The Findings of the Rand Health Insurance Experiment: A Response to Welch et al.

Author(s): Joseph P. Newhouse, Willard G. Manning, Naihua Duan, Carl N. Morris, Emmett B. Keeler, Arleen Leibowitz, M. Susan Marquis, William H. Rogers, Allyson R. Davies, Kathleen N. Lohr, John E. Ware, Jr. and Robert H. Brook

Source: *Medical Care*, Feb., 1987, Vol. 25, No. 2 (Feb., 1987), pp. 157-179

Published by: Lippincott Williams & Wilkins

Stable URL: https://www.jstor.org/stable/3765515

MEDICAL CARE
February 1987, Vol. 25, No. 2

The Findings of the Rand Health Insurance Experiment—A Response to Welch et al.

JOSEPH P. NEWHOUSE,* WILLARD G. MANNING,* NAIHUA DUAN,* CARL N. MORRIS,¶
EMMETT B. KEELER,* ARLEEN LEIBOWITZ,* M. SUSAN MARQUIS,§
WILLIAM H. ROGERS,* ALLYSON R. DAVIES,† KATHLEEN N. LOHR,§
JOHN E. WARE, JR.,† AND ROBERT H. BROOK‡

Welch et al. appear to raise a prodigious number of issues. But their critique ultimately succeeds or fails on the basis of its central implication, which is, to quote them, "whether the differences in health care expenditures and hospitalization between health insurance plans claimed by Rand have, in fact, been shown."

Fortunately, it is easy to demonstrate that these differences exist without becoming mired in a morass of detail. The simplest possible data—per-person use and expenditure in the various insurance plans—provide overwhelming evidence that the differences among plans that we have elsewhere described are real and substantial (Tables 1 and 2).[1–10] In general, we can reject the null hypotheses that there are no differences in use and expenditure with extremely high confidence; in three of the measures the chances are less than one in a trillion that this variation arose by chance! Moreover, the raw data are available in public use files, so that any interested person can verify our findings.

Virtually all of Welch et al.'s critique becomes irrelevant if the differences shown in Tables 1 and 2 are not caused by biased refusal to enroll by plan, biased attrition by plan, biased reporting of data by plan, or methods ("Hawthorne") effects. (Welch et al. do not raise the latter two issues, but they are of obvious importance.) We take up the issues of refusal, attrition, reporting, and methods effects in the next section, to which many readers may wish to skip; the remainder of this section elaborates on the data in Tables 1 and 2.

The columns in Table 1 are ordered by the "noise" in the utilization variable, with noise measured as the coefficient of variation in the free care plan (e.g., .009 = .8/86.8). The chi-square statistics testing the null hypotheses of no difference among the plans fall monotonically as the noise index rises, as would be expected if in fact there were true differences. We also show the probability with which we can reject the null hypothesis of no difference between the free and 95% plans, the two most extreme plans. Even in the case of inpatient expenditure, the noisiest variable, the free-95% contrast is significant at the .03 level using a one-tail *t*-test, which we believe is justified for this comparison.

In general, use and expenditures decline as the coinsurance rate or percentage paid

From the *Economics and Statistics Department, The Rand Corporation; †Behavioral Sciences Department; ‡Systems Sciences Department, The Rand Corporation; §Washington Office, The Rand Corporation. ¶Department of Mathematics, University of Texas.

Preparation of this response was supported by Grant 016B80 from the Department of Health and Human Services.

The views expressed are those of the authors and are not necessarily those of the Rand Corporation of the Department of Health and Human Services.

Address correspondence to: Joseph P. Newhouse, The Rand Corporation, 1700 Main St., Santa Monica, CA 90406.

TABLE 1. Annual Use of Medical Services Per Capita, By Plan[a] (Standard errors in parentheses)

Plan	Likelihood of Any Use (%)	Outpatient Expenditures (1984 $)	Face-to-Face Visits	One or More Admissions (%)	Total Expenditure (1984 $)	Total Admissions	Inpatient Expenditure (1984 $)	Number of Person Years
Free	86.8	340	4.55	10.3	749	.128	409	6822
	(.8)	(11)	(.17)	(0.45)	(38.7)	(.0070)	(32)	
25%	78.7	260	3.33	8.4	634	.105	373	4065
	(1.4)	(15)	(.19)	(0.61)	(52.8)	(.0070)	(43)	
50%	77.2	223	3.03	7.2	674	.092	450	1401
	(2.3)	(17)	(.22)	(0.77)	(143.5)	(.0166)	(139)	
95%	67.7	203	2.73	7.9	518	.099	315	3727
	(1.8)	(12)	(.18)	(0.55)	(44.8)	(.0078)	(37)	
- - - - - - -								
Individual deductible	72.3	235	3.02	9.6	608	.115	373	4175
	(1.5)	(12)	(.17)	(0.55)	(46.0)	(.0076)	(42)	
$\chi^2(4)$[b]	144.7	85.3	68.8	19.5	15.9	11.7	4.1	
P-value for chi-squared	2.8×10^{-30}	1.3×10^{-17}	4.1×10^{-14}	.0006	.003	.02	n.s.	
P-value for free vs. 95 percent contrast[c]	1.6×10^{-22}	2.0×10^{-17}	1.0×10^{-13}	.00037	.000051	.0028	.027	
Noise Index[d]	.009	.032	.037	.044	.052	.055	.078	

[a] The values in the table are sample means by plan. All standard errors are corrected for intertemporal and intrafamily correlation using an approach due to Huber; references 3 and 4 give details of this method.[1] Dollars are expressed in June 1984 dollars. Visits are face-to-face contacts with MD, DO, or other health providers and exclude visits for only radiology, anesthesiology, or pathology services. All data exclude dental services and outpatient psychotherapy services. The sample includes children born into the study except for the year of birth and excludes partial years except for deaths. The data come from reference 2.

[b] Testing null hypothesis of no difference among plans.

[c] p value comes from 1-tail t-test.

