies; and have at least 50 beds. These restrictions limit the frame to the mainstream of US hospitals which may be presumed reasonably comparable, except for factors taken explicitly into consideration, relative to our objectives. Special ad hoc studies would be required to treat the excluded classes adequately (except that Alaska and Hawaii were omitted simply for administrative convenience, to reduce expenses). It may be noted that although the total number of excluded hospitals is large (about 3000), most of them were excluded for smallness, and together they account for less than 15% of all hospitalizations.

The remainder of this paper is divided into seven sections: Introduction, Available Data, A Stratified Sample of Hospitals, Determining the Sample Size of Patients per Hospital, Choosing $T_1$ and $T_2$, Retrospective Chart Review and the Multiple Read System, Techniques for Selecting Charts, and Summary.

2. Available Data

The following information is available for nearly all US hospitals (including hospitals outside the SENIC sampling frame):

a) AHA. The American Hospital Association collects data mainly from an annual questionnaire sent to each hospital. Emphasis is on matters of particular interest to hospital administrators. The data include full identification of each individual hospital, whether affiliated with a medical school, ownership or control (i.e., type of organization managing hospital and type of service provided to the majority of patients), an inventory of facilities and services, size (number of beds) and utilization (e.g., average census), financial matters (revenue, expenses, and assets), and numbers of personnel of various types.

b) RMP. The Regional Medical Program collected data in 1972. Emphasis was on a more medically-oriented inventory of hospital resources and services available.

In addition to these data, there is:

c) NNIS. The National Nosocomial Infections Study collects information on nosocomial infections on an ongoing basis from an irregularly changing panel of 50 to 100 volunteer hospitals. The sample can in no way be regarded as representative, but it may form the basis for a priori estimates of NIRs.

Since there is absolutely no information available on ISCPs for any broad group of US hospitals, it was early decided to obtain:

d) PSQ. CDC sent a SENIC Preliminary Screening Questionnaire (PSQ) to all US general short-stay hospitals in the spring of 1976. The response rate for hospitals within the SENIC sampling frame was over 80%. Some preliminary

tabulations have been made; final analysis will be completed in 1977. The questionnaire is exclusively concerned with hospital programs for surveillance and control of nosocomial infections. Two important summary variables which have been developed from it are: (i) an overall index of extent and intensity of programs for surveillance of nosocomial infections and (ii) an overall index of extent and intensity of programs for control of nosocomial infections.

The following will be collected from sample hospitals:

e) HIS. A Hospital Interview Survey will be conducted by CDC interviewers in each sample hospital using forms developed in collaboration with the Institute for Social Sciences Research at UCLA. Those to be interviewed include the hospital administrator, the chairman of the infection control committee, the infection control nurse, a sample of RNs and LPNs, and other hospital personnel whose positions may include an important involvement in controlling infections. This will provide detailed information on infection surveillance and control programs.

f) MRS. Within each hospital a Medical Records Survey will be conducted, sampling records of discharges from time periods $T_1$ and $T_2$. The information to be abstracted from each record includes demographic characteristics of the patient, service, diagnosis, therapeutic procedures such as catheterization, antibiotics used, results of all cultures performed, and other medical data relevant to the diagnosis of infection. In particular, we will determine whether an infection was present. If so, we will further determine whether it was hospital- or community-acquired, and its site. The essential summary variable to be calculated for each hospital is "change in NIR from $T_1$ to $T_2$".

3. A Stratified Sample of Hospitals

The classifications which were primarily considered for use in stratifying hospitals are as follows:

a) Size of hospital (number of beds)

Stratification on size was needed because separate conclusions and reports were wanted for different size categories. Furthermore, both NIRs and ISCPs were believed to vary significantly with size. In general, larger hospitals tend to have a larger proportion of patients at higher risk of nosocomial infection, hence, higher NIRs and a need for more elaborate ISCPs. In addition, the greater administrative and logistical complexity of larger hospitals renders infection control activities more difficult to carry out. Thus, stratification here would help control variability also.

b) Affiliation (whether affiliated with a medical school or not)

Similar considerations, especially variability, suggested stratification on affiliation, but the need for separate conclusions and reports was less.

c) Quality of ISCP:

(i) Surveillance

(ii) Control

Stratification on quality of ISCP was necessary in order to ensure sufficiently many hospitals with good ISCPs to perform adequately the analyses required for Objectives I and III, particularly since it was suspected that only the best ISCPs - perhaps no more than the upper 5% - would be good enough to have significant influence upon NIRs.

