IMPLICATIONS FOR ESTIMATION OF PROGRAM IMPACTS WHEN BASELINE DATA ARE COLLECTED DIFFERENTLY FROM TREATMENT AND CONTROL GROUPS

Randall S. Brown, Mathematica Policy Research, Inc.
Peter A. Mossel, Columbia University

The National Long Term Care (channeling) Demonstration was established by the U.S. Department of Health and Human Services to evaluate community-based approaches to long term care for the elderly. Specifically, the channeling demonstration is testing two models of organizing community care as alternatives to the current institutionally oriented system. Both offer a central point of intake for individuals in need, systematic assessment of their needs, and ongoing case management to arrange and monitor the provision of services. The basic case management model is designed to manage services currently available to clients; the financial control model is intended to expand the range of publicly financed services available to the client while controlling total costs. Through contracts with the participating states, local agencies in ten communities around the country were selected to implement the demonstration, five implementing each model. The demonstration is designed to determine (1) the impact of these approaches on costs, utilization of services (especially hospitals and nursing homes), informal caregivers, and client well-being; (2) the feasibility of implementing future programs like channeling; and (3) its cost-effectiveness.

In order to maximize the likelihood of obtaining reliable estimates of channeling impacts, an experimental design was used, under which eligible channeling applicants in each of the 10 sites were randomly assigned to the treatment group, which was offered channeling services, or to the control group, which was not. Under the design, estimates of program impacts are obtained by comparison of the post-randomization experience of the two groups.

One aspect of the evaluation design which could, however, raise questions about the accuracy of the estimates of channeling impacts that eventually will be obtained is the fact that the baseline data were collected by different types of interviewers for the two groups. The combination of several factors--conflicts between research needs and good case management practices, budget constraints, and the desire to minimize the burden on sample members--led to the decision that baseline data would be collected by channeling staff for members of the treatment group, and by research interviewers for the control group. For a variety of reasons, this difference in data collection could result in differences between the two groups on observed data for some characteristics, when in fact no real differences exist between the two groups on these baseline characteristics. Estimates of channeling impacts that are obtained from regression models which use these baseline data as auxiliary control variables could then be distorted, because these artificial differences between the two groups are treated as real pre-treatment differences that must be accounted for (netted out) by the statistical procedure.

The purposes of this paper are to determine whether the baseline data for treatments and controls are comparable and, if they are not comparable, what should be done to ensure that regression estimates of channeling impacts are not biased by such differences.

I. THE DATA

To understand the baseline data collection process, some background is first necessary. Elderly individuals in the ten demonstration sites who were referred to channeling were given a screening interview by channeling intake workers (usually by telephone) to assess their eligibility. The screen included questions on the individual's functional ability, need for assistance with various personal care or household activities, and other variables. In addition, there were questions on income, ethnicity, sex, recent hospital and nursing home use, and cognitive impairment. Eligible individuals were then randomly assigned to treatment or control status by research staff.

The screen interview does not, however, contain the comprehensive data that were necessary for either the evaluation or the development of a care plan for channeling clients. A thorough, in-person baseline assessment of treatment group members was required in order for program case managers to develop an appropriate care plan for participants. Because there was a great deal of overlap between the types of data required for research and care planning purposes and because the interview represented a substantial burden on the frail respondents (and on project resources), a single instrument that would serve both purposes was developed. It was considered important that channeling staff members collect the data necessary for developing an appropriate care plan; thus, having the baseline administered by research interviewers for both groups was ruled out. Having channeling staff conduct the baseline interview for both groups was also considered but was rejected because of higher costs, the excessive burden on channeling staff (which could affect the quality of services provided to treatment group members), and the possibility that assessment of the control group by program staff would "contaminate" the control group. Thus, the baseline interview was administered by channeling staff for the treatment group, and by research interviewers for the control group.

The difference between channeling staff and research interviewers in background and experience was considerable. Typically, the case manager assigned to a treatment group member conducted the baseline interview. Case managers were typically social workers or nurses, many with several years of experience working with the elderly. Control group members, on the other hand, were interviewed at baseline by trained interviewers, frequently individuals with substantial experience in conducting research interviews, but no particular knowledge of or experience with assessment for the frail elderly population.

In order to minimize the risk that these differences in background might cause the data for treatments and controls to be noncomparable, research interviewers and channeling staff who would be conducting baseline interviews received similar training in how to administer the baseline, using the training manuals developed by the survey research staff. Nevertheless, the goals of a clinical assessment and a research interview are inherently different, and these differences in emphasis can result in spurious differences in the data that are collected. Data collected by research interviewers are not necessarily better (i.e., more "correct") nor worse than data collected by channeling staff. It is only whether they are different that matters for the analysis. The data collected by channeling staff and by research interviewers could differ, even if the sample members in the two groups did not, for several reasons:

o Incentives of treatment group respondents to misreport their need for or ability to pay for services (since their responses will influence the services the program will try to arrange for them)

- Different backgrounds of research interviewers and channeling staff

- Differences between channeling and survey research in implementing the survey
  - differences in the length of time between screen and baseline
  - different use of proxy respondents
  - differences in the amount of previous experience in administering the baseline

The last three reasons require some further elaboration. Comparing the frequency distributions for the length of time between screen and baseline for treatment and control groups show a very apparent difference between the two groups. The median length of time between screen and baseline was 7.4 days for the treatment group and 12.5 days for the control group, nearly twice as long. While a difference of five days is inconsequential for some variables (e.g., income, age, assets), it may be very significant for other variables for this population. Because many sample members were at a critical point at the time they were referred to channeling, their situation (i.e., unmet needs, ability to perform certain activities, hours of formal or informal care received that week) one week after the screen may have been quite different from their situation nearly two weeks after the screen. This is especially true for the substantial number of sample members (about 19 percent) who were in a hospital or nursing home at the screen, many of whom were soon to be discharged. The difference of a few days could greatly influence their responses, especially if it affected whether they were at home instead of in the hospital by the time of baseline.

