HIDDEN BIAS IN THE USE OF ARCHIVAL DATA

ROSALIND J. DWORKIN
Department of Psychiatry
Baylor College of Medicine

Nonresponses in archival data may violate the missing-at-random assumption in ways difficult to detect. Standard methods of comparing sociodemographics of respondents and nonrespondents are inappropriate when the units of analysis are not also the individuals who maintain the archival record. Under these circumstances, the distribution of missing data may be correlated with the dependent variable and traits of the record keepers. This will distort relationships, especially when listwise deletion of missing values is used in multivariate analysis. Data are used from a large clinical chart study of mentally ill patients to demonstrate the process of identifying hidden bias and the implications of such bias.

EVALUATION & THE HEALTH PROFESSIONS, Vol. 10 No. 2, June 1987 173-185

Missing data create problems that plague researchers in both the behavioral sciences and medical research. Although missing values can diminish sample sizes in simple statistical procedures such as students' t-tests and chi-squares, the problem is most dramatically salient when multiple predictor or multivariate analysis is done. With the preferred listwise deletion of missing data (that is, if a case has a datum missing on one variable, the case is entirely deleted from the analysis), the sample size can be severely diminished, to half or less of the original n. This diminution can alter the attendant probabilities of the resulting statistic and, more important, it can raise substantive questions as to the interpretation of the results. One is forced to ask if the results can be generalized to the conceptual population or even to the original sample drawn.

Because of its importance statistically and substantively, much has been written with regard to ways of dealing with missing data. Most writings have dealt with missing data due to nonrespondents: those sampled subjects who either refuse to participate in the research or those who cannot be located (Bradburn and Sudman, 1979; Schuman and Presser, 1981). Nearly all textbooks in research methodology in the social sciences deal with this issue in two ways (see, for example, Babbie, 1986; Dillman, 1978). The first suggests reducing the refusal rate through the use of experienced interviewers and/or locating individuals through various call-back techniques. The second way is to assess the bias introduced by comparing respondents and nonrespondents.

Of increasing interest is the case in which a sampled individual participates but some of his or her data are missing. Papers and textbooks dealing with this problem have advocated two different approaches to the problem. One approach has been to attempt to assess the bias introduced by missing data. This approach parallels the approach described earlier. It generally suggests that the researcher examine the differences between the full respondent and the partial respondent on a selection of sociodemographic characteristics (Babbie, 1986; Dillman, 1978). Oftentimes this must be done through indirect means. Nevertheless, if there are no discernible differences sociodemographically, one may conclude that there is no reason to suspect that there will be

differences on other variables, including the dependent variables of interest. One would therefore proceed with analysis of the complete cases under the assumption that generalizability has not been compromised (Bridge, 1974).

The second approach is to estimate the values for the missing values through various techniques such as averaging or maximum likelihood strategies. With estimation of the missing values, the sample size is maintained at its maximum through the analysis, and statistical significance will not suffer. The various statistical approaches to the problem all assume that data are missing at random (Beale and Little, 1975; Gleason and Staelin, 1975; Haitovsky, 1968; Hawkins, 1975; Marini et al., 1979; Ruben, 1974, 1976a, 1976b, 1977), meaning that "given particular values on observed variables—the values missing on other variables are missing at random" (Marini et al., 1979: 316). Obviously, the two approaches (that is, assessment and estimation) are not mutually exclusive. After ascertaining the absence of systematic bias one could then estimate the missing data.

Nevertheless, it is the contention of this article that however adequate or inadequate these approaches might be for interview or survey types of data, their use may fail to identify serious bias when certain types of archival data are used. Of particular concern are situations when the individuals who maintain the archival records are not the individuals about whom the archives are kept. Such is the situation when medical records (for example, clinical charts) are the data source to be used. Using one study that drew data from clinical charts as an illustration, this analysis will demonstrate how identifying bias may be a difficult, chancy thing to do. Furthermore, it will be demonstrated that the missing-at-random assumption may not be a valid one, and may introduce serious distortion into known relationships.