Since the two indexes of Surveillance and Control derived from the PSQ had no established or natural breakpoints, we divided each of them at the 20th, 50th, and 80th percentiles, thus producing four categories: Lower 20% (LO),

next 30% (ML), next 30% (MH), and upper 20% (HI). Table 1 shows a cross tabulation of the four variables Size, Affiliation, Surveillance Index, and Control Index for the 3515 respondents to the PSQ which fall within the SENIC sampling frame[5]. In this table, hospitals with fewer than 200 beds are not classified as to affiliation (Y = yes, N = no); very few of them are affiliated with any medical school. Rearrangement of the row totals in Table 1 shows a strong positive relationship between the Surveillance and Control Indexes: very few hospitals are simultaneously low on one index and high on the other. Also, hospitals affiliated with a medical school are rarely low on either index.

A major question was whether the number of hospitals from each cell of the stratification plan should be equal, or proportional to the number in the universe, or whether some compromise between these extremes is preferable. (The answer should depend on the relative variances of changes in NIR between hospitals in the different cells, but we had absolutely no information and almost no intuition regarding this point, and faute de mieux we decided to proceed as if the variances were equal.) Thus we considered the following three alternatives:

| | |
|---|---|
| "Equal" | take the same number of hospitals from each cell sampled (except, of course, if the equal number is greater than the total in the universe for some cell, take "all"). |
| "Compromise" | take the same number of hospitals from each ISCP combination within each size/affiliation combination, but sample the size/affiliation combinations approximately proportionally to their numbers in the universe. |
| "Proportional" | for each cell in the design, take a sample number of hospitals which is approximately proportional to the total number in the universe. |

Proportional allocation has considerable advantages for a descriptive

[5]As of the time of the stratification analysis

TABLE 1

Hospitals within the SENIC Sampling Frame

by Size, Medical School Affiliation, and Surveillance and Control Indices

| SIZE (BEDS) | | 50-74 | 75-99 | 100-149 | 150-199 | 200-299 | | 300-499 | | 500+ | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AFFILIATION | | all | all | all | all | Y | N | Y | N | Y | N | all |
| SURV. | CONTROL | | | | | | | | | | | |
| LO | LO | 113 | 72 | 57 | 29 | 2 | 23 | 3 | 11 | 2 | 1 | 313 |
| LO | ML | 76 | 57 | 56 | 30 | 3 | 26 | 3 | 8 | 7 | 2 | 268 |
| LO | MH | 28 | 19 | 15 | 14 | 0 | 10 | 2 | 2 | 3 | 0 | 93 |
| LO | HI | 8 | 3 | 7 | 1 | 2 | 3 | 3 | 1 | 1 | 0 | 29 |
| ML | LO | 55 | 60 | 55 | 46 | 2 | 30 | 7 | 19 | 0 | 2 | 276 |
| ML | ML | 83 | 76 | 76 | 52 | 3 | 46 | 18 | 44 | 7 | 7 | 412 |
| ML | MH | 45 | 33 | 52 | 24 | 7 | 43 | 23 | 31 | 19 | 9 | 286 |
| ML | HI | 11 | 9 | 6 | 13 | 3 | 8 | 10 | 9 | 11 | 0 | 80 |
| MH | LO | 24 | 22 | 16 | 9 | 1 | 7 | 0 | 4 | 2 | 2 | 87 |
| MH | ML | 44 | 47 | 45 | 19 | 5 | 33 | 18 | 29 | 15 | 5 | 260 |
| MH | MH | 34 | 44 | 77 | 51 | 14 | 82 | 36 | 47 | 38 | 10 | 433 |
| MH | HI | 21 | 20 | 38 | 42 | 8 | 57 | 30 | 26 | 30 | 3 | 275 |
| HI | LO | 7 | 5 | 10 | 1 | 0 | 3 | 1 | 0 | 0 | 0 | 27 |
| HI | ML | 27 | 23 | 19 | 11 | 2 | 14 | 4 | 9 | 3 | 2 | 114 |
| HI | MH | 32 | 30 | 39 | 34 | 3 | 37 | 26 | 20 | 17 | 5 | 243 |
| HI | HI | 41 | 28 | 51 | 40 | 13 | 55 | 30 | 30 | 30 | 1 | 319 |
| Total Responding to PSQ | | 649 | 548 | 619 | 416 | 68 | 477 | 214 | 290 | 185 | 49 | 3515 |
| Total in universe | | 827 | 663 | 759 | 504 | 79 | 549 | 239 | 322 | 198 | 54 | 4194 |