[d] Value shown is coefficient of variation in free care plan.

out of pocket increases. (Recall that the Individual Deductible Plan has free inpatient care; so it does not order naturally with respect to the 25, 50, and 95% plans.) The one exception, the 50% plan, is easily explained by noise in the data, as can be inferred from the large standard errors (SE) associated with the means for that plan. Even for the 50% plan, however, the results are consistent with monotonicity among the four "pure" coinsurance plans, being monotone in the three least noisy cases, less than the midpoint of the 25 and 95% plans in two cases (by 0.6 and 1.2 SE), and greater in two cases (by 0.7 and 0.8 SE).

The estimated magnitudes of the differences in spending among the plans in Table 1 do not generally differ from estimates derived from the results of the four-part model that Welch et al. question.[2] (Mean expenditure on the 50% plan, however, falls between mean expenditure on the 25 and 95% plans when using the four-part model.) Although we have shown that the four-part model estimates are more precise (specifically, have lower mean square error) and hence are preferable to the sample means shown in Tables 1 and 2,[11] for policy purposes it does not matter whether one uses sample means or the four-part model estimates. Also note that the values in Table 1, which are based on data from all sites and all years, are not very different (except for inflation) from values based on the first 40% of the data pub-

TABLE 2. Annual Use of Medical Services Per Capita, Seattle Sample, by HMO and FFS Status[a] (SE in parentheses)

Plan	Likelihood of Any Use (%)	One or More Admissions (%)	Imputed Expenditures with no Covariates[b] (1983 dollars)	Imputed Expenditures with Age-Sex Covariates[b,c] (1983 dollars)	Person Years
HMO Experimental	87.0 (1.0)	7.1 (0.50)	434 (28)	426 (23)	3687
HMO Control	91.1 (0.8)	6.4 (0.55)	432 (34)	465 (47)	2596
Fee-for-service Free	85.3 (1.6)	11.2 (1.17)	640 (81)	612 (66)	1221
T-statistic on free-experimental difference[d]	−0.88	3.24	2.41	2.69	
P-value for t-statistic (2-tail)	n.s.	.0012	.016	.007	

[a] The sample includes participants while they remained in the Seattle area. The sample excludes children born into the study and excludes partial years except for deaths. For HMO Controls and Experimentals, the data include both in-and out-of-plan use. The standard errors are corrected for intertemporal and intrafamily correlation using an approach due to Huber; see references 3 and 4 for details. These numbers differ slightly from those in references 3 and 4 because of minor corrections in the data as well as the use of a less precise but more robust method for calculating standard errors.

[b] See reference 2 for details of imputation method.

[c] The values with age and sex covariates come from using age and sex, along with plan dummies, as covariates in a four-equation model. See reference 4 for further details.

[d] Testing null hypothesis of no difference between HMO Experimental and Free Fee-for-Service plan.

lished in 1981.[5] In the interests of simplicity, the data in Table 1 have not been adjusted for the minor imbalance of plans across sites, but such adjustment scarcely affects the estimates in Table 1, save for the 50% plan, which, after adjustment, falls between the 25 and 95% plans means.[2,12]

The major comparison of interest in Table 2 is between the HMO Experimental Group and the FFS free-care plan. (Welch et al. mistakenly argue that this is not the major comparison of interest; we examine their argument below.) Table 2 shows that the annual percentage admitted to a hospital at the HMO was 36% below the rate in the free FFS plan in Seattle, whereas the likelihood of any use was virtually identical. This reduction in hospitalization is estimated to save 30% in expenditure using the sample means and 28% using the four-part model with only age and sex covariates. (The values reported in references 4 and 7 are those with age and sex covariates.)

The values in Tables 1 and 2 exclude data from individuals with only partial years of participation (partial years represent 3% of total years in Table 1 and 9% of the free FFS and HMO experimental years in Table 2). Such individuals have generally been excluded from the sample for convenience (in particular, the four-part model can only be used with full-year observations because it does not convolute), but their rate of spending does not differ significantly from those with full years of participation, so their inclusion or exclusion does not much affect the estimated mean rate of expenditure. As pointed out in reference 4, however, including these individuals and using ANOVA on raw expenditures reduces the *P*-value on the

free FFS-HMO experimental comparison to 0.08 (two-tail test). This is approximately a lower bound on the *P*-value, because it makes use of no covariates and no transformations and arguably a one-tail test should be used instead of a two-tail test. In any event it is virtually inconceivable that there could be a reduction of more than one third in admissions and approximately the same use of outpatient services without an appreciable reduction in expenditure.

Refusal of the Enrollment Offer, Attrition, Reporting and Methods Effects

We now examine various possibilities of bias in the sample means reported in Tables 1 and 2. Those means are based on a random sample of families in six sites, with some groups (most notably the aged) ineligible and a mild oversampling of the poor, except in Seattle, where there was a proportional sampling.[4,6]

Bias could, of course, arise if the groups in various plans were dissimilar because of nonrandom refusal of the offer to enroll in the experiment or nonrandom withdrawal (attrition) from the experiment. Although the rates of refusal and attrition were both significantly lower on the free plan than on other plans, this would only cause bias if there were a difference in the characteristics of individuals who refused or withdrew on each plan (i.e., if refusal or attrition were nonrandom). For example, if sickly individuals made up a greater proportion of those who refused the offer to enroll in the cost-sharing plans than of those who refused the offer to enroll in the free plan, our estimates of plan differences would overstate the true differences.

A Priori Expectations

We will present data below showing that all measured biases resulting from differential acceptance of the offer by plan and differential attrition by plan are small. Such results are not surprising, because the design of the experiment assured that families on all plans were always better off financially from participating. In particular, families were paid amounts that exceeded the maximum loss they could suffer by changing from their prior insurance plan to the experimental plan. Furthermore, families had an incentive to remain in the experiment to completion because they were paid an amount for completing the experiment that was equal to the maximum possible annual loss. Thus, on any simple rational calculus, there was no reason not to participate in the FFS sample. (The alert reader may ask whether such payments affected the observed response. They did, but only negligibly; the true differences across plans in the absence of payments may be 1% larger than those we estimate.)

The HMO Experimental sample faced somewhat different incentives than the FFS sample because they were asked to change providers (more specifically, change the terms at which they could see their previous FFS providers). Their inducements to enroll and remain in the experiment were that (1) their care was free at the HMO; (2) they received free dental care in the FFS system; in addition to acting as an incentive, this served to keep the benefit package the same as the free FFS plan; (3) they were reimbursed for any premiums they were paying out-of-pocket for a FFS insurance policy in exchange for an assignment of benefits for that policy; in addition to acting as an incentive, the assignment of benefits assured that they were not using their old policy; and (4) they received $120 for completing the experiment. Because the HMO sample had to change the terms at which they could see their old providers, however, there is a less compelling a priori case of no bias than in the case of the FFS plans.

