METHODOLOGICAL PROCEDURES IN A VERY LARGE NATIONAL SURVEY OF PHYSICIANS

T. Donald Rucker, National Association of Blue Shield Plans
and
Martin Taitel, Chicago, Illinois

Introduction

This paper discusses some of the methodological considerations that arose during a large survey of physicians in the United States. The purpose of the study was to determine the relationship between Blue Shield payments and the cost of physician care for selected types of medical services under various contractual arrangements.

The survey was conducted by 54 (out of 75) Blue Shield Plans and covered nearly 90 percent of the payments made by these organizations under their basic contracts. Questionnaires were prepared from selected claims paid during the six-week period ending June 26th, 1964 and they served as the basis for the study. Four hundred seventy thousand forms were mailed to physicians and over 366,000 usable returns (78 percent) had been received by the middle of October. Before referring to the methodological problems which are the subject of this report, we will present a brief outline of the steps which were followed in implementing the project.

Preliminary work was undertaken by the staff at the National Association of Blue Shield Plans. This consisted of the preparation of a prospectus, design and printing of a one-page questionnaire, pilot-testing the survey in two states, and evaluation of the findings. In addition, the staff prepared a detailed Manual, including sample design and selection, which indicated the procedures that all Plans should follow. Finally, the staff conducted regional orientation meetings in five cities to ensure that uniform and correct methods were used by Plan personnel throughout the nation.

The various Blue Shield Plans then performed the following functions: (1) Mailed an orientation letter to all physicians in their areas about ten days before sending the first questionnaires. (2) Selected claims from surgical, maternity, anesthesia, and medical services according to random sampling techniques. Small Plans, generally those with an enrollment under 240,000, participated on a census basis. Plans whose enrollment exceeded this figure selected claims using the terminal digit of the claim control number as specified by a table of random numbers. (3) Prepared questionnaires by inserting information at the top of each form which furnished the doctor with the name of the patient, date, type of service, procedure code, description, and amount paid by Blue Shield for the given service. (4) Prepared an 80-column punch card for each selected claim. This included 22 items such as age, sex-relationship, county and type of physician, amount paid, type of service and so on. (5) Mailed the forms, as imprinted with the identifying information, at intervals of either two or three weeks. As a result of this accumulation process, some doctors received 20 or more questionnaires in a single mailing. This problem arose because a large proportion of their business involved patients with Blue Shield contracts, and because of the distribution of claims by size.

All of the punched-cards, and except for four Plans, all of the questionnaires were sent directly to the trade association headquarters in Chicago. (The four Plans subsequently shipped their returned forms to Chicago en mass.) At this point, the staff at NABSP assumed responsibility for editing the forms (4500 man-hours) and having the information keypunched. These cards, and those submitted by the Plans, were placed on magnetic tape with admissable-code controls specified for each field. The final step included matching the basic data cards with their corresponding informational cards and calculating the weights. From the resulting 115-column records, numerous reports showing Blue Shield's performance have been and will be prepared.

The above overview should provide sufficient background to enable us to proceed with a discussion of selected methodological problems encountered during the survey.

Variates Used to Measure Plan Performance

A measure of Plan "performance" was required that met a number of minimum requirements. These were deemed to be (a) effective differentiation between different levels of performance, (b) application to different types of certificates and services within Plans and between Plans, (c) capable of calculation in terms of operational constraints, (d) simple enough to be understood by a variety of interested parties. Performance, therefore, was subsumed under a single ratio which was obtained by dividing the Blue Shield payment for a given service by the cost of the doctor's care for the same service.

Such a measure is satisfactory if all other things are equal. Yet the survey was conducted in the real world and the variates were influenced by a number of factors. Among these were differences in fee schedules, benefit levels, supply and type of physicians, number of hospital beds per community, subscriber incomes, proportion of premium dollar available for benefit payments, age and sex distributions, proportion of premium contributed by the employer, frequency with which patients in Service Plan areas visited Participating physicians, and so forth. In undertaking the survey, there was no practical way to control for such items. Research will be conducted subsequently, however, to assess the impact of a number of variables on the performance results which were obtained.

The Statistical Unit

Selection of the statistical unit seemed to be confined to three choices - the claim, the illness or the patient. Focusing on either of the latter two would have required access to a number of related records. From an administrative point of view, there was no economical and quick way to pull all the claims together for a particular illness (assuming terminal points could have been defined) or patient. Moreover, some of the financial obligations of the patient might not have been a Blue Shield responsibility. Nor was there any way to overcome the fact that certain claims pertaining to a case may have required a number of months to process. Further, the time lag with which doctors (or a given doctor) submit claims to Blue Shield varies considerably.