analysis such as is required by Objective II, in that it leads to improved efficiency assuming equal variances, and in that it allows simpler calculations because the proportional subsamples are self-weighting. On the other hand, equal allocation will yield more power for a test of hypothesis such as is required by Objective I, and indeed at least some oversampling of good ISCPs is necessary in order to make this test even possible.

When a panel of three outside experts in sampling was convened at CDC to advise SENIC, a tentative plan was produced as follows. In Table 1, ignoring medical school affiliation, there are 7 size x 4 surveillance x 4 control = 112 cells. From each cell sample 2 hospitals, except that from those with combinations LO-LO and HI-HI on Surveillance and Control (the first and last rows) sample 4 hospitals, and from those in the HI-LO and LO-HI rows (which have very few hospitals) sample only 1. This produces a total of 272 hospitals in the sample. It is essentially an "equal" allocation as defined in the preceding paragraph. However, because of the way the stratification on size has been arranged, it is also nearly the "compromise", except for oversampling of the largest hospitals; this is because the numbers of hospitals in the various categories, except for the one above 500 beds, are not far from equal. One additional complication was suggested: for the size categories above 200 beds, subsample each cell as closely as possible according to the distribution with respect to medical school affiliation.

Further study led us to propose a modified plan. In particular, one change was suggested by the analyses detailed in Section 4. It now appeared that we should sample fewer patients per hospital than had been contemplated up to then (early plans had supposed 1000 patients per time period per hospital) and instead should take more hospitals. We thus proposed to sample 3 hospitals per cell rather than 2. In those cells where the tentative plan proposed 4 hospitals,

we proposed 6, and in those where 1 had been proposed, we proposed 2. This gave a total of 364 hospitals, except that, based on the 3515 responses to the PSQ, it appeared that in a few cells there were not quite as many hospitals in the universe as the plan called for: Table 2 shows the resulting plan, with 354 hospitals.

TABLE 2

Proposed Sampling Plan

Size (beds) Categories

| SURV. | CONTROL | 50-74 | 75-99 | 100-149 | 150-199 | 200-299 | | 300-499 | | 500+ | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Y | N | Y | N | Y | N | |
| LO | LO | 6 | 6 | 6 | 6 | 1 | 5 | 1 | 5 | 2* | 1* | 39 |
| LO | ML | 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 | 3 | 0 | 21 |
| LO | MH | 3 | 3 | 3 | 3 | 0 | 3 | 2 | 1 | 3 | 0 | 21 |
| LO | HI | 2 | 2 | 2 | 1* | 1 | 1 | 2 | 0 | 1* | 0* | 12 |
| ML | LO | 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 | 0* | 2* | 20 |
| ML | ML | 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 | 2 | 1 | 21 |
| ML | MH | 3 | 3 | 3 | 3 | 1 | 2 | 1 | 2 | 2 | 1 | 21 |
| ML | HI | 3 | 3 | 3 | 3 | 1 | 2 | 2 | 1 | 3 | 0 | 21 |
| MH | LO | 3 | 3 | 3 | 3 | 0 | 3 | 0 | 3 | 2 | 1 | 21 |
| MH | ML | 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 | 2 | 1 | 21 |
| MH | MH | 3 | 3 | 3 | 3 | 1 | 2 | 1 | 2 | 3 | 0 | 21 |
| MH | HI | 3 | 3 | 3 | 3 | 0 | 3 | 2 | 1 | 3 | 0 | 21 |
| HI | LO | 2 | 2 | 2 | 1* | 0 | 2 | 1* | 0* | 0* | 0* | 10 |
| HI | ML | 3 | 3 | 3 | 3 | 0 | 3 | 1 | 2 | 2 | 1 | 21 |
| HI | MH | 3 | 3 | 3 | 3 | 0 | 3 | 2 | 1 | 2 | 1 | 21 |
| HI | HI | 6 | 6 | 6 | 6 | 1 | 5 | 3 | 3 | 6 | 0 | 42 |
| Sub-total | | 52 | 52 | 52 | 50 | 6 | 46 | 22 | 29 | 36 | 9 | 354 |
| TOTAL | | 52 | 52 | 52 | 50 | 52 | | 51 | | 45 | | 354 |

*Universe contains only as many hospitals as shown here, based on the responses to the PSQ.