Findings on Refusal

As expected, the data on refusal suggest no important biases. Table 3 gives the mean values of several characteristics for those aged 14 and over at enrollment by FFS plan;

one "significant" difference (at the 5% level) would be expected among the 20 comparisons made if there were no true differences, and there is in fact one such difference (in the percent female). The data in Table 3 indicate good comparability among the plans at enrollment, especially for preexperimental visits and hospitalization rates, the most important measures for establishing subsequent differences in use. Similar results were reported for children.[13]

A statistically more powerful test of the null hypothesis of no bias from refusal is to compare those who accepted and refused the offer (Table 4). None of the F-values in the rightmost column is significant at the 5% level; thus, we cannot reject the null hypothesis that those who accepted and those who refused are similar by plan in terms of prior use and health status.

Analogous findings for the HMO sample are reported in Tables 5 and 6. Table 5 shows no significant differences among the three experimentally enrolled groups (GHC Experimental, Free FFS, and Pay FFS); the F-statistics testing the null hypothesis of no differences for both personal (e.g., health status) and family (e.g., income and family size) characteristics are both below 1. (Further detail is available in reference 7.) Table 6 does show statistically significant differences within the Seattle sample between those who accepted and refused the offer (among the four groups) with respect to income, the percentage with group insurance, and the percentage with public insurance. (If one were to make a correction for multiple comparisons, however, the significance of the latter two differences would be problematic.) Adjusting for these quantitatively minor differences would leave our qualitative findings virtually unchanged. For example, the estimated means by plan controlling for most of the covariates shown in Tables 5 and 6 (including income) and using the four-part model do not much differ from those in Table 2 (specifically, the reduction in imputed expenditure is 25% instead of 28%).

In addition to analyzing data on those who enrolled and refused, we compared data on the group that enrolled with the group that completed the baseline interview (our best estimate of the makeup of the population from which we sampled).[15] The group that completed the baseline interview includes not only those who refused the subsequent offer, but also those who were lost for other reasons between the time of the baseline interview and the time of enrollment (e.g., because the employer refused to verify the terms of the insurance policy). Because of the additional population included, this represents a more comprehensive test of whether bias was introduced in the enrollment process.

The results demonstrate that no important bias was introduced at this preenrollment stage (data not shown).[15] Indeed, because of our use of the Finite Selection Model,[16] a generalization of randomized, stratified sampling, the sample we enrolled was better matched with the sample that did not enroll than would be expected with simple random sampling. The one exception is that the percentage of children enrolled is higher than among those not enrolled but eligible for enrollment. Correcting for the higher percentage of children in our sample than in our sites' population would let us generalize to the population of our sites, but because that was not our purpose, we have not done so. We have, however, presented adult- and child-specific results,[4–7,10,12] so a correction would not be difficult to make from our published results.

To summarize the effect of refusal of the enrollment offer: No bias was detected in the FFS sample as a result of different refusal rates by plan. The few differences detected in the HMO sample between those who accepted and those who refused have negligible effects on the results.

Findings on Attrition

Bias could, in principle, also arise from differential attrition; if, for example, those with

TABLE 3. Values for Adults of Demographic and Health Status Measures At Enrollment and Exit, by Plan[a]

Variable and Brief Description[b]	Cost-Sharing Plans				Free Plan	t-Test Value[d] for Difference Between Free and Cost-Sharing Plans
	95%	25 and 50%	Individual Deduction	Total Cost-Sharing		
Number of enrollees ≥14 yr of age	759	1024	881	2664	1294	
Mean age (yr)	32.8	33.8	33.6	33.4	33.3	−0.0
Sex (% female)	56.1	53.5	53.8	54.4	52.2	−2.1
Race (% non-white)	20.8	17.4	18.3	18.9	16.6	−0.5
Mean family income adjusted for family size ($ 1982 thousands)	21.5	22.8	23.3	22.5	22.1	−0.5
Hospitalized in year before enrollment (%)	11.5	11.2	12.0	11.6	11.7	0.1
Mean number of physician visits in year before enrollment	4.49	4.23	4.80	4.51	4.55	0.2
Mean education (yr)	11.9	12.0	12.0	12.0	11.8	−1.4
Taking enrollment-screening examination (%)	59.1	57.8	58.6	58.5	62.5	1.6
Enrolled for 3 yr (%)	69.8	67.4	71.3	69.5	68.9	−0.3
Physical functioning (mean score, 0–100)[c]						
Enrollees	89.6	88.7	89.1	89.1	88.9	−0.2
Analytic sample	89.6	89.0	89.6	89.4	89.0	−0.5
Role functioning (mean score, %)						
Enrollees	94.8	91.9	91.8	92.8	93.1	0.3
Analytic sample	94.8	92.1	92.5	93.1	93.0	−0.2
Mental health (mean score, 0–100)[c]						
Enrollees	73.8	75.0	73.7	74.2	74.7	0.9
Analytic sample	73.8	75.1	73.9	74.3	74.7	0.8
Social contacts (mean score, 0–100)[c]						
Enrollees	72.8	72.1	72.3	72.4	72.5	0.1
Analytic sample	72.6	72.2	72.0	72.2	72.5	0.3

TABLE 3. Continued.

