BIOMATHEMATICS TRAINING PROGRAM

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

by

Peter A. Lachenbruch
Department of Preventive Medicine and Environmental Health
University of Iowa

Richard H. Shachtman and Dana Quade
Department of Biostatistics
University of North Carolina

Institute of Statistics Mimeo Series No. 1130
August 1977

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

PART A: MANAGERIAL CONSIDERATIONS

By Peter A. Lachenbruch[1], Richard H. Shachtman[2], and Dana Quade[2]

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

*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 this paper we are 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 is 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. A companion paper [6] will be more closely tied to specific statistical problems encountered in SENIC. The individual problems considered there, 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 what we have learned during the first two years of SENIC. We function as a statistics and data management center. Our Center's role in the project includes consultation on all aspects of the study, particularly on its design, and also management and statistical analysis of the data. A word about terminology: we refer to the statistician in charge of such a center, assuming he[3] is a statistician (as we believe he should be), as the statistician manager. Reporting to him are the chief statistician, the chief programmer, the chief data manager[4], and an administrative assistant; he in turn reports to the funding agency's project officer, who has both scientific and administrative responsibility for the entire study, including data collection (See Figure 1). The authors of this paper have played the roles of statistician manager (P.A.L. first, and R.H.S. later), chief statistician (D.Q.), and project officer (R.W.H.)

The SENIC Project is designed to measure the magnitude of the nosocomial (hospital-acquired) infection problem in the U.S., to take inventory of the programs currently in effect which are aimed at reducing the problem, to evaluate the efficacy of such programs over the recent past, and to pinpoint reasons for their success or failure. Such goals are to be achieved by a nationwide survey of hospitals in which nosocomial infections will be ascertained by a procedure called retrospective chart review (RCR). This involves application of a complex diagnostic algorithm, further developed and tuned on pilot studies,

---

[3] The pronoun "he" in this paper will be construed to include both masculine and feminine.

[4] This role may be combined with the preceding one.