The above reasons, when combined with the difficulties inherent in a short time period of six weeks, militated against using either the "illness" or "patient" as the unit. Consequently, it was concluded that the least-worst approach to Plan performance was by means of the "claim."

Sampling Design

The prime consideration in the determination of sample size was to obtain results which would be useful in the operations of individual Blue Shield Plans. The precision of national averages was not critical; nor was the precision of Plan averages. Rather, each Plan had to be considered as being partitioned into cells, and it was the precision of each individual cell which was the critical factor. Thus, there was not one overall national or Plan sample but a large number of samples. The exact number was dependent upon the extent and nature of the stratification used for sampling purposes.

One extreme of the possible sampling designs was to have sampling strata and analytical cells in one-to-one correspondence; the other was to take a single sample for a Plan large enough to ensure satisfactory precision for cells within the Plan. Actually, an intermediate design was used.

As indicated above, the design incorporated stratification. But, instead of setting the sample size to provide the required precision for a stratum as a whole, it was increased by an amount estimated to be sufficient to ensure reasonable proportionate representation for the smaller, but sizeable sub-cells within the stratum. In effect, this yielded more than the required precision for the sampling stratum as a whole, and about the required precision for the larger sub-cells within the stratum. Moreover, it ensured satisfactory precision for many averages of marginal distributions of variables not used for the sampling stratification.

The original sampling design indicated a potential of some 2500 different samples - the exact number being unknown because provision was made for Plans to sub-stratify if necessary for operating needs and because the number of null cells was not known. A minimum sample size of 225 was arbitrarily set for small sized populations (about 300 or less). The theoretical maximum sample size for a sampling stratum was computed at about 700, though in certain instances larger samples did actually occur.

Stratification and The Sampling Unit

Within Plan stratification was clearly necessary because of widely different groups and widely different sizes of the groups. It was not possible, however, to achieve the optimum benefits of stratification. To do so would have required an identity between the variables of classification and the variables of stratification. As noted above, there was a potentially large number of variables of classification; at the same time, though, administrative consideration limited the variables of stratification to a very few.

Stratification was based upon the contractual relation between the subscriber and the Plan (with one exception) and upon the generic type of medical service rendered by the doctor. Thus, one variable of stratification was the Basic Certificate which is the major means of differentiating between a great variety of contractual relations; the other was the Type of Service (surgery, anesthesia, etc.) which is the major means of differentiating between a large number of diverse services rendered by physicians. These two had the desirable aspects of being related (or expected to be related to) the performance ratio and of being easily identified and readily available in Plan records. (Some variables - doctor's charge, doctor's specialty - were not generally available at the time samples were selected.)

One feature of the Basic Certificate variable should be noted. It was not the same from Plan to Plan. In order to satisfy the condition of usefulness in Plan operations, it was defined with respect to the Plan rather than uniformly for all Plans. Thus, "Best Certificate" and "Most-Widely-Held Certificate" categories differed from Plan to Plan. But doing so, though, ensured comparisons of the most meaningful kind. The Type of Service variable was uniform over all Plans.

Stratification by size of claim (a substitute for cost of physician care which was not available beforehand) was considered because of the importance of total dollar cost. In addition, the pilot survey had indicated some variation of performance with differences in cost. Further, the proportion of large claims tends to be small; moreover, there was the question of measuring performance on a "per dollar of cost" basis instead of, or as well as, on a "per claim" basis. Obviously, the precision of the regression of "payment per dollar of cost" on "actual cost" and the precision of the "relative frequencies of dollars or claims" by size of claim would have increased without significant change in the sample size. It was not administratively feasible, however, to stratify further than by Certificate and Type of Service.

A suggestion made after survey operations were

under way may be of interest. Professor Nathan Keyfitz suggested that using dollars rather than claims as the sampling unit might have been a better solution. Such a procedure does not necessarily gain the advantages of stratification, and perhaps gain them with smaller samples, without actually stratifying.