Variable and Brief Description[b]	Cost-Sharing Plans				Free Plan	t-Test Value[d] for Difference Between Free and Cost-Sharing Plans
	95%	25 and 50%	Individual Deduction	Total Cost-Sharing		
Health perceptions (mean score, 0–100)[e]						
Enrollees	70.5	71.1	69.4	70.4	69.7	−1.2
Analytic sample	70.4	71.2	69.7	70.4	69.8	−1.2
Smoking scale (mean score, 1–2.20)[e]						
Enrollees	1.29	1.30	1.32	1.30	1.29	−0.7
Analytic sample	1.28	1.29	1.30	1.29	1.29	−0.3
Mean standardized weight (kg)						
Enrollees	71.5	71.3	71.0	71.3	71.3	0.0
Analytic sample	71.6	71.3	71.6	71.5	71.6	0.2
Mean cholesterol level (mg/dl)						
Enrollees	207	205	206	206	202	−1.9
Analytic sample	208	205	207	207	204	−1.5
Mean diastolic blood pressure (mmHg)						
Enrollees	75.2	75.3	75.4	75.3	74.6	−1.4
Analytic sample	76.0	75.4	75.7	75.7	74.7	−1.9
Functional far vision (mean number of lines)[e]						
Enrollees	2.28	2.39	2.42	2.37	2.33	−0.9
Analytic sample	2.28	2.37	2.41	2.35	2.32	−0.9
Risk of dying (mean score)[e]						
Enrollees	0.99	1.05	1.12	1.05	1.04	−0.6
Analytic sample	1.00	1.06	1.13	1.06	1.03	−0.8

[a] Values are adjusted for differences according to site.

[b] For demographic data, table entries include everyone with valid enrollment data. For health measures, the mean score for enrollees excludes persons who did not have valid enrollment data owing to the study design (e.g., they were not assigned to an initial screening examination) or to missing data. The mean score for analytic sample excludes persons with no exit data as well as persons who withdrew and for whom we could not recover data on their subsequent health status (i.e., they did not complete a follow-up questionnaire.) Thus, the difference between the two samples shows the effect of attrition.

[c] Values represent equally weighted averages of the three types of cost-sharing plans.

[d] Value shown is for the difference between free and total cost-sharing plans.

[e] See reference 3 for further details of scale.

TABLE 4. Characteristics of Families That Accepted and Refused Enrollment Offer, by Plan

Characteristic	Free Plan		25 and 50%		95%		Individual Deductible		F Value[a]
	Accept	Refuse	Accept	Refuse	Accept	Refuse	Accept	Refuse	
Education (yr)	12.6	12.1	12.6	11.7*	12.8	12.0*	12.7	11.7*	0.2
Physician visits in past year	4.3	4.6	3.6	3.9	3.7	4.5	4.5	3.8	0.9
Income (in thousand of 1973 dollars)	10.2	12.7	10.0	9.0	10.0	11.3	10.3	10.1	2.2
Hospitalized in past year (%)	10.6	9.2	9.2	13.2	10.0	13.2	12.3	13.7	0.5
Male (%)	49.3	42.6	48.8	40.5	46.0	43.0	49.7	44.2	0.2
Age (yr)	29.1	39.8*	28.8	29.9	28.0	34.1*	28.6	30.6	2.4
Black (%)	0.9	0.0	2.1	0.5	2.6	0.4	2.2	0.0	0.1
Health index[b]	9.9	9.8	9.9	9.8	10.0	9.9	9.9	9.6	0.9
Number of families	332	20	234	66	194	65	264	62	
Families (%)	94	6	78	22	75	25	81	19	

NOTE: Sample consists of families in Seattle and Massachusetts. Dayton and South Carolina data for refusals are not available. Variables are averaged within families; averages presented are averages across families.

[a] The F-statistics test the null hypothesis that there are no differences by plan in those who accept and refuse. They have 3 and n degrees of dreedom, where n ranges between 1186 and 1229, because of missing data. The F value for significance at the 5% level is approximately 2.60 and at the 10% level approximately 2.08.

[b] The health index is the sum of responses to one question about the amount of pain (from 1 = a great deal to 4 = none), one question about the amount of worry (from 1 = a great deal to 4 = none), and one self-rated health question (1 = poor, 2 = fair, 3 = good, 4 = excellent).

* $P < 0.05$ for within-plan comparison.

large medical bills were a higher fraction of those withdrawing from the cost-sharing plans, we would overstate the effect of plan on spending. Fortunately, the differences in attrition rate by plan also did not impart any noteworthy bias to our analysis. The most straightforward test of bias we can make is whether those who withdrew spent at a higher rate while on the study than those who completed the experiment. A comparison of the rates of spending of these two groups while they were participating shows no differences that are statistically significant. Specifically, if we test the hypothesis that there is no plan-related bias from sample loss, the relevant statistic for the (all sites) FFS sample is $\chi^2(4) = 2.67$ and for the HMO sample versus the FFS sample (in Seattle) it is $F(5, 3083) = 0.81$; both values are clearly far from significant.

One may ask whether some families might have withdrawn in anticipation of large medical expenditure. As explained in the opening paragraph of this section, they had no financial incentive to do so, and we have no reason to expect that this happened. No data on utilization subsequent to attrition were collected. We did, however, obtain health status data from over three quarters of those who did not complete the experiment; the analysis of plan effects on health status was unaffected by whether these individuals were included or not.[3]

Table 7 shows the completion rates of the various groups in the HMO portion of the experiment. The difference in the proportion completing normally by plan among the three experimental groups is attributable to differences in the period for which persons were enrolled; 75% of the FFS group was enrolled for 3 years and the remainder for 5 years, whereas half of the experimental HMO group was enrolled for 3 years and the remainder for 5 years. Controlling for the length of the enrollment period, we cannot detect any differences in withdrawal rates among the following four experimental groups: experimental HMO, free FFS, indi-

TABLE 5. Mean Values for Demographic and Health Status Measures at Enrollment in Seattle, by Plan* (SE in parentheses)

	Free Fee-for-Service		HMO Experimentals		HMO Controls		Pay Fee-for-Service	
Variable	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Family variables								
AFDC (%)	6.0	(1.9)	6.1	(1.2)	3.6	(1.2)	4.7	(1.3)
Black (%)	1.9	(1.1)	2.3	(0.7)	4.0	(1.1)	3.2	(1.0)
Family size (number)	2.7	(0.13)	2.6	(0.07)	2.4	(0.08)	2.5	(0.08)
Income (1983 $)[a]	25300	(1200)	22400	(600)	27700	(940)	22900	(740)
Person variables								
Age (yr)	25.2	(0.79)	24.6	(0.45)	26.6	(0.60)	26.1	(0.59)
Female (%)	49.7	(2.4)	51.0	(1.5)	52.7	(1.8)	51.0	(1.79)
Number of chronic complaints[b]	6.5	(0.35)	6.9	(0.22)	7.5	(0.26)	7.0	(0.26)
General Health Index[c,d]	72.9	(0.77)	73.5	(0.47)	72.8	(0.57)	74.5	(0.55)
Mental Health Index[c,e]	75.8	(0.68)	75.8	(0.41)	75.2	(0.51)	75.9	(0.52)
Percent with physical or role limitation[f]	16.8	(1.8)	15.9	(1.1)	14.6	(1.3)	14.1	(1.24)
MD visits[g]	3.8	(0.28)	4.0	(0.17)	4.5	(0.21)	3.6	(0.21)
Hospitalized in previous year (%)[g]	10.3	(1.48)	10.5	(0.92)	9.4	(1.09)	10.0	(1.08)
Education (yr)[h]	12.7	(0.12)	12.8	(0.07)	13.8	(0.09)	12.8	(0.08)
Number enrolled	431		1149		733		782	