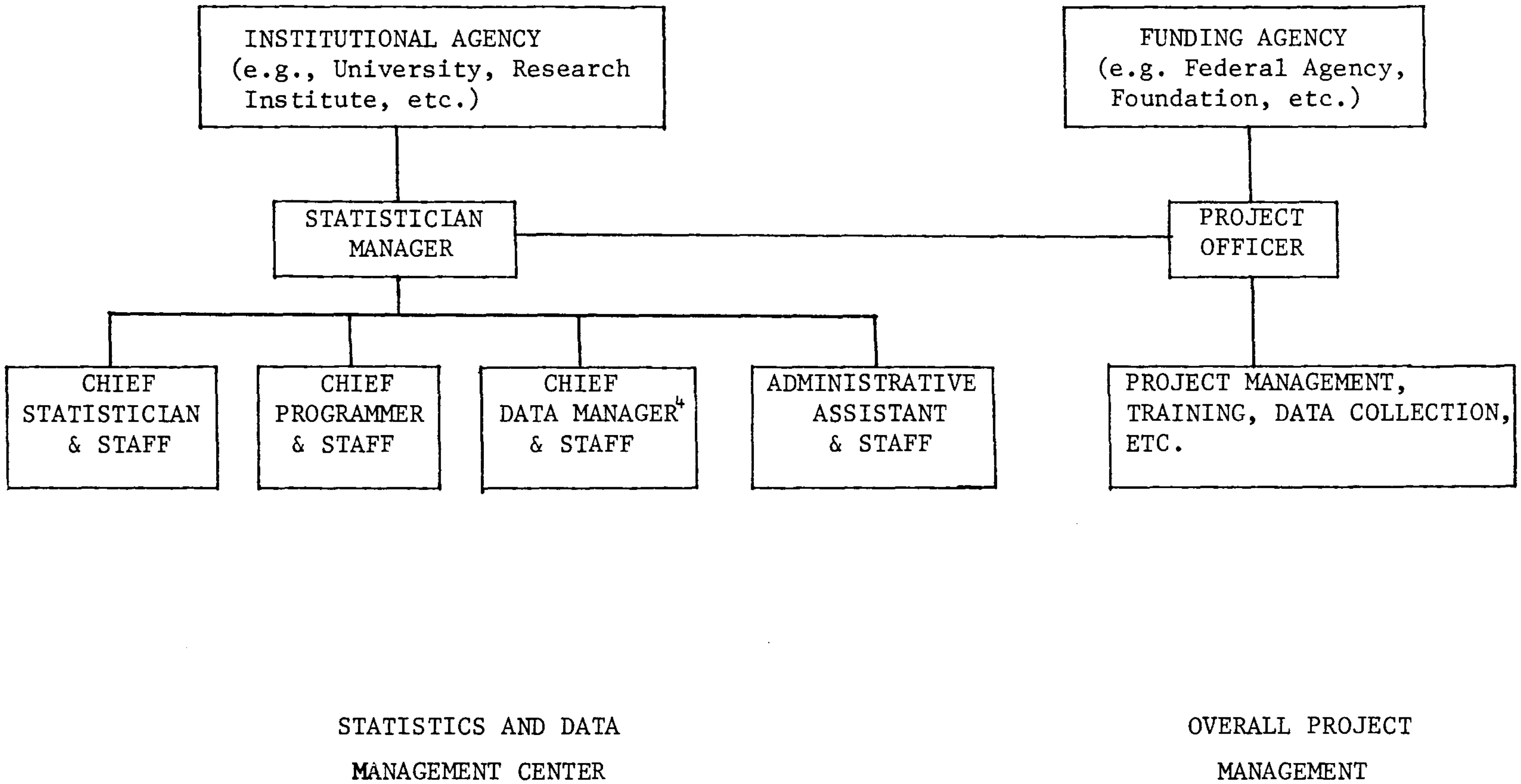

FIGURE 1: ORGANIZATIONAL LINES OF RESPONSIBILITY

to data abstracted from patient records. An important requirement of the study is to obtain timely results on which the Center for Disease Control (CDC), the sponsoring agency, may base recommendations for health care policy. (A more statistically oriented description of SENIC will be given in [6]).

The dimensions of the project are staggering. Initial analysis suggested that for accurate estimates of infection rates we would need about 2000 patients per hospital and 200 hospitals, stratified according to size, infection control practices, and medical school affiliation (a surrogate measure of the proportion of patients at high risk of infection). The sampling plan has since been refined (see [6]). The current plan is to sample about 350 hospitals from a highly stratified frame. From each of these hospitals approximately 1000 discharges will be reviewed, some of which will be multiply read to ensure a satisfactory level of accuracy. To perform the chart reviewing process we have estimated that about 250 chart reviewers are needed over a period of some 18 months. Counting pilot studies, and multiple readings, some 500,000 forms will be received at the Center for processing, to include organizing, editing, converting to machine-readable format, and archiving. Computer processing, including application of the diagnostic algorithm to each form, will be a vast undertaking, and statistical analysis will not be completed for several years. The Center staff numbered about 50 at its peak, and at one point jumped from 10 to 40 within two months. As is true of many projects, we operate under a budget limited in both dollars and time. Because of the nature of SENIC, and its funding mechanism, time constraints have often been even more severe than dollar constraints during the first two years.

The remainder of this paper is divided into eight sections: Getting Started, Finances and Personnel, Data Management, Programming, Pretesting, Communications, Collaborative Design, and A Final Word.

2. Getting Started

As a statistician about to become a statistician manager, you may foresee your job as mainly technical, consonant with your academic training - yes, there will be some bothersome administrative duties, but my problems will be basically statistical. You are in for a rude awakening.

The first administrative duty to be coped with is planning and budgeting. You may be accustomed to following your research interests, not knowing where they may lead, and to taking on consulting problems as they come. But sponsoring agencies want to know exactly what you will be doing for a year or more in advance, and how much it will cost them to have you do it. Two methods which we have found useful in making projections are Gantt charting and PERT charting (see[9]). These techniques help you organize the activities to be carried out and their relationships to each other, to estimate the time and resources required for them, to schedule personnel and equipment acquisition, and to project feasible performance deadlines, thus yielding a basis for estimates of costs. Thorough study of Kendall and Stuart provides no enlightenment on these topics.