Random sampling of dollars, of course, gives claim-selection probabilities proportional to claim-size in place of equal probabilities when sampling is by claim. This is an inherent rigidity not present in stratification. In addition, the number of times a claim is selected becomes a random variable; the comparable item under stratification is a combination of the sampling ratio and the size of the claim - neither being subject to sampling errors. Finally, it may be noted that observations for claims are independent; those for dollars from the same claim are not; hence, it is the number of claims, rather than the number of dollars, which determines the precision of the sample statistics. And to set the sample size, therefore, requires an error element - average size of claim - to be introduced which is not present in stratification.

Systematic sampling of dollars from randomly ordered claims (the assumption used in the survey) would lessen the relative shortcomings of sampling dollars. Let R equal sampling ratio and X equal claim size, then every claim for which $X \geq (1/R)$ would be selected at least once. Other claims would essentially be sampled at random.

It appears, therefore, that random sampling of dollars would have advantages only if (a) it was appropriate to have claim-selection probabilities proportionate to claim size, and (b) frequencies of large claims - say those greater than 1/R - were relatively small.

From the administrative viewpoint, both stratification by size and sampling by dollars would have required large increases in cost. Stratification would have involved them at the point of sample selection, primarily, but also later, though it would have reduced the total cases in the survey. Sampling of dollars would have necessitated sorting and collating of "dollars" selected in order to reduce the data to unduplicated claim records which included a count of the number of times selected.

Sample Size: Specified Precision

The initial precision specification for the survey was in terms of the average performance per claim (ratio of the Blue Shield payment to the actual charge made by the doctor) for a sampling stratum. A 95 percent confidence interval of plus or minus 0.05 was specified when the average was considered as a measurement of the true average performance for the actual finite population (and not for a hypothetical infinite population).

The upper limits for sample size to achieve this precision on the assumption of 100 percent response were as indicated below. Such a routine application of the finite population sampling formula does not, of course, take account of various deviations from the assumption underlying such an application.

Assumed Probability Distribution	V(X/Y)	Upper Limit For Sample Size
$P(X/Y = 1) = P(X/Y = 1) = \frac{1}{2}$	$\frac{1}{4}$	400
$P(X/Y) = d(X/Y)$	1/12	135
$B(3/2,3/2)$	1/16	100

Sample Size: Response Rate

In this survey, expected non-response was the easiest element for which to adjust sample size. Assuming non-responses were at random, division of the theoretical sample size by the expected response rate provided the necessary adjustment. The estimate of an 80 percent response did, in fact, turn out to be of the same order of magnitude as the actual response rate which was achieved. It reflected the results of the pilot survey for which a higher estimate was used (and again achieved).

While the validity of the assumption of randomness of non-response cannot be fully determined, the survey will provide at least some significant indications. These will be derived from comparisons of characteristics of the non-response and the response parts of each sample. (As noted above, a punch-card exists for each non-response. It contains 22 variables, many of which will be invaluable for analytical purposes.) It may also be possible to derive additional indications from between-sample comparisons.

The possibility of biases resulting from differences between the response and non-response parts of a sample was recognized, and provision for offset was made through an increase in the variance estimate. Thus, the precision specification was not applied to the sampling error itself, but rather applied as an asymetrical range for the sampling error plus an estimate of bias.

Sample Size: Estimates of Population Size

Population sizes were not known, but had to be estimated beforehand. These estimates were subject to large errors, because the number of claims submitted or approved during relatively short periods is subject to wide fluctuations. Moreover, the varying ability of Plan personnel also contributed to the estimating problem.

To prevent serious errors in sample sizes arising from this source, two steps were taken. First, in making the estimate of population variance, this element was given consideration; such consideration was designed to offset only the smaller errors resulting from errors in population estimates. Second, and much more important when needed, was provision for modification of the sampling rate during the course of the survey, mailing period by mailing period. Only very marked deviations were covered by this provision; and, when used, each mailing period became a sampling sub-stratum within the original sampling stratum (with independently computed weights in the subsequent tabulations). Fortunately, only a few occasions developed in which sampling rates were modified during the course of the survey.

Sample Size: Non-Homogeneity of Sampling Frame

It was not possible to establish a sampling frame which included only the population to be surveyed. Subsequently, however, it was possible to delete claims which did not belong to the proper population. Two considerations were involved in the provisions made to offset this condition.

First, population size estimates were discounted by the expected proportion of deletions. This by itself could be relied upon to provide the sample sizes necessary to meet the individual sampling stratum precision requirement. However, there was the further consideration that the sample then provided the only estimate of the population size. This second consideration, under conditions where deletions would be substantial, was the more important one.