4. Determining the Sample Size of Patients Per Hospital

In this section we discuss sample size determination. It was decided to increase the concentration of infections by excluding certain categories of patients who are at relatively low risk of nosocomial infection, including those on the psychiatric, pediatric, and obstetric services (except Caesarean sections). Also, newborns were excluded as representing a group requiring special study.

For any given stratum, let

$M$ = number of hospitals in the population

$m$ = number of hospitals in the sample

$N$ = number of patients per hospital per time period in the population (i.e., number of discharges per annum)

$n$ = number of patients per hospital per time period in the sample

$S_1^2$ = variance in change in NIR between hospitals within stratum

$S_2^2$ = variance in change in NIR within hospital (on a per-patient basis)

$C_1$ = marginal cost of adding a hospital to the sample

$C_2$ = marginal cost of adding a patient record to the sample

Then the cost of sampling the stratum is

$$C = mC_1 + mnC_2$$

and the variance of the estimated average change in NIR within the stratum, following Cochran [1] is

$$V = \frac{1}{m}(S_1^2 - \frac{S_2^2}{N}) + \frac{1}{mn} S_2^2 - \frac{1}{M}S_1^2 .$$

The marginal cost of an additional hospital was estimated as:

| | |
|---|---|
| Recruitment and enrollment | $ 300 |
| Conducting HIS | 750 |
| Travel for 14 field workers for MRS | 1400 |
| Total | $2450 = $C_1$ |

The marginal cost per patient was estimated to include a field cost per form of about $8, and a cost for printing, data processing, and data analysis of about $2, giving a total of $10 per form. Then, with 1.3 forms per chart (see Section 6), and two time periods, the marginal cost for adding one patient in each time period is $10 x 1.3 x 2 = $26 = $C_2$. The cost ratio $C_1/C_2$ is thus about 100.

To estimate the between-hospital within-stratum variance in the change in NIR, consider two extreme examples: (1) if half the hospitals experience no change ($p_2 = p_1$) and the other half a drop of .05 ($p_2 = p_1 - .05$), then $S_1^2 = .000625$; (2) if the changes within the stratum are uniformly spread between 0 and .05, then $S_1^2 = .00020833$. We therefore supposed that for most strata, $.0002 < S_1^2 < .0006$. The within-hospital variance per observation, given perfect chart reviewers, would be $S_2^2 = p_1(1-p_1) + p_2(1-p_2)$. Extreme situations might be:

(1) if the initial NIR is high and does not change, e.g.

$p_1 = p_2 = .1$, then $S_2^2 = .18$;

(2) if initial NIR is low and drops 40%, e.g.

$p_1 = .05$ and $p_2 = .03$, then $S_2^2 = .0766$.

Allowing for a possible doubling because of sensitivity and specificity errors, we supposed that, for most hospitals, $.07 < S_2^2 < .36$. Since the combination ($S_1^2$ high, $S_2^2$ low) is unlikely, it appeared that the variance ratio $S_1^2/S_2^2$ would lie between .0006 and .0028.

The number of discharges, N, depends on the size of the hospital. The smallest hospitals eligible for the sample, with just over 50 beds, may have some 2000 discharges in a year; the largest hospitals have more than 30,000 discharges, a population size which is effectively infinite.

The following formula gives the value of n which both minimizes the variance for a fixed cost and also minimizes the cost for a fixed variance:

$$n^* = \sqrt{\frac{C_1/C_2}{S_1/S_2 - 1/N}} \quad .$$

Suppose the allowable cost per stratum is fixed at C; then it follows that the number of hospitals to be sampled must be

$$m^* = \frac{C}{C_1 + nC_2} \quad .$$

We were able to propose a value of C to be used from the following considerations. For many months all discussion of the total size of the MRS had assumed there would be about 500,000 review forms in all, and, as noted above, the cost per form is about \$10, giving \$5,000,000. In addition, it had been tentatively decided (see Section 3) to sample 272 hospitals (at \$2450 each) adding \$666,400 for a total "variable cost" (not counting fixed costs for pilot studies, development of computer programs, etc., which do not change with sample size) of HIS and MRS of \$5,666,400. Assuming there are 112 strata (see Section 3) gives the cost per stratum as C = \$50,000.