We also examined the distribution of the treatment and control groups by use of proxy respondents at baseline. The results indicate that in both the basic and financial control models, someone in addition to the sample member was present at the baseline interview for 75 percent of the treatment group, compared to only 59 percent of the control group. The different mix of sample member and proxy respondents for the treatment and control groups could lead to differences in the data obtained.

Finally, it is also the case that on average, at the time of a given baseline, the interviewer administering the baseline to a control group member had conducted many more baselines than the channeling staff member administering the baseline to a treatment group member. Forty-five percent of the control group was given a baseline by an interviewer who already had conducted 50 or more previous baseline interviews, while this could be said of only 14 percent of the treatment group. The average number of baseline's previously completed by the person administering a given baseline was twice as high for the treatment group as for the control group (53 for the treatment group compared to 28 for the control group at baseline). This difference between treatment and controls on the average experience of the interviewer at baseline could affect the data collected.

II. RESEARCH PLAN AND COMPARISON OF SCREEN CHARACTERISTICS

Perhaps the first reference that occurs to the analyst when confronted with questions about the measurement of data items is the literature on measurement error, especially the work by Joreskog and others (e.g., see Joreskog, 1973). Joreskog's LISREL model was developed as a way of directly confronting the effects of measurement error by jointly modelling the regression equations of interest and the relationship between the measured and true values of regressors. However for a number of reasons the measurement problems faced here do not lend themselves well to a modeling approach of this type. First, the concern here is not with the effects of measurement error per se but rather with the effects of systematic differences in measurement for treatment and control groups. Second, the literature on measurement error is addressed to the problem of random measurement error. However, we are also (especially) concerned with systematic measurement error, e.g., a tendency to overreport actual needs by treatment group members, that differs for the two groups. Third, we have many potentially affected baseline variables, too many to model jointly. Fourth, there are a variety of mechanisms by which measurement may be noncomparable, not all of which imply that the measured difference will be distorted in the same direction. Fifth, many of the baseline variables are qualitative rather than continuous, as the LISREL model and measurement error literature specify. Sixth, even if a complex model could be specified to address these issues, the large number of dependent variables to be examined for evidence of program impacts form a system of equations that would be much too large to handle. Seventh, given the need to estimate subgroup impacts as well, we need a procedure that guarantees that truly comparable groups are being compared. These problems suggest that an econometric or modeling solution to the problem is infeasible. What is required is a procedure that will enable us to determine for each of a large set of variables whether the data are comparably measured for treatment and control groups.

The fact that the screen is administered to all members of the research sample and in a uniform manner prior to assignment to treatment or control status makes it an ideal source for assessing the comparability of the treatment and control groups before and after attrition at baseline. The approach that we have taken is as follows:

1. Use the screen data on the full sample to determine whether the randomization process produced comparable treatment and control groups.

2. Compare treatment and control groups on screen characteristics for baseline responders to determine whether differential attrition has taken place at baseline--i.e., should the baseline data be expected to be similar for the two groups or did attrition at baseline distort the equivalence of the two groups.

3. Develop ad hoc tests to determine whether specific baseline variables are likely to distort estimates of program impacts, exploiting the fact that some of these variables are measured at screen as well.

Below we first describe the results from the first two steps, then present the testing procedures and results.

A. THE EFFECTS OF ATTRITION ON COMPARABILITY OF THE TREATMENT AND CONTROL GROUPS

Comparison of treatment and control groups on screen characteristics showed that, as expected, randomization produced groups that were very similar to each other, once the unbalanced design (unequal distribution of treatments and controls across sites) is accounted for. The results are displayed in Table 1.

The next step in assessing the comparability of baseline data for the two groups is to determine whether there is differential attrition at baseline. This is assessed here in two ways: first, by comparing screen characteristics of treatments and controls for the baseline respondents only, to determine whether significant differences exist where none did for the full sample, and second, by estimating a model of the probability of attrition at baseline as a function of screen characteristics, separately for treatment and control groups, and testing whether different patterns of attrition