METHODOLOGY OF THE FIVE CLINICS STUDY

The Five Clinics Study is a retrospective study using the clinical charts of outpatients attending five community mental health

centers. Conforming to proper methodological practice, the coders were not clinic employees coding their own archives. Rather, data were coded by a team of workers trained in the social sciences and especially employed for this job (Esrov, 1981). The coders had two days of itensive training on the coding protocol and associated psychiatric vocabulary before beginning the work. Once in the field 10% of the records were double coded for reliability (Esrov, 1981). There was constant in-field supervision with retraining done as needed (Esrov, 1981). Hence, one can be reasonably confident that missing information and concomitant effects are the result of incomplete charts rather than unreliable coding.

The coding protocol was designed to abstract data from the standard forms found in the charts (Hendrickson and Myers, 1973). With one exception, open-ended qualitative data noted in the charts were not used. Thus if data were missing it was because it was not supplied on the form, thereby avoiding uninterpretable omissions in the qualitative progress notes. Since all the clinics were part of one centralized system, all the forms used were identical across sites (Murnaghan and White, 1971). Qualitative progress notes were utilized for only one group of variables (those dealing with hospitalizations). To verify these data, each manager was asked to compare our coding with his or her own estimates for each appropriate case.

A saturated sample was drawn of all active case files in March 1983. At that time, 1752 were coded. One year later (in February 1984) the same charts were sought for follow-up coding. When charts were no longer in the original center, the other facilities were searched to find the still-active cases. Inactivated cases were located in the centralized medical records library. Using this method, over 90% of the original charts were included in the follow-up (n = 1620).

This group of 1620 cases will constitute the working population for this analysis. A total of 272 variables were collected on this working population. The completeness of data in this set ranges from some variables with responses on every case to variables that have nearly 40% of the cases missing.

THE SEARCH FOR BIAS

For this demonstration, let us hypothesize a theory to be tested consisting of a multiple regression model to predict retention of patients in the centers, using a selection of predictors. The predictors were not chosen arbitrarily, but on the basis of the established literature. These independent variables include sociodemographic variables (age, ethnicity, sex, education); social psychological variables (insight and motivation); social support (living situation, source of financial support) and patient residential density (a measure based upon the plots of addresses); and various treatment variables (diagnosis, years in treatment, number of different treatment modalities, number of hospitalizations and days hospitalized, number of medications, ethnic similarity to case manager). Of the 17 variables in the model, 4 (retention status, ethnicity, sex, and number of medications) had information on all patients (n = 1620). Two other variables had only minimal missing values (age, n = 1619; and diagnosis, n = 1616). The n's of the other variables ranged from 1494 (92%) for years in treatment to only 1112 (69%) for motivation.

Of the total working population, 34.3% of the patients had no missing values on any variable in the model. In all, 80% had three or fewer missing values. However, over 5% of the charts had 11 values missing. The mean number of missing values among patients in the working population was 2.2. When the proposed multiple regression was computed, the listwise deletion of missing values diminished the n to 556, or 43% of the working population. Clearly any statistic computed on such a drastically reduced sample would be suspect. Equally clear is the fact that there is no direct way to assess the extent of bias introduced by these missing cases.

However, to assess the extent of bias caused by the extensive number of missing values, two demonstration analyses were done. First, analyses were done to ascertain the effects of the missing values upon the sociodemographics of the sample. It has been argued that if no differences were found between the sociodemographic characteristics of the responding sample and

the nonresponding group, there is no reason to suspect that the missing values will introduce bias to the dependent variable (Bridge, 1974). Hence, four variables (age, education, sex, and ethnicity) were chosen as index variables. The four were chosen because they are four commonly used sociodemographic traits available in this data base and are often utilized in other studies for these purposes. (Income, another sociodemographic characteristic often used, was not available in this data base.)

Each of the four index variables were examined for significant differences between the sample with complete data and the sample with missing data. This was done for all of the independent test variables taken individually. For example, the effect on mean age of having diagnosis missing was assessed by comparing the age of the sample with diagnosis coded with the sample of patients with diagnosis missing. The effects of missing diagnosis was examined not only for age but also for education, sex, and ethnicity. Each of the index variables was examined separately for each of the independent test variables. For the effects on age and education means, students' t-tests were computed; for the effects on sex and ethnicity, chi-square tests were used. The results are compiled in Table 1.

The most striking finding in the table is the inconsistency of results. In this study, education and sex are not significantly affected by missing values on any of the variables examined. However, the age and ethnicity variables do show some significant effects of missing values, but they differ somewhat on which missing variables have an effect. In addition, missing data on two test variables (diagnosis and insight) did not significantly change any of the four index variables. One would expect missing diagnosis to have little effect since there were very few missing cases on that variable (.2%). However, insight had over 30.5% missing cases, but its missing cases also made no significant difference.