Just this initial task should convince you that what you need most, and right away, is an administrative assistant. This may be the most useful employee of your center. He and his staff will see to the details of position creation, space, purchasing, and travel; he will monitor expenses for payroll, equipment, computer time, and whatever; and he will produce necessary internal and external audit reports. It is particularly important that he be aware of the regulations imposed by the agencies which have authority over your administrative acts, and be effective in reducing the delays which tend to accompany the satisfaction of such regulations. Routine administrative matters should not require the constant attention of a statistician manager; turn them over to your administrative assistant. Remember also that he will be ready to help when contract renewal time rolls around, and planning and budgeting must be done again.

One other piece of advice: obtain as much expert opinion and help as possible in the early stages of the project. We hope this paper will be of value, but it certainly won't qualify you as a manager. The references may be helpful also; Mintzberg [4] should be an eye-opener. But nothing will be as useful as discussions with others who have been along this route before you.

3. Finances and Personnel

Be sure that you have secured a budget adequate for performing the required tasks. Our Center's history shows continual underestimation of programming tasks, computer costs, and data entry requirements. For example, we began with one programmer and two data processors and projected only a modest increase in staff over the life of the project. There are now 16 programmers and 11 data processors! And our budget for computer time has jumped from about $10,000 to $200,000 per year. In fact, the initial budget projection for SENIC data management was vastly under the actual costs. The first four months were funded at a modest level to support statistical design and planning for the Center. The next year's budget tripled that rate of expenditure, and the present budget triples <u>that</u> rate. To avoid such underestimates, you should consult with data management experts who have dealt with the entering and processing of <u>large</u> data sets. Estimates of needs may be based on your requirements and their experience, but you still need to project for the possibility of non-standard systems; see Section 5, Programming. Moreover, you should exercise sufficient supervision and quality control over programming tasks, especially if they have been subcontracted elsewhere.

Thus a large center's budget may well involve hundreds of thousands of dollars. This necessitates careful monitoring of all expenses, including payroll, purchases, and computer costs. Have your administrative assistant prepare summaries of expenses at least monthly, with projections to the end of the budget period. (Keep in mind that these projections must usually be nonlinear.) And by all means stay on good terms with the relevant fiscal and personnel offices.

It should go without saying that recruitment of the right people for your center is an extremely vital early task. You may be surprised, however, at how long it can take to set up a position and then fill it, and the myriad of regulations which accompany the process.

Be prepared for the inevitable turnover in personnel. Sometimes this will involve a key person. It is important to consider the lead time needed to find a suitable replacement. In SENIC, our first chief programmer left after six months. Because of university regulations, no recruitment was possible until formal notice of resignation had been given. We were fortunate in finding an excellent replacement in only one month. At the end of the first year, the statistician manager resigned in order to head a biostatistics group elsewhere. Fortunately, there was a nine-month lead time to recruit a new manager.

Supervisors should not make promises concerning promotions or reclassifications to staff until the manager has considered the implications of such changes; individual enthusiasm may have to yield to overall morale and the realities of budget constraints.

4. Data Management

In large data management centers, it is necessary to have formalized procedures to handle the inundation of forms arriving. There may be only one form for many cases or there may be many different forms for each case. For each form, there are a number of steps to be followed.

a) Log in. A detailed log of all forms should be maintained (sometimes simplifications are possible if records are kept in batches). Whenever forms are shipped from the collection site, a notification of shipment should be sent under separate cover; and when they are received at the center, an acknowledgment should be returned. Then any loss in shipment will be detected and steps can be taken to trace it. Copies of the forms may be retained at the collection site until their receipt at the center is verified. If periodic reports are expected, a tickler file is useful.

b) Microfilming. This serves two purposes. First, it is a precaution to protect against loss during data entry; this implies, of course, that it must be done immediately after log in, and the forms held in abeyance until the film has been developed and checked for readability. Second, microfilming provides compact long-term storage. Typically, after data editing, there is no need to retain the original forms, and they can be destroyed.

c) Data entry. Since all large studies must use computers for data processing, there must be a way to convert the data from the data forms to machine-interpretable form. In SENIC, almost 200 million characters must be entered, verified, and edited. Three methods of data entry were considered.