* Means and SE calculated with the family as unit of observation for family variables and the person as unit of observation for person variables.

[a] The mean of the Free FFS sample is significantly different from the mean of the experimental sample at the 5% level.

[b] Applies to individuals 14+ at enrollment. See reference 14 for details.

[c] A higher value reflects better health.

[d] See reference 3 for details.

[e] Applies to individuals 5+ at enrollment. See reference 3 for details.

[f] Physical limitations are called personal limitations in reference 3.

[g] Year prior to the beginning of the Health Insurance Study. We have no explanation other than chance for the seemingly high GHC Control hospitalization rate (relative to later years) that Welch et al. find anomalous.

[h] Own education if age 18 or older, otherwise education of female (if present, otherwise male) head of household.

vidual deductible FFS, and all other pay FFS ($\chi^2(3) = 0.30$).[17] (The free FFS versus experimental HMO contrast has a t-statistic of 0.50.) The difference shown in Table 7 in attrition rates between the HMO experimental and control groups is largely the effect of a design feature; if a HMO control group member lost eligibility for HMO coverage (e.g., because of a job change), the individual was dropped from the experiment, whereas a HMO experimental member could not lose eligibility in this fashion. Some of the difference also occurs because the entire control group was enrolled for 5 years.

In sum, the most important test of whether attrition affected the utilization results is whether those who withdrew had the same mean rate of spending as those who did not. We cannot detect a difference between these two groups; for that reason and because the percentage withdrawing (for reasons other than moving from the Seattle area) is small

TABLE 6. Average Characteristics at Baseline of Those Who Accepted and Refused Offer to Enroll in Seattle, by Plan*
(SE in parentheses)

	Free Fee-for-Service		Pay Fee-for-Service		HMO Experimentals		HMO Controls		
Variable	Accept	Refuse	Accept	Refuse	Accept	Refuse	Accept	Refuse	*F*-Statistic[a]
Family variables									
Number of families	154	10	311	77	434	111	128	43	
Family size (number)	2.6 (0.1)	2.3 (0.5)	2.4 (0.1)	2.4 (0.2)	2.5 (0.1)	2.8 (0.1)	2.4 (0.1)	2.7 (0.2)	1.0
Income in 1974 ($1000)	10.7 (0.5)	14.6 (2.1)	10.4 (0.3)	10.8 (0.8)	10.3 (0.3)	13.6 (0.6)	12.2 (0.4)	12.2 (1.1)	7.9
Person variables									
Number of persons	431	23	782	191	1149	310	733	145	
Age (yr)	25.2 (1.1)	34.5 (4.7)	25.8 (0.7)	26.8 (1.6)	24.4 (0.6)	28.7 (1.3)	25.6 (0.8)	26.7 (2.0)	5.0
Female (%)	51.1 (2.6)	47.8 (10.6)	50.4 (1.9)	57.8 (3.6)	51.0 (1.6)	50.0 (2.9)	52.8 (2.1)	52.6 (4.7)	0.7
Black (%)	1.7 (1.1)	0.0 (0.0)	2.9 (0.8)	1.7 (1.2)	1.6 (0.7)	3.3 (1.3)	4.0 (1.2)	5.9 (4.8)	1.0
Education of persons 21 years and over (yr)	13.0 (0.2)	13.2 (0.7)	13.1 (0.1)	12.3 (0.3)	13.1 (0.1)	12.8 (0.2)	13.9 (0.2)	14.3 (0.4)	1.8
Self-rated health status[b]	1.5 (0.0)	1.7 (0.2)	1.5 (0.0)	1.5 (0.1)	1.5 (0.0)	1.5 (0.1)	1.4 (0.0)	1.5 (0.1)	0.8
Amount of pain due to health[c]	3.3 (0.0)	3.1 (0.3)	3.3 (0.0)	3.4 (0.1)	3.3 (0.0)	3.3 (0.1)	3.2 (0.0)	3.2 (0.1)	1.5
Amount of worry due to health[c]	3.2 (0.1)	3.5 (0.2)	3.3 (0.0)	3.2 (0.1)	3.2 (0.0)	3.4 (0.1)	3.2 (0.0)	3.1 (0.1)	1.7
Physician visits in previous year	3.9 (0.4)	2.5 (0.4)	3.5 (0.2)	4.1 (0.5)	4.0 (0.2)	3.4 (0.3)	4.5 (0.3)	5.4 (0.9)	1.7
Percentage hospitalized in previous year	10.6 (1.6)	4.3 (10.0)	9.8 (1.4)	14.3 (3.1)	10.4 (1.1)	8.4 (1.7)	9.3 (1.3)	11.4 (3.7)	1.1
Dental visits in previous year	2.1 (0.2)	2.4 (0.4)	2.0 (0.2)	1.7 (0.2)	1.8 (0.1)	2.0 (0.2)	2.2 (0.2)	1.9 (0.3)	0.9
Percent on AFDC program	12.0 (2.0)	0.0 (0.0)	6.1 (1.2)	11.0 (3.1)	5.2 (1.0)	3.9 (1.5)	2.2 (0.8)	1.8 (2.4)	0.7
Percent with group insurance	77.6 (3.4)	78.3 (13.3)	77.8 (2.4)	80.8 (4.5)	78.7 (2.1)	88.0 (2.8)	73.4 (2.9)	60.5 (7.1)	3.8
Percent with nongroup insurance	10.4 (2.7)	17.4 (10.0)	14.0 (2.0)	9.3 (4.0)	12.0 (1.6)	10.3 (2.7)	37.7 (3.1)	51.8 (7.5)	0.4
Percent with public insurance	9.6 (1.8)	0.0 (0.0)	5.1 (1.1)	15.5 (3.7)	4.7 (1.0)	3.9 (1.5)	2.4 (0.8)	0.9 (0.4)	3.0

* The values in this table are based on baseline variables and baseline family compositions for those who accepted (enrolled) and refused. For all but the Control group, missing values comprised about 5% of the data; the figure is about 20% for the Control group.