TABLE 1

TREATMENT/CONTROL DIFFERENCES ON SCREEN CHARACTERISTICS FOR FULL SCREEN SAMPLE AND FOR BASELINE RESPONDENTS

	Full Screen Sample				Baseline Respondents Only			
	Basic Case Management		Financial Control		Basic Case Management		Financial Control	
Screen Characteristics	Treatment Group	T/C Difference	Treatment Group	T/C Difference	Treatment Group	T/C Difference	Treatment Group	T/C Difference
Demographics								
Age (%):								
65 to 74 29.7	0.2	26.0	-1.3	29.9	0.2	25.8	-1.4	
75 to 79 21.9	0.0	19.7	-0.7	22.3	0.0	19.7	-0.9	
80 to 84 22.4	-1.2	25.1	0.3	21.6	-1.6	25.2	0.7	
85 and over	26.0	1.0	29.3	1.7	26.2	1.5	29.3	1.6
Mean age 79.1	0.1	80.1	0.3	79.1	0.2	80.1	0.3	
Male (percent)	28.6	0.0	29.2	1.6	27.7	-0.1	28.9	2.2
Ethnic Background (%):								
Black (not of Hispanic origin)	21.9	-1.8	23.2	-1.1	22.0	-3.2**	23.3	-2.6
Hispanic 1.9	0.0	5.2	0.0	2.0	-0.2	5.3	0.0	
White and other	76.2	1.7	71.6	1.2	76.0	3.4**	71.4	2.6
Financial Resources								
Income (%):								
Less than $500	57.9	-1.1	59.3	-0.3	58.8	-1.5	59.2	-1.7
$500 to $999	33.7	-0.2	35.4	1.9	32.9	0.1	35.5	3.1
$1,000 or more	8.4	1.3	5.4	-1.6*	8.3	1.4	5.3	-1.4
Mean monthly income	529.5	-13.0	508.5	-13.0	526.3	5.2	531.2	-10.1
Insurance Coverage (%):								
Medicare, not Medicaid	77.1	-0.8	76.9	0.1	76.7	-0.1	77.2	1.8
Medicaid 20.2	0.7	23.0	-0.1	20.7	0.4	22.8	-1.8	
Neither Medicare or Medicaid	2.7	0.1	0.1	0.0	2.7	-0.3	0.1	0.0
Living Arrangement								
Type of Living Arrangement (%):								
Nursing home or LTC facility	3.6	0.1	1.5	-0.1	3.5	0.0	1.5	0.0
Hospital 15.3	-1.7	25.1	-1.7	14.7	-1.5	24.3	-1.3	
Community	81.1	1.6	73.4	1.8	81.9	1.5	74.2	1.3
Community Living Arrangement (%):								
Alone 35.8	-0.7	38.9	-0.9	35.5	0.3	38.8	-1.1	
With spouse	31.1	1.8	31.4	1.4	30.8	1.3	31.3	1.4
With child, but not spouse	20.6	-0.8	18.8	-0.9	21.2	-0.7	19.2	-1.2
With others	12.5	-0.3	10.8	0.3	12.5	-0.2	10.7	0.9
Health and Functioning								
Activities of Daily Living (%):								
Impaired on eating	21.2	-1.5	26.4	-0.1	20.4	-2.6*	26.3	0.7
Impaired on transfer	53.8	-0.6	58.0	-0.7	53.1	-0.8	57.7	1.7
Impaired on toileting	56.2	-1.3	61.3	2.4	55.5	-2.2	60.8	2.1
Impaired on dressing	69.3	-0.9	71.8	0.2	68.6	-1.4	71.5	0.6
Impaired on bathing	90.0	-0.4	93.2	1.7*	89.9	-0.2	93.5	2.2**
Impaired on continence	59.1	0.2	57.5	-0.7	58.8	-0.4	57.3	-0.7
Cognitive Impairments Affecting Functioning (%)	58.7	0.7	60.0	0.3	58.2	0.9	60.1	1.0
Number of Unmet Needs (%):								
0-1 7.6	-0.3	3.9	-0.2	7.5	-0.6	3.8	0.1	
2-3 58.2	-0.8	66.2	-0.6	58.2	-0.6	67.2	0.0	
4-5 34.2	1.0	29.9	0.8	34.2	1.2	29.0	-0.1	
Mean number of unmet needs	3.0	0.0	3.0	0.0	3.0	0.0	3.0	0.0
Existing Care and Contacts								
Currently Receiving Help with Services (%):								
Meal preparation	68.2	-1.8	73.3	-2.2	67.8	-1.2	73.6	-1.3
Housework/shopping	72.9	-1.1	76.4	-1.5	72.7	-1.3	76.7	-1.1
Taking medicine	45.8	-1.2	51.9	-1.2	45.6	-0.7	51.8	-1.1
Medical treatments	29.8	1.5	37.8	-0.4	29.9	1.4	37.8	0.2
Personal care	61.4	-1.0	69.5	-2.7	61.3	0.3	70.0	-1.8
Proxy Use (%)	65.1	-0.4	67.5	-0.7	64.5	-0.7	67.3	-0.3
Applied for Admission to Nursing Home	11.0	1.3	7.2	0.3	10.4	0.6	7.1	0.5
Maximum Sample Size		3123		3202		2757		2870

NOTE: Treatment/control differences are estimated using multiple regression to control for the different distribution of the two groups across sites.

*Significantly different from zero statistically at the 10 percent significance level using a two-tailed test.
**Significantly different from zero statistically at the 5 percent significance level using a two-tailed test.

occurred for the two groups. The results of these two analyses are discussed below.

Response rates at baseline were considerably lower for the control group (about 83 percent) then for the treatment group (93 percent), in both models. However, that does not necessarily imply that different types of individuals drop out of the sample in the two groups. When treatment/ control differences are reestimated for only the portion of the sample responding at baseline, the results, displayed in Table 1, show treatment/control differences that are in general very similar to those found for the full sample. There are only two differences that are statistically significant for responders, percent black and percent impaired on eating, and neither differences is substantially greater for the responders than for the full sample. Thus, these results suggest that there is very little evidence of differential attrition in either channeling model. The two groups continue to be composed of comparable individuals in both models.

An alternative approach to assessing whether patterns of attrition were different for the two groups is to estimate a model of the probability of response at baseline as a function of screen characteristics, separately for treatment and control groups, and test for differences between the two sets of coefficients. Probit models of the probability of response at baseline as functions of screen characteristics were estimated and the tests performed. A likelihood ratio test showed that the hypothesis of equality of the two sets of coefficients was rejected at the .05 level but not at the .01 level. However, there were statistically significant differences in coefficients for the two groups for only 3 of the 31 variables in the model. Thus, the results generally support those contained in Table 1. (See Brown and Mossel, 1984 for details.)

B. COMPARISON OF TREATMENT AND CONTROL GROUPS ON BASELINE CHARACTERISTICS

The preceding section indicated that we should expect relatively few differences between treatment and control groups at baseline. To test this hypothesis, treatment/control differences in means on a variety of baseline variables are estimated, using the same regression model used above to control for the difference between the groups in the distribution of observations across sites.

The results, presented in Table 2, are striking. (To preserve space we present only those variables for which statistically significant differences were obtained.) Many of the treatment/control differences are statistically significant, frequently for both models. Also, the differences are often large and significant at even the 1 percent level. The results can be conveniently summarized by dividing the variables examined into three categories:

No Significant Differences in Either Model

Age
Sex
Marital status
Living arrangement
Days restricted to bed
Hours of informal care per month
Percent with formal care of various types
Number of physician visits
Nursing home use
Life satisfaction
Loneliness

Significant Differences in Only One Model

Education
Assets
Insurance
IADL
Mental functioning
Self-rating of health
Percent with home-delivered meals
Attitude toward nursing homes
Whether institutional baseline given

Significant Differences in Both Models

Percent black
Income
Sources of income
In hospital/nursing home
ADL
Medical conditions
Unmet needs
Receiving informal care
Number of informal visits per month
Treated for medical condition
Received case management
Hospital use
Social isolation (contacts per week)
Proxy use
Length of time to complete interview
Interviewer rating of reliability

Given that relatively minor differences between treatments and controls were expected, the number of variables for which significant differences are found in one or both models and the size of these differences suggest that the data for some variables may not be comparably measured for the two groups. Several aspects of the results support this inference:

1. Several variables for which significant treatment/control differences are observed at baseline exhibited no significant differences when the corresponding screen measure was examined for the same set of individuals.