Keypunching to cards is the data entry method most familiar to statisticians. It is quite satisfactory for small or moderate-sized studies; but, for large studies, the problems of card storage and the extra read-in step required make it less desirable than key-to-disk or optical character recognition. In a key-to-disk (or key-to-tape) system no cards are used; the data are keyed directly onto a disk (or tape). Some keying services feed their key units through a mini-computer which can perform range checks on fields as they are keyed in, thus speeding up the editing process.

Optical character recognition (OCR) offers the possibility of great speed in data entry, but also places stern requirements on the forms and the manner in which data is recorded. The forms must be printed on a heavier weight, higher-quality bond paper, with a special ink which is not detected by the OCR reader, and the printing must meet close tolerances. The forms must be in good physical condition (e.g., without tears and dog-ears or stray marks and stains) to be fed through the reader. In addition, the handwriting must be carefully done in a prescribed style, requiring special training for the data recorders, or data may be typed using a special font; otherwise there are problems with nonrecognition and misrecognition of characters. The advantage of OCR as compared with keying systems is also affected by the density (in terms of characters per sheet) of the forms. Initially, SENIC had planned to use OCR. However, when the pilot studies provided data on the cost of printing, form density, and the quality of scanning of handwritten material, and these data were combined with the specter of technical problems, including the unavailability of a backup OCR reader in case of failure, the subsequent cost-effectiveness analysis resulted in our switching to a key-to-tape system.

d) Editing. There are two main types of editing procedures: range checks (or more generally, legitimate code checks), and consistency checks. For example, (range check) if sex is coded 1 for male and 2 for female then it must not have the value 3; or (consistency checks) patients cannot be discharged from a hospital before they are admitted, and hysterectomies cannot be done on men.

Not all fields need editing. To completely edit every field in large studies would be wasteful, expensive, and probably impossible. Moreover, not all errors found by a computer edit should be manually recoded. In one pretest, we found that 95% of the forms had at least one error. To correct these errors would have tied up our staff for far longer than we could afford. Thus, the editing decision generally is not how to correct a field, but whether to impute a value, or assign an "unknown" code. A very few important fields may be manually corrected, but the total number of manual checks must be small.

e) Tape management. Because computers occasionally break down, it is imperative to protect data tapes by (i) having duplicates, (ii) having a cycle of tapes which have the data from several weeks of accrual. Thus, Tape 1 might be the most recent completely edited set, Tape 2 the previous week's, and so forth. When a new updated edited tape is ready, it takes its place at the head of the line, and all other tapes move down one place. In addition, there should always be backup tapes of the original raw unedited data and of at least the first post-edit generation.

Good reference sources for further information in this area are Helms [3], Naus[5], Simmons [7], and Smith and Helms [8].

5. Programming

This area has already been alluded to several times because of its great importance. Here, we suspect, is where too many statisticians come to grief. Of course, as a statistician manager you are inexperienced, but this excuse works only once; so get advice from competent sources. This will not solve your problems completely, but may help you to avoid at least the two fundamental errors described below.

a) You may greatly underestimate the magnitude of the task, for at least three reasons.

i) At the start of the project, you have not identified all the programs which will be necessary. For example, you are probably familiar with the need for an edit system and a statistical analysis package, but you should also consider: a system for auditing the quality of data collection; a system to account for all forms at each stage from initial entry as raw unedited data to final archival in the definitive data base; a hierarchical file management system for manipulation of data prior to statistical analysis; and a system for keeping records and statistics on all programs executed. (See also [2]). The fact is that the full range of programming and data processing tasks cannot become evident until the project has evolved.