A calculation made on the assumption of deletions running about 15 percent indicated some 200 cases would be necessary to give a 95 percent confidence interval of 5 percent for deletions, in contrast to some 100 cases in the absence of deletions. Further, reduction of the size of the confidence interval could only be achieved by relatively large increases in sample size.

Sample Size: Tabulating Cells Different From Sampling Strata

Precision for tabulating cells which were subclassifications of a sampling stratum would, of course, be less than for the sampling stratum itself (except possibly under very unusual conditions). No direct provision was made on this account, because such large increases of sample size would have been required. Thus, assuming a sub-cell to be 1/10th of the sampling stratum, and illustrative calculation indicated that an increase from around 100 to 1200 would be necessary to provide the specified precision for this size sub-cell. It may be noted that the number in the sub-cell sampling, under these conditions, is subject to sampling error as well as the averages.

Precision for tabulating cells which cut across sampling strata might, of course, be greater or less than for the sampling strata, depending upon the size of such cells and other factors. Again, no direct provision was made on this account because of uncertainty with regard to the terms of the problem. Instead, reliance was placed upon the provision made for deletions because of non-homogeneity in the sampling frame.

Sample Size: Final Estimation

Based in part upon judgment, the final formula used in actual calculations was:

$$1/n = [(1/N) + (3/1600)](4/5)$$

where n is the size of the sample and N is the size of the finite population. The "4/5" factor represents a discounting for reasons of non-response; and the "3/1600," the ratio of

$V(\overline{X/Y})$ = Estimated variance of sample average performance necessary to achieve the precision specified

to

$V(X/Y)$ = Estimated variance of the parent finite population

Calculations based upon this formula are shown in Table 1.

Adaptation to actual operating conditions required one further modification; i.e., the population sizes to which the sampling ratios were to be applied were rounded up to round figures and the sampling ratio was applied to the minimum for the range. The final figures are reported in Table 1. This adaptation added an extra safety factor, of varying importance from place to place, on the population-size scale.

The above procedure wherein the sampling ratio was based on the lower limit of the expected number of claims in a given class interval tended to inflate the size of the sample. To take hypothetical illustration, the required n where the number of claims fell between 35,000 but under 70,000 was equal to 700. If, however, 50,000 claims occurred, 1,000 sample observations would have been generated by applying the 2 percent rate as called for in the table used by Plan personnel in drawing the sample. Because the last two digits of the claim number were used to select sample cases in the larger cells, (like the one in our example) there was no simple way to write a program which incorporated a selection rate of 1.4 percent which was necessary to yield the desired 700 claims. Thus the survey specified a 2 percent rate and thereby inflated the n in our illustrative cell by about 43 percent.

Table 1

Sampling Ratio	Calculated		Modified	
	N	n	N	n
0.01	66,000	660	70,000 & over	700-
0.02	32,667	653	35,000 but under 70,000	700-1400
0.03	21,556	647	24,000 but under 35,000	720-1050
0.04	16,000	640	18,000 but under 24,000	720-960
0.05	12,667	633	14,000 but under 18,000	700-900
0.06	10,444	627	12,000 but under 14,000	720-824
0.07	8,857	620	10,000 but under 12,000	700-720
0.08	7,667	613	8,000 but under 10,000	640-800
0.09	6,741	607	7,000 but under 8,000	630-720
0.10	6,000	600	6,000 but under 7,000	600-700
0.15*	3,773	566	4,000 but under 6,000	600-900
0.20	2,667	533	3,000 but under 4,000	600-800
0.25*	2,000	500	2,000 but under 3,000	500-750
0.30	1,556	467	1,500 but under 2,000	450-600
0.40	1,000	400	1,000 but under 1,500	400-600
0.50	667	333	700 but under 1,000	350-500
0.60	444	267	500 but under 700	300-420
0.70	286	200	400 but under 500	280-350
0.80	168	133	300 but under 400	224-320
0.90	74	67	275 but under 300	248-270
1.00	1-73	1-73	Under 275	N

* These optional steps were included so that the Plan, if it chose to do so, might limit the number of claims drawn in the survey. The disadvantage from an operational point of view was that these cells, like those with more than 7,000 claims, required that selection be based upon the last two digits of the claim number rather than the terminal digit as indicated for all other conditions.

Because the number of cells where N exceeded 3,000 claims was not great, and because there was no a priori reason to indicate that, on the average, the actual N would tend to exceed the midpoint of the class interval, it is estimated that the inflationary factor increased the over-all sample size by, at best, 10 percent. This was construed to be a salutary feature since precision would be increased and administrative considerations suggested no better alternative.