The optimal n and the corresponding variance could then be calculated from the preceding formulae. Using the extreme values of $S_1^2/S_2^2$ and N yields n* ranging in value from about 200 to 400 (eliminating the case $S_1^2/S_2^2 = .00055$ and N = 2000, since $S_1^2/S_2^2 = .00055$ assumes p = .1,

which is unlikely for small hospitals). Noting that the number of hospitals sampled must be an integer, we find that the optimal design ranges from 7 hospitals per stratum with 200 patients per hospital per time period to 4 hospitals per stratum with 400 patients per hospital per time period.

Another consideration for sample size determination is that a report is to be made to each sample hospital of its own NIR. This is in order to help obtain participation in the study. If there is to be any hope of having meaningful results for individual hospitals, the number of patients sampled must be several hundred, probably not much less than 500. But 500 happens to be administratively convenient, since it was estimated that a chart reviewing team should be able to complete the reading of about 1300 forms - 500 per time period, with 30% rereading (See Section 6)- in a week's time. If n were to be increased beyond 500, either the team would have to stay longer and so incur weekend per diem costs, or the size of the team would have to increase, perhaps leading to inefficiency due to crowding and administrative problems; this also might put an undue burden on the sample hospitals. Thus we suggested a design of n= 500 patients per time period per hospital and m = 3 hospitals per stratum. This yields a total variable cost for the study of $5,623,800, essentially the same as calculated earlier. The per stratum variance of the estimated drop in NIR for the proposed design is increased from 10% to at most 35% as compared with a design using the optimal n*.

We considered the possibility of taking different sample sizes from different hospitals. If one could make a good initial estimate of the NIR, one might adjust the sample size accordingly. Thus if absolute errors in estimating NIR were more important, one would take more charts where a high NIR was expected; if the percentage errors were more important, one would take more where a low

NIR was expected. Such a prediction might be based very simply, but with little precision, on the size of the hospital. However, even if a good estimate were not possible, one might adjust the sample size in accordance with the results actually found in the first few charts reviewed. For example, if it were desired to have larger samples from hospitals with low rates, a "double-sampling" strategy might be used: read an initial sample of fixed size (say 400), and then add a smaller supplementary sample (say 200) if the rate found in the first sample is small. A more extreme "inverse sampling" strategy would require sampling to continue until a prescribed number of infected patients is found. Such a strategy makes the number of charts read be inversely proportional to the NIR. Strategies which allow for variable sample sizes have obvious disadvantages, however. They add to the administrative complexity, and hence also the cost, of both field work and data processing. For example, teams of chart readers must be sent to hospitals without prior information of how many charts they will read, and hence how long they must remain. Furthermore, the advantage which might be gained is again limited to within-hospital variance, which makes up considerably less than half, perhaps only 20%, of total variance. For these reasons we recommended that a single sample size be used for all hospitals in the MRS.

It has been assumed so far that the same number of patients would be sampled in each time period. Statistical considerations suggested that a departure from this might be desirable. For example, consider sampling 400 patients from $T_1$, and 600 from $T_2$. The cost would remain the same, or perhaps even drop slightly, as patient records from time $T_1$ might be more costly to read (for example, they are more likely to be on microfilm). While the variance of the estimated change in NIR would increase slightly, this would be outweighed by

the improvement in variance of the estimate of the NIR for $T_2$. This in turn would improve the results for Objective II. For example, for $p_1 = .1$ and $p_2 = .08$, the increase in within-hospital variance for estimated change in NIR is only 7%, while the decrease in the estimated variance of the NIR at $T_2$ is 17%. When one takes into consideration the between-hospital variance, which is essentially unaffected by the relative sample sizes at $T_1$ and $T_2$, this advantage becomes relatively minor for the study as a whole; but it enhances the value of the report to be prepared for each individual MRS hospital. Finally, however, it was decided to sample 500 patients from each time period. In addition to the above, this was partly because Objective I was determined to be overriding and partly because this plan would be more easily administered and more readily accepted by less statistically sophisticated reviewers.