This inconsistency of effects on sociodemographic index variables is troubling. Under most circumstances only two or three variables are selected for such comparisons. A chance selection of index variables and test variables may yield conclusions that are misleading. Unknowingly one could choose only

TABLE 1
The Effect of Missing Values Upon Selected Sociodemographic Variables: Comparisons of Data Missing and Data Present Subsamples

	Socio-Demographic Index Variables			
Independent Test Variables	AGE t (d.f.)	EDUCATION t (d.f.)	SEX (d.f.=1)	ETHNICITY (d.f.=2)
Diagnosis	-1.28 (3.1)	-.36 (3.)	.080	.430
Education	-4.30 (511.5)	-- --	1.345	10.839*
Insight	-1.28 (986.7)	-.85 (560.0)	.639	.537
Motivation	-1.96* (1038.1)	-1.23 (582.8)	.150	2.339
N Treatments	-1.73 (332.7)	-1.59 (146.7)	1.340	8.141*
Years	-3.74* (149.1)	-.51 (20.6)	1.920	.204
Living Situation	-4.51* (609.7)	1.13 (165.1)	2.033	6.061*
Financial Support	-4.46* (635.9)	.23 (218.9)	2.816	16.214*
Case Manager	-2.06* (152.3)	-1.32 (45.7)	.040	.311
Patient Density	-.56 (345.7)	1.20 (193.4)	.131	12.859*
N Hospitalizations	-2.21* (486.2)	-1.39 (258.0)	.6925	17.945*
Hospital Days	-2.53* (507.3)	-1.18 (270.8)	.897	16.566*

*Tests for differences between data present and data missing subsample; $p < .05$..AE less than or equal to .AF.

those variables that would support the hypothesis of nonbias when such a hypothesis may, in fact, be untenable. Only by examining all the variables available can we begin to question the missing-at-random assumption.

When the data are examined in another way, we can see further evidence that the missing-at-random assumption is suspect. To demonstrate this, let us focus upon potential dependent variables rather than upon sociodemographics.

Months in Treatment (1983-1984) and Number of Medications were chosen as potential dependent variables because of likely substantive interest. Moreover, these two variables themselves

had no missing values and thus a *true mean* of the working population could be used as a standard by which to compare the means of these variables under the conditions of present and absent data on the independent variables.

As we can see in Table 2, the true mean of Months in Treatment is 11.438. When the mean is recalculated on the sample that excludes cases with age missing, the mean is 11.437 (n = 1619), a nonsignificant departure from the true mean. For many independent variables tested, the difference between mean Months in Treatment of those with missing data and those with complete data is not significant. However, there were five variables with significant differences. For example, when patients missing on days hospitalized are excluded, the mean deviates by .318, a statistically significant departure. Moreover, the direction of the bias varies. In two of the five significant variables, the inclusion of only complete cases deflated the means. In the other three, the means were inflated.

Such inflation was the case even more consistently when Number of Medications was the dependent variable. However, it is important to note that the missing values of different independent variables tend to affect the mean Number of Medications moreso than was the case with Months in Treatment. Furthermore, Table 2 indicates the effect on the means of the dependent variables when a listwise deletion is used. Using listwise deletion, the various biases do not cancel each other out, but tend to cumulate. Moreover, there is no significant correlation between the size of the deviation from the true mean and the sample size. Therefore, we cannot know the extent of bias from the extent of missing data. To summarize this demonstration, we cannot predict with any significant accuracy the amount of deviation, the direction of the bias, or the variables having a biasing effect without actually making variable by variable comparisons.