[a] F-statistics test the null hypothesis that there are no differences by plan between those who accept and refuse (4 degrees of freedom in the numerator). An F of 2.9 is significant at the 5% level.

[b] Response to: How would you rate your health? 1-Excellent; 2-Good; 3-Fair; 4-Poor.

[c] Responses to: How much pain (worry) does your health cause you? 1-Lots; 2-Some; 3-A little; 4-None.

relative to the percentage not withdrawing, especially for the HMO Experimental-Free FFS comparison, we conclude that there is no appreciable bias from attrition.

Findings on Biased Reporting and Methods Effects

Although Welch et al. do not raise the issue of biased reporting, we have shown elsewhere that this bias is also small.[18] Specifically, outpatient physician expenditure on the 95% plan is understated by about 5–10% relative to the free plan because more visits were made on that plan for which claims were not filed. This means that the true reduction in outpatient expenditure caused by the 95% plan is closer to 35% than the 40% shown in Table 1. Because HMO data were abstracted from the medical record, there is no issue of underfiling at the HMO; hence, the true HMO-FFS difference may be larger than we describe. There could in principle be underfiling of out-of-plan use by the HMO sample; however, because our out-of-plan use rates are similar to those in other studies, there is no reason to suspect underreporting. Even if out-of-plan use were underreported, it is such a small amount as to be immaterial to our overall conclusions.[4,7]

The HIS design also included a number of subexperiments to test for methods effects.[19] These subexperiments showed methods effects had negligible effect on levels of use and plan comparisons.[6,12]

Summary of Discussion on Bias

The data in Tables 1–7 establish the following facts: (1) The various groups studied in the experiment were assigned by unbiased methods and were approximately comparable at enrollment in terms of prior use, health status, and sociodemographic characteristics; (2) attrition during the experiment did not importantly alter that comparability; and (3) the groups had quite different rates of use of medical care services during the experiment. Additionally, reporting and methods effects are small relative to the size of effects found. If these facts are accepted, Welch et al. have no basis for questioning that both cost sharing and the Health Maintenance Organization reduced the use of services relative to free FFS care and that the magnitude of the reduction was as our previous publications have described.[4–10]

TABLE 7. Reasons for Noncompletion of Study in the Seattle Area, by Plan (%)

Reason	Free Fee-for-Service	Pay Fee-for-Service	HMO Experimental	HMO Control
Voluntarily withdrew	0.2	9.6	4.4	11.9
Terminated because of failure to meet study obligations	3.0	3.2	2.5	20.6[a]
Died	0.2	0.8	0.3	0.3
Moved from Seattle area	20.7[b]	12.9[b]	22.3[b]	10.5
Other	0.2	0	0.5	0.8
Completing normally	75.6	73.5	70.1	55.9

[a] Loss of eligibility for the HMO (for example, because of employment change) is included in this value.
[b] These individuals were kept in the experiment; the HMO Experimentals, however, were switched to the Free Fee-for-Service plan once they moved from the Seattle area. Once HMO Control families moved away from the Seattle area, they were dropped.

Some Canards

Data Inaccessibility

Welch et al. complain of information inadequacy and inaccessibility to data. Although we think they had no basis for complaint in the first place, public use files have now been released with accompanying documentation,[20–33] so this issue is settled. (The documents referenced are those released as of the date of publication. More are in preparation.) We caution, however, that users of these files will not be able to replicate exactly our published results because, in preparing public use files, we have corrected errors that

came to light after we published those findings. The effect of these corrections should be quite minor; however, for reasons of economy, we have not generally repeated our analyses with the corrected data.

Bias

We are accused of bias by Welch et al. because our focus was a comparison of the HMO Experimentals with the Free FFS plan. They argue that cost sharing in FFS can reduce utilization to the levels observed in HMO. Although they concede that we pointed this out, it is their contention that we did not give it enough emphasis.

Their accusation of bias has no substance. Two points are relevant:

(1) From a scientific point of view, we wished to estimate the effect of the HMO on utilization holding all else constant. In particular, we wanted to keep the scope of benefits and amount of cost sharing the same between the HMO and the FFS plan we used for comparison. This dictated use of the free-care plan as the comparison group. In short, the comparison we made measured "the HMO effect" on use. Welch et al. want to compare the HMO without cost sharing to FFS with cost sharing, a comparison anyone can make from the data we present,[4,7] but nonetheless a hybrid comparison. In particular, it is likely that HMO utilization would have been still lower had the HMO elected to use cost sharing at the time of use, as some HMOs do. For example, the Kaiser-Permanente Health Plan uses copayment ranging as high as $10 per visit.

(2) In the paper about which Welch et al. seem to complain most,[7] we sought to demonstrate that with comparable cost sharing (i.e., none), the HMO had less use. We did not try to show that on average HMO care was better or preferred to FFS care, which Welch et al. seem to assume was our intent. As we said in this paper, any overall judgment should at least consider health status and satisfaction effects; in fact, these present a somewhat mixed picture.[34,35] In addition, of course, generalizing to the universe of HMOs based on experience at one HMO is decidedly hazardous, as we have pointed out in many places.

Cost Versus Price

Welch et al. argue that not only we, but also Enthoven,[36] and the press have ignored the distinction between the cost of services to the HMO and the premium the HMO charges. They imply that the reductions in use at the HMO are not translated into savings for the consumer. We make four points in response:

(1) As Welch et al. concede, capital, organizational, and marketing costs of HMOs have not been quantified. Hence, their argument is conjectural. Nonetheless, suppose that in the absence of data one were forced to choose whether FFS or HMOs had higher administrative costs. We think most readers would guess that, if anything, the combination of private insurers and private hospitals and physicians, not HMOs, had the higher administrative and marketing expenses. One basis for such a guess is that there may well be some economies (of scope) to combining the insurer and provider roles (e.g., claim forms do not have to be prepared, processed, or paid). In any event, it seems to us implausible that any difference in marketing or administrative expenses at the HMO could compensate for the estimated 28% difference in medical care expenditures that we found.