ii) You discover that reprogramming is necessary. After a program has been put into use, you may find that it does not perform properly, or it is too expensive, or it does not do all that is needed, or it does not mesh with the rest of the system. The task must then be reprogrammed making the necessary changes. Brooks [1] suggests that you should plan to discard the first program that is written, as it will have to be redone anyway. For example, once pilot data became available, we modified the specifications for the computerized diagnostic algorithm to improve its accuracy. We knew in advance, of course, that such reprogramming would be needed.

iii) You undertake ad hoc programs which delay completion of scheduled programs. Ad hoc tasks that "only take a day" have a nasty habit of taking a week, and causing delivery dates to slip. Heroic efforts to cope with crises may be necessary on occasion, but cannot be a regular operating procedure. One possible solution to this is to have an experienced programmer available, with no regularly assigned duties, who can provide backup help for crisis situations and work on any task until additional help can be provided. He may also serve as a general consultant for debugging or alleviating pressure situations near deadline times. This becomes feasible, of course, only when your programming staff exceeds 4 or 5.

Since the dimensions of your programming tasks have probably been greatly underestimated initially, it follows that you must have built in budgetary flexibility so that expansion of staff will be possible. To prevent slippages, the chief programmer must be able to recognize when additional help is needed, and to arrange for new programmers easily. Anticipation is a key here.

b) You fail to understand the nature of programming.

Brooks [1] notes that there are certain tasks which cannot be assigned to two programmers. Some tasks are indivisible, and only trouble ensues if they are split. On a small scale, we can understand the difficulty of having two programmers write one subroutine. On a systems scale, the same problems arise, particularly in the design phase. With multiple minds working on the same system, there will be inconsistencies, duplications, and omissions. For this reason, your chief programmer should be at the Computer Systems Analyst level; note that even an excellent programmer may not be suited for this position. So let one competent and experienced person design each system and write its specifications, and then have the chief programmer and chief statistician review it thoroughly.

Too much splitting of programming tasks is an error of Type I. But there is also an error of Type II. Make sure that after the design is completed and approved, there are sufficient programmers to handle the individual programs and subsystems. Then these tasks can be parceled out to the other programmers under supervision of the system designers.

The complexity of the programming system for SENIC is far greater today than was envisioned only a year ago. We hope that others can profit from our mistakes!

6. Pretesting

In any study, pretesting is an absolute requirement. It must be conducted to eliminate bugs in the design of the forms, inconsistencies and inadequacy of the protocol, and to check on coordination of the various elements in the data management system. In SENIC several pilot studies were carried out. The first pilot was too large and too rushed; it flooded the Center with forms which we were not yet ready to handle, and we soon discovered many defects that might have been worked out in a smaller and more carefully planned pre-pilot. A second pilot provided far more satisfactory results, and a third pilot actually constituted (for statistical purposes) the first part of the main study, but conducted at a slower pace to allow some fine tuning of procedures.

The main objective of the pilots was to test the accuracy of the retrospective chart review (RCR) method of collecting data from medical records. In additon, we used the pilots to review forms design, the data processing system for forms flow and data entry, and the efficacy and efficiency of the programming system. We discovered many problems.

As a result of the pilots,

1) We instituted a more sophisticated co-ordination system for field activities, supervision from the project officer's staff, and data processing activities at our Center;

2) We improved selection of data collectors and training methods;

3) We placed a quality control/systems manager with each team of data collectors in the field, and stressed that any change in coding conventions must be immediately checked through the Center;

4) We completely redesigned data management procedures ranging from forms shipping procedures and handprint quality to co-ordination between the

technical supervision of OCR and specifications for the printer, which had been done through too many intermediaries;

5) We eventually changed the data entry method (as indicated previously); and

6) We designed a more complex set of programming subsystems, especially the edit system.

A key point is to make certain that the pretests cover <u>all</u> aspects of the system. For example, if the data entry and editing systems are not ready there can be no pretest for them at that time; it is possible to gather the data, and later run through data entry and editing, but this checks only the entry and editing system and <u>not</u> the interface between data collection and data entry which may be the more crucial step.