If the missing-at-random assumption has been violated, then there will be biases not only in the univariate descriptions of the sample and the dependent variable, but very likely the relationships found in any multivariate model may be distorted. To continue the demonstration of the problem, let us examine the changes in a known bivariate relationship when listwise deletion

TABLE 2
Means of Dependent Variables for Condition of Present Data on Independent Variables (N = 1620)

		Dependent Variables	
		Months in Treatment	Number of Medications
No Exclusions: True Mean		11.438	1.404
Independent Variable	n		
Age	1619	11.437	1.404
Education	1274	11.443	1.440 *
Diagnosis	1616	11.440	1.405
Insight	1126	11.442	1.421
Motiviation	1112	11.468	1.422
N Treatments	1379	11.413	1.434 *
Years in Treatment	1494	11.349 *	1.408
Living Situation	1224	11.485	1.429 *
Financial Support	1220	11.563 *	1.424
Case Manager Similarity	1491	11.378 *	1.411
Patient Density	1359	11.503	1.417
Hospital Episodes	1299	11.734 *	1.454 *
Days Hospitalized	1290	11.756 *	1.454 *
Listwise Deletion	556	12.196	1.462
r(p) #		-.173 (p=.57)	-.536 (p=.06)

*Students' t difference between means of data present subsample and data absent subsample; $p < .05$. .AE less than or equal to .AF.
#Correlation between deviations from the true mean, and subsample size.

of missing values is done as one would in a multivariate model. As the standard, the true relationship between retention and ethnicity will be used. This is a known relationship in this population since there are complete data on both variables. The relationship is significant ($\chi^2 = 19.408$; d.f. = 4; $p = .0007$) when the full working population (n = 1620) is included in the contingency table. However, if we utilize a listwise deletion of missing values for the same independent variables appearing in previous demonstrations, the n diminishes to 556 and the chi-square is reduced to nonsignificance ($\chi^2 = 2.569$; d.f. = 4; $p = .632$).

Moreover, the impact of this missing bias is not evident when missing cases on single variables are deleted. When the relationship is examined under the condition of complete data on each variable taken individually, the relationship is maintained (see Table 3). It is only when the missing values cumulate that the

TABLE 3
Chi-Square Tests: Retention by Ethnicity Under Varying Inclusions

Inclusionary Condition *	n	Chisquare	d.f.	p
All included	1620	19.408	4	<.001
Independent Variable Inclusions				
Age	1619	19.631	4	<.001
Diagnosis	1616	19.903	4	<.001
Education	1274	12.593	4	.013
Insight	1126	10.610	4	.031
Motivation	1112	9.864	4	.043
N of Treatment	1379	21.674	4	<.001
Years in Treatment	1494	18.289	4	.001
Living Situation	1224	11.640	4	.020
Financial Support	1220	18.071	4	.001
Case Manager Similarity	1491	19.999	4	<.001
Patient Density	1359	15.746	4	.003
N of Hospitalizations	1299	10.160	4	.038
Days Hospitalized Missing	1290	10.233	4	.037
Listwise Deletion	556	2.569	4	.632

*Sex and number of medications had no missing values.

relationship is lost. Unfortunately, it is under this latter condition that the analysis would ordinarily be done, in order to keep the sample size consistent throughout the analysis. Thus for the entire population there is a relationship between retention in treatment and ethnicity. However, when confined to only those charts where there are complete data for the model there is no such relationship. If when using listwise deletion there are distortions in the relationship between retention and ethnicity, we can only assume that there will be similar—albeit hidden—distortions in related multiple predictor models subject to listwise deletion.

Indeed, the evidence strongly suggests that the missing-at-random assumption has been compromised. Table 4 lends further credence to this. There is a highly significant relationship between chart completeness and retention. As can be seen in the table, 24% of those with incomplete data withdrew from treatment, compared with only 10% of those with complete data. Thus missing values are associated with the dependent variable. Contrary to being missing at random, this is an example of systematic bias introduced by missing data.

TABLE 4
Cross Tabulation: Retention Status by Chart Completion in Percentages (n = 1620)

	Complete Data	Incomplete Data
Active	85.6	70.6
Referred	4.5	5.4
Withdrew	9.9	24.2
Total	100.0%	100.0%

ChiSquare = 49.887 d.f.=2 p<.001

DISCUSSION AND CONCLUSION

When one uses sociodemographic variables to verify the missing-at-random assumption, one implicitly assumes that the supplier of the data is the same as the subject of the analysis. This is generally the case in sample survey techniques. The individual interviewed provides data about him or herself and the unit of analysis is the interviewed individual. However, with archival data such as clinical charts, this is usually not the situation.

The ultimate origin of most of the data is the patient. However, the chart is written and maintained not by the patient, who is the unit of analysis, but by staff, often a case worker, who is not ordinarily part of the analysis. The completeness of the file is not a function of the patient, but a function of the case worker (or other staff) whose responsibility it is to maintain the record. Therefore, it is the characteristics of the case worker rather than the patient that influence the completeness of the data. Thus when we examine patient characteristics we may not find violations of missing-at-random assumptions because we are looking in the wrong place. We must look to see if there are systematic hidden biases due to missing values vis-à-vis the staff as well as the patient.