(2) Welch et al. present no valid evidence for their claim. Their data, a comparison of HMO and Blue Cross premiums, are irrelevant because of possibly different risk mixes in the two groups (i.e., selection effects). For example, those who purchase a Blue Cross policy with a $1,000 deductible (and scope of benefit differences from the HMO) may well be healthier than those who enroll at the HMO. Our finding that the HMO population is similar to the FFS population[7] does not imply it is similar to a subset of FFS users on a particular insurance plan (i.e., on the Blue Cross $1,000 deductible plan).

(3) In a competitive market price will approximate cost; thus, even if Welch et al. were correct about the past, one would expect their assertion not to hold if the medical services market becomes more competitive.

(4) We compare the value of services rendered, using the same price per unit for FFS and the HMO (e.g., the same price of a brief visit in both systems). To the degree that the HMO used a less costly mix of inputs (e.g., physicians assistants and nurse practitioners

instead of physicians), our estimated cost difference is understated.

Review

The Department of Health and Human Services is chided by Welch et al. for inadequate review procedures. This criticism is misguided, and in part it is factually wrong. At the request of the federal government one outside group and a second partially outside group reviewed the design at the beginning of the project. Furthermore, the design was published with a critique and response.[37–39] We believe the design of the project was exposed to much greater prior review than is customary in health services research, although we believe that the scale of the undertaking warranted such attention.

Welch et al., however, seem less concerned with the design than with the analysis, which has been published in many articles in distinguished journals and Rand publications.[1–19,34–35,37,40–65] Taking these papers as a group, approximately 100 referees from diverse disciplines have been asked to review our methods, results, and conclusions. The results have also been presented at many conferences and other forums with the usual complement of discussants.

We believe this process has functioned well, and we find simply not credible Welch et al.'s belief that a DHHS-commissioned review group would find something important that this large group of referees did not find. Welch et al. apparently believe that one must read all or nearly all of the publications in order to appraise any individual one, but this is clearly not the case. Although space limitations on journal articles have precluded publishing some methodologic details in those articles, consider that (1) prior to accepting the manuscripts for publication, journal referees and editors have frequently asked for details that were not contained in the manuscripts; and (2) more detail has generally been available in the Rand monographs.[1,3,4,6,8,13,40,42,49,65] We think the reader can judge the likelihood that an additional DHHS review would have been worth the time and money. In any event, Welch et al.'s concluding charge that the allegedly inadequate review process led to "immense" costs in "human life and suffering" is inconsistent with their opening claim that differences among treatments have not been demonstrated. How can choice among supposedly similar treatments matter?

Statistical Methods

They criticize our statistical methods. Although most of our response to these arguments is relegated to the appendix, it is germane to point out here that a coauthor of the Welch et al. critique refutes Welch et al.'s conclusions in another paper! We quote from the abstract of that paper: "The results suggest that despite theoretical concerns with the Rand Two Part Model, it actually outperforms the two comparison Sample Selection Model variants in terms of Mean Square Error of parameter estimate (sic)."[66]

Welch et al. raise a host of minor points; those we regard as more technical or more substantial, we respond to in the appendix.

References

1. Duan N, Manning WG, Morris CN, et al. A comparison of alternative models of the demand for medical care. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2754-HHS).

2. Manning WG, Newhouse JP, Duan N, et al. Health insurance and the demand for medical care: evidence from a randomized experiment. American Economic Review, forthcoming.

3. Brook RH, Ware JE Jr, Rogers WH, et al. The effect of coinsurance on the health of adults. Santa Monica, CA: The Rand Corporation, 1984 (Publ. no. R-3055-HHS).

4. Manning WG, Leibowitz A, Goldberg G, et al. A controlled trial of the effect of a prepaid group practice on use of services. Santa Monica, CA: The Rand Corporation, 1984 (Publ. no. R-3029-HHS).

5. Newhouse JP, Manning WG, Morris, CN, et al. Some interim results from a controlled trial of cost sharing in health insurance. N Engl J Med 1981;305:1501.

6. Newhouse JP, Manning WG, Morris CN, et al. Some interim results from a controlled trial of cost sharing in health insurance. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2847-HHS).

7. Manning WG, Leibowitz A, Goldberg G, et al. A

controlled trial of the effect of a prepaid group practice on use of services. N Engl J Med 1984;310:1505.

8. Keeler, EB, Rolph JE, Duan N, et al. The demand for episodes of medical services: interim results from the health insurance experiment. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2829-HHS).

9. Keeler, EB, Rolph JE. How cost sharing reduced medical spending of participants in the health insurance experiment. JAMA 1983;249:2220.

10. Lohr KN, Brook RH, Kamberg CJ, et al. Use of medical care in the Rand health insurance experiment: diagnosis- and service-specific analyses in a randomized controlled trial. Med Care 1986;24(Suppl):1.

11. Duan N, Manning WG, Morris CN, et al. A comparison of alternative models of the demand for medical care. Journal of Business and Economic Statistics 1983;1: 115.

12. Manning WG, Newhouse JP, Duan N, et al. Health insurance and the demand for medical care: evidence from a randomized experiment. Santa Monica, CA: The Rand Corporation, forthcoming (Publ. no. R-3476-HHS).

13. Valdez RB. The consequences of cost sharing for children's health. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. R-3270-HHS).

14. Manning WG, Newhouse JP, Ware JE Jr. The status of health in demand estimation: beyond excellent, good, fair, and poor. In: VR Fuchs, ed. Economic Aspects of Health. Chicago: University of Chicago Press, 1982.

15. Morris CN. Sample selection in the health insurance experiment: comparing the enrolled and nonenrolled populations. Santa Monica, CA: The Rand Corporation, 1985 (Publ. no. N-2354-HHS).

16. Morris CN. A finite selection model for experimental design of the Health Insurance Study. Journal of Econometrics 1979;11:43.