2. Most of the demographic variables for which measurement differences in the data are unlikely (for any of the potential reasons given earlier) in fact exhibit no significant treatment/control differences at baseline.

3. The effect of the treatment/control difference in the length of time between randomization and baseline (presented earlier in Table 1) is obvious in some of these variables.

4. Most of the significant treatment/control differences on baseline variables are in the direction that might be expected if they were due to noncomparable data, based on the reasons given earlier for why data might be noncomparably measured.

C. THE REINTERVIEW SAMPLE

Finally, a small sample (400) of treatment group members were administered a second baseline, this time by research staff, a short time (two weeks, on average) after completing the original baseline given by program staff for the sole purpose of assessing the comparability of data collected by the two types of interviewers. Comparison of sample member's responses at baseline and reinterview indicated patterns of differences that were very similar to those obtained in Table 1. This special sample is not as valuable as might be supposed for assessing comparability, primarily because of the two week interval, and the fact that this comparison does not enable us to distinguish differences due to the unreliability of the questions from those due to interviewer differences. Nonetheless, as further evidence of the noncomparability of some data items, we note without

TABLE 2

TREATMENT GROUP MEANS AND TREATMENT/CONTROL DIFFERENCES ON BASELINE CHARACTERISTICS

	Basic Case Management Model		Financial Control Model	
Baseline Characteristics	Treatment Group	T/C Difference	Treatment Group	T/C Difference
General Demographics				
Ethnic Background (%):				
Black (not of Hispanic origin)	22.3	-2.9**	23.7	-2.7*
Hispanic	2.0	-0.8	5.2	-0.4
White or other	75.7	3.7***	71.1	3.1*
Financial Resources				
Total Monthly Income (%):				
Less than $500	54.8	-4.5**	51.8	-6.2***
Between $500 and $1,000	34.1	2.1	39.0	6.4***
Over $1,000	11.2	2.4**	9.2	-0.2
Mean monthly income	567.5	46.0***	571.9	32.7**
Living Arrangement				
Living Arrangement Type (%):				
LTC facility	3.7	-0.7	1.4	-1.0*
Hospital	8.3	2.2**	15.1	4.6***
Supportive housing	1.2	-0.9*	2.1	0.1
Community	86.9	-0.5	81.4	-3.7**
Health and Functioning				
Activities of Daily Living (ADL) (%):				
Impaired on eating	23.2	-3.7**	26.0	-2.0
Impaired on transfer	51.3	4.9**	53.8	4.6**
Impaired on toileting	54.6	0.8	57.8	3.5*
Impaired on dressing	59.4	3.0	61.5	5.1***
Impaired on bathing	77.6	-1.1	799.3	3.5**
Impaired on continence	52.2	1.2	53.1	4.5**
Medical Conditions (%):				
Life threatening	65.7	-4.3**	66.3	2.3
Chronic disabilities	92.9	-0.1	93.9	2.0**
Acute problems	15.2	-0.1	15.3	0.3
Number of Unmet Needs (%):				
(0-1)	23.9	-8.8***	13.4	-15.5**
(2-3)	30.1	3.8**	27.2	-1.1
(4-5)	46.0	5.0**	59.4	16.6***
Mean number of unmet needs	3.3	0.3***	4.0	0.8***
Current Utilization of Health and Social Services				
Help From Informal Providers:				
Average visits per month	18.4	-3.0**	19.7	2.6**
Average time per visit (hrs)	1.8	0.0	1.9	0.0
Total hours informal care per month	52.1	-1.9	51.1	-1.3
Services Arranged for by Case Management-Type Provider (%)	8.8	-13.2***	16.9	-13.4***
Hospital Use:				
Any admissions last six months (%)	63.1	-2.0	67.9	-3.8**
Number of admissions last six months	1.1	-0.1**	1.2	-0.1
Number of days (last two months)	9.4	-0.4	10.9	1.3**
Respondent Attitudes				
Number of social contacts per week (%)				
None	9.4	-0.8	10.2	4.4***
One	6.3	-3.5***	7.2	1.7
2-6	27.8	-2.1	27.1	-6.1***
7 or more	56.6	6.5***	55.5	0.0
Methodological				
Type of Respondent (%):				
Individual only	42.6	-4.4**	38.2	-2.0
Individual and proxy	30.7	7.5***	36.0	11.7***
Proxy only	26.7	-3.1*	25.7	-9.7***
Interview Completion Time (minutes)	80.0	12.9***	78.6	7.3***
Interviewer Rating of Reliability (%):				
Highly reliable	26.5	-20.2***	22.5	-20.3***
Moderately reliable	41.9	9.2	49.0	9.2***
Unreliable	26.7	11.0***	25.0	10.1***
Totally unreliable	4.9	0.0	3.5	1.0
Maximum Sample Size	1,638	2,757	1,815	2,870

NOTE: Treatment/control differences are estimated using multiple regression to control for the unequal distribution across sites of the two groups. Asterisks denote significantly different from zero at the .10 level (*), .05 level (**), or .01 level (***).

discussion the general agreement of results obtained on this sample with those cited earlier. The results and a discussion of the problems with this sample are presented in Brown and Mossel (1984).