Sampling Variance and Teleology

The survey provides an illustration of situations in which there is not necessarily a unique sampling variance. Thus, depending upon the definition of the parent population or, stated otherwise, upon the use of the statistic involved, the sampling variance may be upon the basis of a finite population or upon the basis of an infinite population of one specification or another.

The observations actually obtained were "random samples from a finite population." Considering the sample $\overline{X/Y}$ as a measure of the true average performance of that finite population, the sampling variance is appropriately computed according to the usual finite population formula. However, the sample may also be considered as the result of a two-stage sampling process so that it is selected, not from a population, but from a first-stage sample generated under given conditions from an infinite population. In this latter case, assuming random generation of the sample, the sampling variance is appropriately computed according to the usual infinite population formula, so that it provides a measure of error when the sample results are used as measures of the hypothetical infinite population.

Finally, it may be noted, sample results may also be considered in other ways; for example, there may be an absence of randomness in the first-stage sample, even though the second-stage sample is taken at random; or, again, the hypothetical population for which random sampling may be assumed appropriate is not the hypothetical population with reference to which the sample results may be used, thus raising the question of measuring biases. In such cases as the latter ones, the usual formulae have, of course, to be

modified in terms of special considerations.

In the survey under discussion, sampling variances have been computed both on the finite and infinite population basis using the assumption of random selection. It is recognized that, in transferring from the finite to the infinite population assumption, more than a formula change is ncessary. Such items as differences in relevant conditions over time, including fee schedule adjustments, benefit levels, medical risk characteristics, months of the year, and so forth, must be considered before the extent of sampling and other errors may be judged.

Validation By Means of a Patient Survey

Formal validation of the survey was attempted to obtain a general indication of the reliability of the results. Some of the factors that might have contributed to respondent bias in the project were as follows: Many physicians have a heavy patient load and the doctor, in his haste to complete the form(s),[1] might have inadvertently supplied incorrect information. A secondary source, such as a nurse, could have been responsible for completing all or part of the questionnaire. Moreover, recorded data, either in the doctor's files or furnished on the form, could have led to an erroneous response. Finally, deliberately biased answers could not be discounted since one of the purposes of the survey was to test whether Blue Shield fee schedules were realistic in terms of contemporary costs of physician care.

Consequently, in order to indicate the magnitude of any such bias, the Plans were requested to draw a second sample for the purpose of determining which patients would be contacted. This was done by taking a systematic sample from those claims previously drawn in the doctor survey. One of the pilot-study Plans selected claims at the rate of 1/11 while most of the others used a rate of 1/20. Theoretically, the former seemed more desirable, but from an administrative point of view, the taking of only one out of every twenty was close to the maximum that could be achieved under existing conditions.[2]

The Blue Shield Plans, therefore, prepared a questionnaire similar to the one sent to the physician which included the appropriate identifying data needed by the patient. These were mailed about 45 days after the last doctor forms had been forwarded. This timing was specified so that the doctor would have time to bill the patient, if contemplated, and still minimize the possibility that patients would misplace their health care records, move away, expire, etc. Some of the Plans sent a second, duplicate questionnaire to those patients who had not replied within 40 days (as indicated by NABSP records in Chicago). No follow-up, however, was undertaken in the physician survey.

It is not anticipated that every returned patient questionnaire will confirm the cost information on the similar form submitted by the physician. Among other reasons, many patients may find it difficult to isolate the cost of the procedure covered in the survey from related expenses for the same illness. It is expected, nevertheless, that a sizeable majority will confirm the results obtained in the primary survey, that the differences will be randomly distributed and, to a large extent, offsetting.

Summary

This paper has outlined some of the methodological questions that arose during a large, national survey of physicians. The discussion dealt with statistical techniques that were inherent in such a project and attempted to focus on the relationship between theoretical and operational considerations which formed the basis of the survey. In none of the sections was it intended that the treatment should be characterized as "exhaustive." It is hoped, however, that the issues covered will be of interest to some who labor in the field of applied statistics.

---

[1] Although the distribution of doctors according to the number of questionnaires received, in total and by mailing period, is not yet known, it is true that a large number of physicians did get more than ten forms. A study of the propensity to response under such conditions should be of great interest.

[2] Substantial resources, financial and personnel, already had been committed by the Plans to this project. Moreover, professional relations considerations militated against a large survey of patients.