5. Choosing $T_1$ and $T_2$

The time periods $T_1$ and $T_2$ were fixed to be one year in length, so as to minimize the effect of any seasonal variation on the occurrence of "mini-epidemics" of nosocomial infection in sample hospitals and to assure a population base large enough to provide a sufficient number of patients even in the smallest hospitals.

The year $T_1$, our "before" year, must of course be chosen before the widespread adoption of modern ISCPs, but not so far back that the patient records would be unavailable or not comparable to $T_2$ records with respect to methods of charting and quality of care. A careful study of the PSQ data suggested that 1971 was the first year when an appreciable number of hospitals began to adopt ISCPs. We felt sure that no hospital would have destroyed such recent records, although we had no actual data, and hence chose $T_1$ tentatively to be 1970. Upon investigation, however, we discovered that many hospitals omit the nurses' notes when microfilming their older records. These notes are apparently seen as superfluous to the permanent record by the hospitals, but they are absolutely essential for RCR, and it seemed that they would be lacking for 10 to 15% of hospitals. (This figure includes some hospitals whose records were unavailable for other reasons, such as flood damage.) There was also the problem caused by the small but still significant number of hospitals which had initiated at least partial ISCPs as early as 1970.

It was therefore decided to allow different years to be used for $T_1$ in different hospitals, as follows. For any sample hospital which has 1970 records available and which had not yet implemented an ISCP, $T_1$ will be 1970. Otherwise, we consider 1971, then 1969, then 1972, then 1968, choosing the first of these years which meets the two criteria (i) records available and (ii) ISCP not yet initiated. If for any sample hospital there is no year in

the period from 1968 to 1972 which meets the two criteria, then it will be excluded from the sample and a hospital chosen randomly from the same stratum substituted. The decision to limit variation from 1970 to at most two years was made on an intuitive basis; when further results are available it will be reviewed.

With varying $T_1$ as indicated we will have data for five different years. Since it is possible that a secular trend in NIRs was operating over that period, we will want to take account of it. Our policy will tend to produce a bias toward earlier $T_1$ in hospitals which eventually adopted better ISCPs (and hospitals which still have no ISCPs will certainly not have $T_1$ any earlier than 1970), and the limited evidence we have suggests that NIRs in hospitals without ISCPs have been rising. so that such a bias would tend to reduce the difference between hospitals with and without ISCPs in change in NIR from $T_1$ to $T_2$.

If the difference is still significant, so much the better. On the other hand, we may sharpen the test by adjusting for the secular trend in NIRs using an analysis of covariance in which year is the covariable. Of course we must first determine whether there is any relationship between the availability of records, the adoption of ISCPs, and other variables of interest. This is being done using the AHA and PSQ data.

The year $T_2$ must also be chosen with care, and two tentative decisions have already been changed. As a first thought one might want to choose the year immediately preceding the gathering of the detailed HIS data, which are based on interviews and therefore subject to telescoping effects and memory lapses. This might suggest 1976 as a good choice. But one theme which recurred repeatedly during preliminary interviewing was that the receipt in March 1976 of the PSQ in hospitals, coupled with rumors of the adoption of

tougher standards for accreditation of hospitals (which indeed took place later in the year), triggered the institution of various infection control measures in hospital after hospital. It was the classical case of the investigation of a phenomenon itself influencing that phenomenon. Thus $T_2$ has been pushed back and fixed as the 12 month period ending 31 March 1976.

6. Retrospective Chart Review and the Multiple Read System

As mentioned above, it had been decided early on to carry out a retrospective study, establishing infection rates from patient records. Although a prospective study of ISCPs might seem preferable, it was found not feasible, for several reasons: there are ethical questions involved in any attempt at manipulation of ISCPs, since surveillance activities definitely benefit some individual patients regardless of the overall influence on infection rates under study here; and statistical questions concerning the fact that the study could not be conducted "blind," and that the data-gathering process in itself would very likely influence infection rates, especially in control hospitals. Beyond these, it became clear that the cost of a prospective study large enough to satisfy the objectives would be prohibitive.