III. A PROCEDURE FOR IDENTIFYING WHICH VARIABLES ARE NOT MEASURED COMPARABLY

The preceding section provided fairly strong evidence that some baseline variables are differently measured for treatment and control variables. Thus, we are left with the following choices for selecting a set of control variables for our regression models:

- o Use only the screen data (i.e., ignore all baseline data)
- o Use baseline data freely
- o Use variables from the baseline only if there are no discernable treatment/control differences that cannot be tied to differences on screen characteristics. For variables exhibiting an unexplained difference at baseline, substitute the corresponding variable from the screen if there is one.

Using only the screen would ensure that the control variables are comparably measured. However, this would result in the loss of some baseline variables for which we have no evidence or expectation of non-comparable measurements. Given that the baseline contains some variables that are not contained in the screen and, perhaps, provides more comprehensive and reliable measures of variables with screen counterparts, this seems an overly extreme approach.

On the other hand, ignoring the evidence of non-comparability seems unwise. The advantages of the baseline relative to the screen are not sufficiently great to justify introducing bias into the program impact estimates. Thus, we settled on the third approach.

The procedure for determining whether to include specific baseline variables as control variables in the outcome regressions was based on statistical tests. Two sets of tests are required, one for the set of variables which have screen counterparts and one for variables that do not. To motivate the test that were used, consider the following simple regression model for estimating channeling impacts:

$$(1) \qquad Y = a_B T_B + a_F T_F + b_1 S_1 + b_2 S_2 + \dots + b_{10} S_{10} + u_1,$$

where Y is an outcome variable that channeling is hypothesized to influence; T_B and T_F are binary variables equal to one for treatment group members in the basic and financial control sites, respectively; S_I is a binary variable equal to one for sample members in the i^{th} site; a_B and a_F are parameters that represent channeling's impact on Y in the two models; $b_1, \dots, b_{10}$ are the coefficients on the site variables; and u_1 is a random disturbance term. If the random assignment process produced equivalent treatment and control groups, and the comparability of these two groups was not distorted by attrition from the sample, regression estimates of a_B and a_F, which are easily shown to be simply weighted averages of the treatment/control differences in means at the five sites implementing the program, would provide unbiased estimates of channeling impacts.

In order to account for the (minor) differences observed between treatments and controls on screen characteristics and increase the precision of our estimates of a_B and a_F, we include additional explanatory variables in our regression model. Assuming these variables are drawn from the screen we have

$$(2) \qquad Y = a_B T_B + a_F T_F + Sb + Xc + u_2,$$

where X is a vector of screen characteristics (such as income, ADL impairment, living arrangement, unmet needs, age, race, sex, etc.), c is a vector of the corresponding coefficients, and S and b are vector representations of the binary site variables and their coefficients. The relationship between the regression estimate of a_B from equation (1) and the estimate from equation (2) is:

$$(3) \qquad \tilde{a}_B^S = \tilde{a}_B^L + P_{1T}\tilde{c}_1 + P_{2T}\tilde{c}_2 + \dots + P_{KT}\tilde{c}_K,$$

where $\tilde{a}_B^S$ is the estimate of a_B from equation (1) (the "short" regression), $\tilde{a}_B^L$ is the estimate of a_B from equation (2) (the "long" regression), $\tilde{c}_I$ is the regression estimate of c_I, and P_{iT} is the regression estimate of the coefficient on T_B in an auxiliary regression of X_i on T_B, T_F, and S. The estimate P_{iT} then is simply the estimate of the treatment/control difference in the screen characteristics X_i after accounting for the different distribution of the two groups across sites. These P_{iT}'s are the estimated differences reported in Table 1.

These P_{iT} terms help clarify the effects that noncomparable data have on estimates of channeling impacts. Suppose, for example, that treatments and controls had identical mean values on all of the variables in X. In this case all of the P_{iT}'s would be zero and $\tilde{a}_B^L$ would exactly equal $\tilde{a}_B^S$. Suppose that the regression was then reestimated, but with alternative measures of the characteristics in X substituted for the original ones, and that on these alternative measures there were differences in the way they were measured for treatments and controls, such that there was now a difference between the two groups in the mean values for the new X, even though the groups were actually equivalent. In this case, the P_{iT}'s would not be zero and $\tilde{a}_B^L$ would differ from the correct value, $\tilde{a}_B^S$.

This is essentially the situation faced in estimating channeling impacts with baseline variables rather than screen variables. Although the P_{iT}'s for screen variables are not exactly zero, they are small and almost never significantly different from zero. However, for the baseline version of these same (and other) variables, the site-adjusted treatment/ control differences (P_{iT}'s) are often large and frequently significant (as shown in Table 2). Thus, we would expect $\tilde{a}_B^L$ to differ more widely from $\tilde{a}_B^S$ when baseline measures are used as control variables than when the screen measures are used.

The expression in (3) makes clear the difficulty in deciding what should be done about baseline variables that exhibit treatment/control differences. If they represent real differences between treatments and controls on characteristics (due to differential attrition, say), it is important that they be controlled for. In this case, $\tilde{a}_B^L$ would be the correct estimate, and it would differ from $\tilde{a}_B^S$. On the other hand, if observed characteristics are different solely because of measurement differences, it is important not to use such variables in the regression because they will cause $\tilde{a}_B^L$ to differ from $\tilde{a}_B^S$, which is the correct estimate when there are no real treatment/control differences. The bias in impact estimates caused by deleting from the regression a control variable with real treatment/control differences is of exactly the same magnitude (but in the opposite direction) as the bias introduced by including in the regression control variables that differ only because of measurement differences.

For baseline variables that have a screen counterpart there is a reasonable solution to this problem: test whether the P_{iT}'s obtained for the baseline variables are significantly different from the P_{iT}'s obtained for the corresponding screen variables; i.e., test whether there is a statistically significant difference between the estimated screen treatment/control difference and the estimated baseline treatment/control difference. This is readily estimated for each variable by taking the observations with both measures available, subtracting the baseline value from the screen for each observation, and regressing this difference on T_B, T_F, and the binary site variables. If we reject the hypothesis that the coefficients on T_B and T_F in this regression are equal to zero, then the screen measure will be used; failure to reject this hypothesis suggests that the baseline differences are not so different from the screen differences and therefore that the baseline data can be used for the variable being examined.