In this particular demonstration, the relationship between retention and record completeness is most probably due to the

common link provided by the staff maintaining the record. Characteristics of the staff may affect both variables. Training, professionalism, job commitment, and experience may turn out to have a combined influence upon record completeness and retention of the patient. Other studies, although differing in specifics, may be subject to similar biasing effects. Although our demonstration suggests this interpretation, further analysis of other data sets would enhance generalizability.

In conclusion, we have demonstrated that missing values may impose a hidden bias upon the data that is difficult to locate partly because its source is external to the unit of analysis and partly because of its uneven appearance. Although individual variables may appear to be missing at random, the cumulative effect may be otherwise. Thus estimation techniques based upon a missing-at-random assumption must be used cautiously. Furthermore, the researcher must be mindful of the origin of the data, especially when secondary sources are used. Hidden bias must be sought not only in the traditional way of looking for differences in the sociodemographics of the patient, but also by considering the characteristics of the record keeper, especially when those traits may also have impact upon the patients' dependent variable. Searching for such bias in a casewise fashion is subject to the chance selection of test variables. A preferable method seems to be to create the theoretical or predictive model ultimately to be tested, and then to inspect for missing value bias using all model variables, both individually and using a listwise method. Furthermore, a cross tabulation of the dependent variable by chart completeness having a significant chi-square would indicate a failure of the missing-at-random assumption. If, however, there is a high proportion of missing values on the dependent variable, results of such a cross tabulation would be inconclusive. Of additional aid would be the comparison of variables and relationship as they exist within the listwise subsample, the full sample, and any population parameters available.

REFERENCES

BABBIE, E. R. (1986) Survey Research Methods. Belmont, CA: Wadsworth.

BEALE, E.M.L. and R.J.A. LITTLE (1975) "Missing values in multivariate analysis." J. of the Royal Statistical Society 3: 129-146.

BRADBURN, N. M. and S. SUDMAN (1979) Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass.

BRIDGE, R. G. (1974) Nonresponse Bias in Mail Surveys: The Case of the Department of Defense Post-Service Survey. Defense Advanced Research Project Agency, ARPA, R-1501.

DILLMAN, D. A. (1978) Mail and Telephone Surveys: The Total Design Method. New York: John Wiley.

ESROV, L. V. (1981) "Assuring data quality in services evaluation," pp. 23-39 in P. M. Worthman (ed.) Methods for Evaluating Health Services. Newbury Park, CA: Sage.

GLEASON, T. C. and R. STAELIN (1975) "A proposal for handling missing data." Psychometrika 40: 229-252.

HAITOVSKY, Y. (1968) "Missing data in regression analysis." J. of the Royal Statistical Society Series B 30: 67-82.

HAWKINS, D. F. (1975) "Estimation of nonresponse bias." Soc. Methods and Research 3: 461-485.

HENDRICKSON, L. and J. MYERS (1973) "Some sources and potential consequences of errors in medical data recording." Methods of Information Medicine 12: 38-45.

MARINI, M. M., A. R. OLSEN, and D. B. RUBIN (1979) "Maximum-likelihood estimation in panel studies with missing data," pp. 314-357 in K. F. Schuessler (ed.) Sociological Methodology 1980. San Francisco: Jossey-Bass.

MURNAGHAN, J. H. and K. L. WHITE (1971) "Hospital patient statistics, problems, and prospects." New England J. of Medicine 284: 822-828.

RUBIN, D. B. (1974) "Characterizing the estimation of parameters in incomplete-data problems." J. of the Amer. Statistical Assn. 69: 467-474.

RUBIN, D. B. (1976a) "Inference and missing data." Biometrika 63: 581-592.

RUBIN, D. B. (1976b) "Comparing regressions when some predictor values are missing." Technometrics 18: 201-206.

RUBIN, D. B. (1977) "Formalizing subjective notions about the effect of non-respondents in sample surveys." J. of the Amer. Statistical Assn. 72: 538-543.

SCHUMAN, H. and S. PRESSER (1981) Questions and Answers in Attitude Surveys. New York: Academic Press.