But patient records do not regularly include explicit notation of hospital-acquired infections, which can generally be determined only by detective work based upon clues such as fevers recorded or antibiotics prescribed which may appear unrelated to the patient's major diagnosis. Pilot studies in four hospitals were conducted in which all patients admitted during defined 10 week periods were followed prospectively by special nurses assigned to the wards by CDC, and diagnoses of nosocomial infections were made by a physician expert in the field. The prospective data collection team worked unobtrusively to avoid influencing the building of the medical records. Then the RCR technique was applied to the records of these same patients after their discharge, to see whether it reached the same conclusions. Each record was reviewed by two different chart readers acting independently. (It is conceded that prospective surveillance of patients might influence infection rates - this was a major argument against conducting a prospective study - but the question of

concern here is whether it influences the detectability of infections from patient records, when conducted for only 10 weeks, and we concluded that any such effect would be negligible.) Results from these pilot studies might be used to correct estimates of NIRs in hospitals where RCR was the only technique used.

A more theoretical look at the RCR technique is as follows: A chart reviewer may make errors which lead to the improper classification of patients. More precisely, let $u$ be the probability that, given the chart of a patient who had a nosocomial infection, the reviewer will recognize that infection (or record the necessary information for recognition by the computer); then $u$ is called the sensitivity of chart review. Let $v$ be the probability that, given the chart of a patient who did not have a nosocomial infection, the reviewer will not record a spurious infection; $v$ is called the specificity. With perfect chart reviewers, $u$ and $v$ would both equal 1.

Preliminary analysis of the data from the pilot studies described above suggests that sensitivity in the MRS may be 80% if it is defined in terms of the reviewer's diagnosis, or 90% if in terms of a computer's diagnosis based on the detailed information he records; specificity appears likely to be above 95%, perhaps as high as 98% for computer-made decisions. However, even with such apparently great accuracy serious biases can occur. For example, if the true NIR, i. e. the probability of a patient having a nosocomial infection, is $p = .05$ (this is quite reasonable; it is the estimated NIR for U.S. hospitals overall), and if $u = .9$ and $v = .98$, then the probability of a patient being declared to have a nosocomial infection is

$$pu + (1-p)(1-v) = (.05)(.9) + (.95)(.02) = .064$$

which overestimates the NIR by 28%. This is optimistic; with $u = .8$

and v = .95 we get .0875, or a 75% overestimate.

To overcome errors of this type, we devised a system of multiply reading charts. After a chart has been reviewed for the first time, a second reader is assigned to give an independent review, if the first reader recorded any infection (hospital- or community-acquired), or with 5% probability even if he did not. The two readers' statements as to site and type of infection are then compared, and if there is any discrepancy, a third reader is assigned also. On those rare occasions when no two of three readers produce consistent findings, the field supervisor calls them into a conference at which they compare notes and must arrive at a consensus. This system has the advantage of concentrating the re-reading on the more difficult charts, those which refer to infected patients; in addition, the conferences serve a direct quality control purpose in making the chart readers in the field aware of errors they have just committed.

After the multiple-read system as just described had been used in the fifth pilot study hospital, certain revisions were suggested. A problem had appeared in that the re-readers soon came to realize that their charts were nearly all difficult. This awareness seriously affected chart reviewer morale and was suspected of introducing bias into the second reads. Also, it was questioned whether conferences should be postponed until after three independent readings. In the final pilot study, under way as of this writing, a conference is being called immediately whenever a discrepancy is found between two readings of a chart. Furthermore, the reviewing process has been reorganized so that a reader does not know whether he is rendering the first or the second review. This not only reduces the morale and bias problems, but also eliminates the necessity for re-reading a sample of charts on which no infection was found at first reading. The 5% re-read of negatives

would have picked up almost no missed infections because of their low frequency among negatives. Rereading charts classified by first reading as having community-acquired infections gives a more sensitive and efficient check on missing nosocomial infections.

Using estimates of sensitivity and specificity based partly on results from the pilot studies conducted so far and partly on the best current opinion at CDC as to what can be achieved with improved procedures, we have concluded that the result of applying the multiple read system will be to achieve a variance in estimating the NIR of a hospital which is not more than twice what could be done with a single reading by a "perfect" chart reviewer, at an average cost of approximately 1.3 readings per chart.