Once this set of variables for which screen counterparts exist have been examined and the decisions made regarding whether the screen or baseline measure will be used, we can then use a related procedure to determine which of the baseline variables that have no screen counterparts should be retained. Without these additional variables the regression that would be estimated is:

$$(4) \qquad Y = a_B T_B + a_F T_F + Sb + X^* c + u_3,$$

where X^* is the set of screen or (comparable) baseline variables selected in the first step described above. Adding additional variables from the baseline (those without screen counterparts) would yield the model

$$(5) \qquad Y = a_B T_B + a_F T_F + Sb + X^* c + Zd + u_4,$$

where Z is the set of baseline variables without screen counterparts. Using the same breakdown as employed earlier, the relationship between the estimates of a_B from equations (4) and (5) is

$$(6) \qquad \tilde{a}_B^S = \tilde{a}_B^L + Q_{1T}\tilde{d}_1 + Q_{2T}\tilde{d}_2 + \dots + Q_{MT}\tilde{d}_M,$$

where Q_{iT} is the coefficient on T_B from an auxiliary regression of Z_i on T_B, T_F, S, and X^*. If there are no treatment/control differences in Z_i that are not explained by treatment/control differences in the site distributions or in X^*, then Q_{iT} should be close to zero and the estimate of treatment effects is expected to be relatively unaffected by including Z_i (i.e., $\tilde{a}_B^S$ will be roughly equal to $\tilde{a}_B^L$). Given our earlier conclusion that major differences in Z are due to noncomparable measures, we will exclude from the set of control variables those variables Z_i for which we reject the hypothesis that Q_{iT} equals zero, and will include (retain) Z_i if this hypothesis cannot be rejected.

With these decision rules, the set of admissible control variables can be selected. Before turning to the results of this selection process, however, two additional technical details about how the tests are to be conducted must be made clear.

The first point is determination of the significance level to use in the tests of whether P_{iT} and Q_{iT} are equal to zero. Our goal is to have criteria for selection of variables which are unlikely to result in the inclusion of baseline variables that are not comparably measured for treatments and controls. This suggests that contrary to the usual case, the conservative approach is to use a significance level higher than normal. However, making the significance level too high, given the large sample sizes, would mean that even trivial differences that were probably due to chance would result in rejection of the hypothesis of equality. Therefore, we conduct the tests at the .10 significance level, and indicate where use of a .20 level would lead to different conclusions. In such cases, the decision about whether to include or exclude the variable will rest on corroborating evidence and a priori expectations about the likelihood of data noncomparability for the specific variable being examined.

The second technical point is that we need testing criteria that will not lead to differences across models with regard to conclusions about which baseline variables are comparable. Although it is possible that the baseline data on a given variable are comparable for one model but not for the other, this seems unlikely. Moreover, allowing the set of control variables to differ by model would double the computational burden and would call into question whether any observed differences between the models in estimated impacts are due to the different regression specifications rather than to actual differences in the effects of the two channeling models. Thus, baseline variables will not be considered to be comparable (and therefore not considered usable as control variables) unless the tests of equality of means indicate no difference between treatments and controls for either model. This approach is again consistent with the basic strategy of being conservative with respect to including baseline variables that may not be comparably measured.

IV. RESULTS

In this section the baseline variables used in a preliminary report on channeling impacts at 6-month follow-up are tested for whether they pass the criteria established above. Results are presented separately for baseline variables with screen counterparts and those without such counterparts.

A. BASELINE VARIABLES WITH SCREEN COUNTERPARTS

For baseline variables that have screen counterparts, the availability of a measure that is known to be comparable for treatments and controls provides a fairly firm basis for assessing data comparability. Above we showed that treatment/control differences at screen for baseline respondents were statistically insignificant for all but a few of the variables examined, but that baseline versions for many of these variables exhibited differences that were statistically significant. The important question is whether the treatment/control differences on individual variables at baseline, although significant, are really substantially different from the treatment/control differences reported at screen, and therefore would have a substantially different effect on regression estimates with channeling impacts.

Tests were performed for all of the baseline variables used as control variables in our preliminary analysis of channeling impacts (see Kemper et al., 1984) that have comparable screen measures. Each of these characteristics is represented by a categorical variable with a discrete set of possible values and is converted into a set of mutually exclusive and exhaustive binary variables. For each characteristic, inferences are based on multivariate F-tests of whether the set of coefficients on treatment status are jointly equal to zero. Furthermore, since we will want to select the same version (screen or baseline) for certain groups of related variables (e.g., those measuring impairments), these sets of variables are each tested with joint F-tests.

The coefficients on treatment status in these auxiliary regressions and the corresponding t-statistics are not presented here in order to save space, but are available from the authors. The significance level of the F-statistics for the joint test of whether all treatment/control differences on a given characteristic are equal to zero are presented in Table 3.

Using a 10 percent level of significance as our criterion, we find significant treatment/control differences in screen-baseline differences for either the basic or financial control model on the following variables:

TABLE 3

TESTS OF TREATMENT/CONTROL EQUIVALENCE ON SCREEN-BASELINE DIFFERENCES AND REINTERVIEW SAMPLE RESULTS

	Significance Level of F-Statistic		Baseline-Reinterview Comparison	
Baseline Characteristics	Basic Model	Financial Control Model	Percent of Sample with Differences	Significance Level for Test of Symmetry
VARIABLES WITH SCREEN COUNTERPARTS				
Age	0.64	0.46	5.3	0.19
Sex	0.45	0.48	1.0	1.00
Ethnicity	0.04	0.22	0.8	0.19
Income	0.06	0.01	7.5	0.46
Insurance	0.85	0.29	5.0	0.82
Hospital/Nursing Home Occupancy	0.03	0.00	13.7	0.03
Nature of Living Arrangement	0.47	0.17	7.8	0.72
Activities for Daily Living	0.01	0.03		
Impaired on eating			13.1	0.72
Impaired on transfer			18.3	0.04
Impaired on toileting			17.8	0.24
Impaired on dressing			12.7	0.12
Impaired on bathing			15.4	0.07
Continence			22.6	0.75
Help with IADL Tasks	0.71	0.06		
Preparing meals			11.9	0.77
Housework/shopping			3.3	0.58
Taking medicine			14.8	0.69
Medical treatments at home			26.5	0.44
Unmet Needs	0.00	0.00		
Meal preparation			36.5	0.00
Housework/shopping			28.8	0.00
Taking medicine			20.4	0.15
Medical treatments at home			13.9	0.79
Personal care			35.0	0.00
Nursing Home Application	0.21	0.65	6.1	0.82
VARIABLES WITHOUT SCREEN COUNTERPARTS				
Educational Background	0.72	0.02	11.9	0.44
Assets	0.65	0.00	28.2	0.92
Home Ownership	0.44	0.49	2.5	0.75
Impairment on Instrumental Activities of Daily Living (IADL)	0.11	0.00		
Traveling			10.2	0.43
Money management			18.4	0.13
Telephone use			18.1	0.08
Short Portable Mental Status Questionnaire (SPMSQ)	0.07	0.27	42.6	0.00
Self-Rating of Overall Health	0.03	0.40	34.4	0.80
Needs More Help with Traveling	0.02	0.00	33.4	0.00
Stressful Life Events	0.74	0.94	15.5	0.90
Global Life Satisfaction	0.59	0.21	38.6	0.82
Attitude Toward Nursing Home	0.10	0.00	30.4	0.02

NOTE: Significance levels for F-tests are for tests of whether estimated treatment/control differences for all subcategories of each variable are jointly equal to zero. The symmetry test for categorical variables is for whether baseline-reinterview differences in one direction are comparable in size to differences in the opposite direction.

- income
- hospital/nursing home/community occupancy
- ADL
- continence
- unmet needs
- ethnicity
- IADL

Significant differences were found for both models for income, hospital/nursing home occupancy, ADL, and unmet needs.

For the remaining variables, i.e.:

- age
- sex
- insurance
- nature of living arrangement
- nursing home application,

no significant differentials were found in screen-baseline differences at the 10 percent level, nor, with the exception of one variable for one model, at the 20 percent level. Thus, the conclusions are essentially unaffected by the choice of significance levels.

To determine whether these results make intuitive sense, consider the likelihood that each of these variables would be affected by the five reasons given on page 2. One would not expect to find major treatment/control differences in age, sex, insurance, nursing home application, or living arrangement at baseline if none were found at the screen for any of these reasons, and none were found. Thus, the test results that indicate no problems with noncomparability are in accord with expectations.

The test results indicating that the data on other variables are not comparable are also in general agreement with prior expectations. For example, the larger screen-baseline differences observed for treatments than for controls on ADL and IADL impairments and on unmet needs are consistent with channeling clients' incentives to overreport needs and impairments. On the other hand, the income results are less clearly in agreement with expectations. The larger decrease from screen to baseline observed for the treatment group than for controls in the percent with income under 500 dollars is inconsistent with the expectation that treatment group members may have a greater incentive to underreport income. On the other hand, there are other factors that could cause the income data to be noncomparable that do not necessarily imply that the difference would be in a particular direction (e.g., reluctance of channeling staff to probe on questions they view as intrusive; different use of proxy respondents).

The other baseline variables on which treatment/control differences were found include whether in a hospital or nursing home at the time of interview and ethnicity. Hospital/nursing home occupancy differences are clearly due to the treatment/control differences in the length of time between screen and baseline (for clients in a hospital or nursing home at screen, as noted earlier). The results for ethnic distribution are somewhat surprising since this is one variable that would seem to be less affected by the interviewer and clearly not affected by timing, proxy use, or incentives. However, for financial control sites, there are significant differences between the two groups in screen-baseline differences in the percent Hispanic, with the treatment group percentages the same (2.0 percent) whether screen or baseline measures are used, but the control group percentage changing from 2.2 percent to 2.8 percent when moving from the screen to the baseline measure. This problem occurs in virtually all of the 10 sites, and suggests that channeling staff, both at the screen and at baseline, may have been reluctant to explicitly probe for Hispanic origin and instead were more likely to guess (e.g., based on surname or appearance, both of which may be misleading) than research interviewers. Channeling staff also may have actually used the screen interview to fill in this information, rather than ask for it in the in-person baseline. Thus, even the measurement of ethnicity appears to be affected by the type of interviewer.

We also examined the reinterview sample for corroborating evidence of whether these variables were measured comparably, using a test of "symmetry" for these categorical variables. The symmetry test examines whether the number of cases classified in, say, Category A at baseline and Category B at reinterview is significantly different from the number of cases for which there are discrepancies between the two interviews in the opposite direction. (See Bishop et al., 1975, pp. 282-296 for a description of tests of symmetry.) Rejection of symmetry implies differential measurement between the two interviews, although we may also be concerned about cases where symmetry is not rejected but the proportion of cases classified differently by the two interviews is substantial.

The results, presented in Table 3 alongside the results from the F-tests, show general agreement between the two tests. The two exceptions are for IADL tasks and income. However, the proportion of cases with differences between baseline and reinterview is relatively high. Thus, we view the reinterview sample results as confirming evidence of our earlier results despite the problems of interpretation with this small sample.

B. VARIABLES WITHOUT SCREEN COUNTERPARTS

For baseline variables without screen counterparts we have less information on which to base our assessment of comparability. Therefore, as argued in Section III, we will base our conclusions about which variables suffer from comparability problems on the assumption that any treatment/control differences at baseline that are not due to differences at the screen indicate noncomparable data for treatments and controls. Therefore, we test the baseline variables without screen counterparts by regressing them on treatment status, site, and the variables with screen counterparts. In these regressions, the control variables are the baseline version of the variables with screen counterparts for those baseline variables determined in the previous section to be comparably measured, and the screen version of those variables for which the baseline version was found to be noncomparable. If the coefficients on treatment status in these regressions are not zero, then the impact estimates will be distorted (under the assumption of no real differences between treatments and controls beyond those explainable by differences at screen). Hence, variables for which the (set of) coefficients on treatment status are significantly different from zero will be considered noncomparable and will be excluded from the set of baseline control variables in regression analyses used to estimate the impacts of channeling.