7. Communications

There are two basic points which we want to make. First, keep in frequent contact both with the project officer on the one hand and your section chiefs on the other. And second, document every step you take. This will result in an avalanche of paper on your desk; set up a prioritized filing system and use it.

Frequent contact between center management and project management is a necessity. This is not difficult to accomplish if the two groups are located in the same facility; if not, get yourself a speaker phone and prepare to make full use of it. Schedule short conferences almost daily to review any current operational problems and overall project strategy. Periodic longer conferences are needed to evaluate progress and set new goals. It is easy to slip on this and become engrossed in daily operational problems. Suddenly, weeks have passed and there is an overly large amount of material to review at the time of contact. This is particularly true when beginning field operations. Many unexpected problems arise and must be solved promptly. For example, one field consultant in a SENIC pilot study adopted a coding convention contrary to that agreed upon by the overall project management. This was not discovered for several weeks, and when it was, all the affected forms had to be manually recoded.

But even frequent contact is still insufficient if it is only superficial. Center staff must keep project management informed of their thoughts, and get their responses. A large part of this involves educating project managment people who may have little or no idea of the complexities of statistical analysis or software development. For example, in another pretest, our project management changed the instructions for filling out a form but failed to notify the Center. This caused every form to be rejected at an edit check.

Even more importantly, members of the center must make every effort to become familiar with the complexities of the subject matter in the field of research. For example, one of our junior statisticians, who had already studied in detail the characteristics of the hospitals in our sample, underwent the rigorous training required to become a certified SENIC chart reviewer. The technical expertise which he gained thereby, combined with his statistical knowledge, makes him an invaluable on-site resource of our Center.

The iterative nature of communication is also well-known. In SENIC, some of our best ideas surfaced only after problems had been discussed for months. Continual review of past decisions is useful, particularly in areas where problems recur.

All agreements with the overall project management should be in writing. The implications of any such agreement should be indicated. For example, a new programming task may use several weeks of programmer time and cause other tasks to slip. If this is not known by the project management, they may be expecting a product when none is ready.

The need for documentation is particularly evident when you consider that new staff will be hired, old staff will leave, and memories will be faulty. For example, every computer program should be documented so that a new programmer can use it, continue its development, or debug it. Standards for documentation are available. We advise a fixed documentation format for simplicity of use. But programs are not the only items for which internal documentation is necessary. For example, minutes of staff meetings, statistical notes, and summaries of communications with project management should all be routinely distributed. Status reports distributed to all staff at frequent intervals are valuable not only for informational purposes but also for morale. Finally, it should be obvious that no staff member can make commitments for tasks to be done without the knowledge and consent of the statistician manager.

8. Collaborative Design

The companion paper [6] will discuss the details of the statistical considerations for SENIC, but here we point out that the actual process of constructing the design of a large study is not a straightforward extension of the process you may be used to following for consulting problems on a smaller scale. Its outline is not all that different to begin with: statement of the problem, characterization of specific objectives, and statistical questions; then determination of sample size and stratification, especially as a consequence of statistical analyses planned and of reports specified as needed. However, the greater scope of the study, and the complexity of the interrelations among its statistical requirements, administrative considerations, and its funding mechanism all dictate that the design process be a collaborative effort: it cannot be a purely statistical matter, involving only yourself and the other statisticians on your staff, but project management will be a full participant. Depending on your areas of expertise, it may be cost-effective to call in consultants. For example, we had some valuable advice from nationally known survey sampling experts. And finally the design must be reviewed and approved at higher levels in your funding agency: SENIC required clearance from the Office of Management and Budget (OMB).