7. Techniques for Selecting Charts

Another area requiring a decision involving both statistical and non-statistical consideration was the question of how to select which charts would be read in a sample hospital. We have already indicated that 500 discharges would be required from each of $T_1$ and $T_2$. Not all hospitals can provide convenient frames for sampling discharges as required. It is now common for hospitals to have computerized record keeping, which is ideal for our purposes, but in few cases are the cards or tapes available for purposes of sampling. Information as to service is not always included. In many instances the hospital can provide only an admissions list; it was early decided that this would be an acceptable substitute for discharges.

A uniform procedure adopted for selecting patient samples is as follows. Each hospital which participates in MRS will be required to provide CDC with discharge lists for the two time periods $T_1$ and $T_2$, or admissions lists if discharge lists are not available. These lists must include the name of each patient, the date of his discharge (including death) or admission, his hospital chart number, and any other information necessary to request his chart from the record room. If possible, the lists should also include other information useful for identification of a patient or for determining his eligibility for the study, such as surgical operations, diagnoses, and especially service. Lists should be in machine-readable format, e.g., punch cards or magnetic tape, where available; otherwise computer listings or photocopies of hand-written listings will have to be accepted. In some instances hospitals can furnish only the beginning and ending numbers of the admissions list for a year, and we must generate the remaining numbers and make exclusions on site.

Assuming a given list is not in machine-readable format, a preliminary inspection will be made to determine how many nonexcludable names it contains. A "nonexcludable" name is one which cannot be excluded from the sample (for wrong service or whatever reason) on the basis of the information in the list. A rough extimate will be made of how many eligible names there are among the nonexcludables: e.g., if the list contains little information to permit exclusions, then presumably a high proportion of its nonexcludables will be found ineligible upon actual inspection of their charts. If the number of nonexcludables is small enough, they will all be keyed into the computer. Otherwise, a sample will be chosen for keying, large enough so that the total number keyed is estimated to include at least 2000 eligible names. This will be done by taking a systematic sample of the days or pages of the list.

The computer, starting from the entire list if it was machine-readable, or otherwise from whatever portion was keyed in, will choose random samples of 500 names for $T_1$ and $T_2$, and print them out as sampling lists. These will be in numerical or terminal digit order for the hospital's convenience in pulling the charts, although this order may be related to the occurrence of infection in some unknown manner. Because charts may not be locatable or some patients on the list may be ineligible, supplementary random samples will also be needed. These must be in random order, in spite of the inconvenience for chart pulling, so that the randomness of the total sample is still assured when one name on the list is taken after the other until the quota of 500 is obtained. (An alternative scheme considered was to provide sorted supplementary batches of names, with the stipulation that the entire batch must be taken even if only one name from the batch is required to make up the quota.) Since we have as yet no national data on the proportion of hospital patients who are on the excluded services or whose records will be

unlocatable, it is difficult to judge how many names are required on the supplementary sampling lists.

## 8. Summary

A variety of statistical questions have been discussed here in. The most important lesson to be learned from them is that practical necessities always influence and sometimes override statistical niceties. In SENIC the decision to do a retrospective study was based on conceptual as well as ethical requirements, the method of obtaining samples of patients had to be tailored to allow for the vagaries of the practical world, and the procedures used in multiply reading charts were responses to practical considerations. A second lesson is that the form of the study will evolve. One cannot anticipate all of the problems (in the subject matter, in operations, or in statistics) in the early stages of a study. This of course is no reason to ignore any problem area, as some may affect the final data analysis in an uncorrectable adverse way. But be prepared to rethink every decision and to redo every analysis.

## ACKNOWLEDGMENTS

We would like to acknowledge stimulating discussions with John Bennett, Howard Freeman, James Grizzle, and Edward Wagner. The SENIC Steering Committee at the Center for Disease Control has provided initiative, enthusiasm, and continuing scientific direction. The stratification/ sampling plan was reviewed by Daniel Horvitz, Ray Jessen, and Seymour Sudman, who made a number of helpful suggestions. We also owe a debt of gratitude to Fredrick Whaley and Donna McClish for assistance throughout.

## REFERENCES

[1] Cochran, W. G.: Sampling Techniques, Second Edition. John Wiley & Sons, 1963.

[2] Lachenbruch, P. A., Shachtman, R. H., and Quade, D.: "Managing the Statistical Center for a Large National Study in the Developmental Stage: The SENIC Project, Part A: Managerial Considerations". Institute of Statistics Mimeo Series No. 1130, University of North Carolina, 1977.