The baseline variables examined here are those that were used as control variables in Kemper et al. (1984) and that have no screen counterparts. The results, contained in the lower panel of Table 3, indicate that, as expected from the results in Section II, a substantial number of these baseline variables may not be comparable. However, the evidence is not nearly so clear-cut as for the variables with screen counterparts. In many cases, the difference is statistically significant only for one of the two channeling models, and of the opposite sign from the difference for the other model. The following list summarizes these results ("same," "mixed," and "different" refer to whether estimated differences are in the same direction, a different direction, or both for the channeling models):

Significant Difference for Both Channeling Models

Unmet travel needs (same)
Attitude toward nursing home (same)

Significant Differences for Only One Model (B = Basic, F = Financial Control)

Education (F; same)
Assets (F; same)
IADL (F; mixed)
SPMSQ (B; same)
Medical conditions (B,F; mixed)
Self-rating of health (B; different)
Restricted days (F; different)
Hospital days (F; different)
Nursing home days (F; different)

Significant Differences for Neither Model

Home ownership
Stressful events
Hours of informal care
Hours of formal care
Physician visits
Global life satisfaction

For unmet travel needs and attitudes toward placement in a nursing home, the results are unambiguous. Unmet travel needs are greater for treatment group members, which is consistent with the incentive that exists for treatment group sample members or their proxies to overreport the number of unmet needs, but could be due to other factors as well. This result is also consistent with the findings on other unmet needs in the previous section of this chapter.

Treatments are more likely than controls to say they would not go to a nursing home. Again, this is consistent with the incentives of clients to overstate the strength of their antipathy for nursing homes and with their possible anticipation that channeling will enable them to remain in the community.

Other variables for which the results are unambiguous are those for which significant differences were found for neither channeling model. For all six such variables, changing the testing criteria to 20 percent significance levels would have led to the same conclusion of no significant differences between treatments and controls in either model. Thus, again the results are insensitive to the choice of significance level used. Home ownership, stressful events, hours of formal and informal care, and physician visits are all objective rather than subjective phenomena and thus the measures are perhaps less likely to be affected by the factors identified earlier that lead to noncomparable data.

The list of variables for which significant differences are found for only one of the channeling models contains a mixture of variables for which we might or might not expect problems with comparability. All of these variables may be affected by differential use of proxies, which could lead to treatment/control differences in either direction. The difference between the two groups in the length of time between screen and baseline could also have caused the data to be noncomparable for some variables (e.g., SPMSQ, IADL impairment, hospital days). Some of these and other variables may also be affected by incentives of sample members to overreport (IADL impairment) or underreport (assets), or by the different backgrounds of the channeling staff and research interviewers. The mixture of reasons for why differences might arise and the uncertainty about the expected direction of differences are consistent with the ambiguous results obtained.

In order to help guide our decision on these variables, we again examine the reinterview sample. For discrete variables, the results presented in the last two columns of Table 3 indicate whether the differences between baseline and reinterview samples are symmetric, and the percent of cases for which baseline/reinterview differences are obtained, respectively.

For the two variables that seemed to point unambiguously toward comparability problems, and for the variables that regression results indicated were unambiguously free of comparability problems, the reinterview sample results again tend to confirm the findings. For the more ambiguous results in Table 3, the reinterview sample results tend to indicate that for most of these variables there is little evidence of systematic differences, but a substantial amount of absolute differences. For example, even for an objective concept such as assets, which does not usually change substantially in the span of a week, 28 percent of the reinterview sample gave a different categorical response at reinterview from that given at baseline. The self-rating of health status was different for over one-third of the sample (34 percent), and even education was differently reported for 12 percent of the sample. All of these results could be due to differential use of proxy respondents at baseline and reinterview.

The large number of discrepancies between the two responses suggests that the data are either differently measured for the groups (though perhaps not systematically in one direction), or contain so much random measurement error as to render these variables relatively useless as control variables at best and a potential source of bias in impact estimates at worst. Bias would be especially likely if the variable were used to create subgroups. Thus, we adhere to the criteria established in Section III; i.e., variables with a significant coefficient on treatment status for either model in the auxiliary regressions should be excluded as a control variable from future regression analyses of channeling impacts.

V. CONCLUSION

Occasionally, there are pressures, budgetary or other, to collect data differently for treatment and control groups in an experimental design. Our results suggest that this is likely to be a mistake in most cases. We find evidence of substantial noncomparability of the data collected for treatment and control groups at baseline, despite the considerable care taken to minimize it. If data cannot be collected by the same type of interviewer in the same way over the same time frame for the experimental and control groups, serious thought should be given to not collecting it at all, especially for data which are likely to be influenced by small timing differences or for which the experimental subject has an incentive to misreport.

REFERENCES

Bishop, Y.M.M.; Fienberg, S.E., and P.W. Holland. Discrete Multivariate Analysis. Cambridge, MA: MIT Press, 1975.

Brown, R. and P. Mossel. "Examination of the Equivalence of Treatment and Control Groups and the Comparability of Baseline Data." Channeling Evaluation Supplemental Report Number 3, prepared for Department of Health and Human Services, October 1984.

Joreskog, K.G. "A General Method for Estimating a Linear Structural Equation System." In A.S. Goldberger and O.D. Duncan (Eds.): Structured Equation Models in the Social Sciences. New York: Seminar Press, 85-112, 1973.

Kemper, P. et al. Channeling Effects for an Early Sample at 6-Month Follow-up. Channeling Evaluation Preliminary Report Number 2. Princeton, NJ: Mathematica Policy Research, June 1984.