For example, there is controversy surrounding the accuracy of the RCR mechanism as compared with prospective surveillance. RCR is subject to errors in the medical records themselves, in the process of coding them onto forms and in the decision as to whether or not the forms indicate infections, besides the usual sampling errors with respect to patients and hospitals. On the other hand, there are ethical problems with a prospective study, not to mention the effect of surveillance on the phenomenon under consideration (Hawthorne effect - see [6]). The parameters needed to determine sample size and stratification should have come from the pilot studies. But time was so constrained that we had to run sensitivity analyses on various designs with respect to the power of the resulting tests and the ability of RCR to correctly identify infections. The results must be reviewed by OMB before the full survey can be completed.

The hardest problem is not the final construction of the design itself, but the preparation of its specifications. In SENIC, do we mainly need a test of hypothesis on the difference in rates? What descriptive results are necessary? How do these requirements compete with the need to test for associations? How accurate is accurate -- one must specify either a dollar limit or a constraint on variance to determine the design, but generally neither the project officer nor the statistician has such numbers handed to him at the outset. Thus in SENIC we had to make priority judgments concerning which objectives were more important, since they competed in that they dictated different optimal sampling schemes; see [6]. This required coordinating the opinions of the physician-epidemiologists at the funding agency, the statisticians at our Center, and the consulting survey sampling experts. Meetings were held to debate the merits of various designs in view of the aforementioned priorities as well as administrative considerations such as: the airfare and per diem costs for chart review teams, the ability of hospitals and their medical records rooms to accomodate the teams, the need to prepare individual reports for each of the hospitals, and so on.

With day-to-day demands competing for your time, the planning function should not be overlooked. The project officer must commit to paper an initial set of specifications. The statisticians must respond on paper, with the first design. Subsequent sets of "specs" and designs are iterated until a final design is ready. But don't wait until then to involve your chief programmer; let him start planning early for the inevitable system design and coding needs. Your data manager cannot wait till the last minute, either.

9. A Final Word

We conclude our advice to the statistician become manager with two final points.

First: coordination and management of large studies is a very difficult, time consuming task. It will occupy about 120 percent of your time. Thus while the project is current you will get very little other work done. In particular, your institution must accept the fact that personnel heavily involved with such a project may have little published evidence of productivity for a considerable period of time.

But second: watching a large, complex project take shape and begin running smoothly, observing milestones, seeing completed documents, and developing a staff that functions well with high morale, will be among the great rewards possible for this type of research.

ACKNOWLEDGMENTS

We would like to acknowledge stimulating discussions with John Bennett, Howard Freeman, James Grizzle, Robert Haley, Ronald Helms, Don Mitchell, Wendell Smith, and Edward Wagner. We also owe a debt of gratitude to Fredrick Whaley and Donna McClish for assistance throughout.

REFERENCES

[1] Brooks, F.: The Mythical Man-Month. Addison-Wesley, 1975.

[2] Francis, I. and Sedransk, J.: "Software Requirements for the Analysis of Surveys". Proc. 9th Int. Biom. Conf., 1976, pp. 228-252.

[3] Helms, R. W.: "Data Analysis Procedures for a Coordinating Center for a Large Cooperative Study". Institute of Statistics Mimeo Series No. 1003, University of North Carolina, 1975.

[4] Mintzberg, H.: "Folklore and Fact". Harvard Business Review, 53 (1976), pp. 49 - 61.

[5] Naus, J.: Data Quality Control and Editing. M. Dekker, 1975.

[6] Quade, D., Lachenbruch, P. A., Shachtman, R. H., and Haley, R. W.: "Managing the Statistical Center for a Large National Study in the Developmental Stage: The SENIC Project, Part B: Statistical Considerations". Institute of Statistics Mimeo Series No. 1131, University of North Carolina, 1977.

[7] Simmons, W.: "Operational Control of Sample Surveys". Laboratories for Population Statistics, Manual Series No. 2, University of North Carolina, 1972.

[8] Smith, W. D., and Helms, R. W.: "Data Management in Large Studies: Data Acquisition, Data Entry, Data Manipulation". American Public Health Association Meeting, Miami Beach, 1976.

[9] Wiest, J. D. and Levy, F. K.: A Management Guide to PERT/CPM. Prentice-Hall, 1969.