17. Manning WG, Wells KB. Preliminary results of a controlled trial of the effect of a prepaid group practice on the outpatient use of mental health services. Journal of Human Resources 1986;21:293.

18. Rogers WH, Newhouse JP. Measuring unfiled claims in the health insurance experiment. In: Burstein L, Freeman HE, Rossi PH, eds. Collecting Evaluation Data: Problems and Solutions. Beverly Hills: Sage, 1985.

19. Newhouse JP, Marquis KH, Morris CN, et al. Design improvements in the second generation of social experiments: the health insurance study. Journal of Econometrics 1979;11:117.

20. Polich SM, Taylor CD. Master sample series: codebook for eligibility-family changes file. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2264/ 1-HHS).

21. Polich SM, Campbell NF, Taylor CD, et al. Master sample series: codebook for full sample demographic file. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2264/2-HHS).

22. Peterson CE, Nelsen M, Bloomfield ES. Aggregated claims series: codebook for FFS annual expenditures and visit counts. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2360/1-HHS).

23. Peterson CE, Nelsen M, Wesley DL, et al. Aggregated claims series: Codebooks for fee-for-service visits—outpatient, inpatient, and dental. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2360/ 2-HHS).

24. Peterson CE, Taylor CD, Bloomfield, ES. Aggregated claims series: codebooks for fee-for-service treatment episodes and annual episode counts. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2360/ 3-HHS).

25. Peterson CE, Nelsen M, Wesley DL. Aggregated claims series: codebooks for HMO and Seattle FFS visits—outpatient and inpatient. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2360/4-HHS).

26. Peterson CE, Nelsen M, Wesley DL. Aggregated claims series: codebook for HMO and Seattle annual expenditures. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2360/5-HHS).

27. Peterson CE, Nelsen M, Welsey DL. Claims line-item series: codebook for FFS claims. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2347/1-HHS).

28. Peterson CE, Nelsen M, Wesley DL. Claims line-item series: codebook for FFS claims. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2347/2-HHS).

29. Peterson CE, Nelsen M, Wesley DL. Claims line-item series: codebook for FFS claims for comparison with HMO claims. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2347/3-HHS).

30. Nelsen M, Edwards CA. HIE references: codes used in HIE claims. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2349/1-HHS).

31. Edwards CA, Holland AB, Weissler LY. Medical history questionnaire series: codebooks for adults at enrollment and exit, form A. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2485/1-HHS).

32. Edwards CA, Holland AB, Weissler LY. Medical history questionnaire series: codebooks for adults at enrollment and exit, form B. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2485/2-HHS).

33. Edwards CA, Bell AM, Wesley DL. Medical history questionnaire series: codebooks for children at enrollment and exit. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. N-2485/3-HHS).

34. Ware JE Jr, Brook RH, Rogers WH, et al. Comparison of health outcomes at a health maintenance organization with those of FFS care. Lancet 1986;1(848): 1017.

35. Davies AR, Ware JE, Brook RH, et al. Consumers' attitudes toward prepaid and FFS medical care: results from a randomized trial. Health Services Research 1986;21:429.

36. Enthoven A. The Rand experiment and economical health care. N Engl J Med 1984;310:1528.

37. Newhouse JP. A design for a health insurance experiment. Inquiry 1974;11:5.

38. Hester J, Leveson I. The health insurance study: a critical appraisal. Inquiry 1974;11:53.

39. Newhouse JP. Reply to Hester and Leveson. Inquiry 1974;11:236.

40. Wells KB, Manning WG, Duan N, et al. Cost

sharing and the demand for ambulatory mental health services. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2960-HHS).

41. Brook RH, Ware JE Jr, Rogers WH, et al. Does free care improve adults' health? Results from a randomized controlled trial. N Engl J Med 1983;309:1426.

42. Manning WG, Morris CN, Newhouse JP, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. eds. Perlman M, van der Gaag J, Health Economics, and Health Economics. Amsterdam: North Holland, 1981.

43. Manning WG, Wells KB, Duan N, et al. Cost sharing and the use of ambulatory mental health services. Am Psychol 1984;39:1077.

44. Ware JE Jr, Manning WG, Duan N. Health status and the use of outpatient mental health services. Am Psychol 1984;39:1090.

45. Bailit HL, Brook RH, Kamberg C, et al. The effect of cost sharing on the quality of dental care. J Dent Educ 1984;48:597.

46. Wells KB, Manning W, Duan N, et al. The sensitivity of mental health care use and cost estimates to methods effects. Med Care 1984;22:783.

47. Valdez RB, Brook RH, Rogers WH, et al. The consequences of cost sharing for children's health. Ped 1985;75:952.

48. Leibowitz A, Manning WG, Keeler EB, et al. The effect of cost sharing on the use of medical services by children: interim results from a randomized controlled trial. Ped 1985;75:942.

49. Leibowitz A, Manning WG, Keeler EB, et al. The effect of cost sharing on the use of medical services by children: interim results from a randomized controlled trial. Santa Monica, CA: The Rand Corporation, 1985 (Publ. no. R-3287-HHS).

50. Bailit H, Newhouse J, Brook R, et al. Does more generous dental insurance coverage improve oral health? J Am Dent Assoc 1985;110:701.

51. Manning WG, Bailit HL, Benjamin B, et al. The demand for dental care: evidence from a randomized trial in health insurance. J Am Dent Assoc 1985;110: 895.

52. Manning WG, Bailit HL, Benjamin B, et al. The demand for dental care: evidence from a randomized trial in health insurance. Santa Monica, CA: The Rand Corporation, (Publ. no. R-3225-HHS), 1986.

53. O'Grady KF, Manning WG, Newhouse JP, et al. The impact of cost sharing on emergency department use. N Engl J Med 1985;313:484.

54. O'Grady KF, Manning WG, Newhouse JP, et al. The impact of cost sharing on emergency department use. Santa Monica, CA: The Rand Corporation, 1985 (Publ. no. N-2376-HHS).

55. Keeler EB, Brook RH, Goldberg GA, et al. How free care reduced hypertension in the Health Insurance Experiment. JAMA 1985;254:1926.

56. Keeler EB, Brook RH, Goldberg GA, et al. How free care reduced hypertension in the Health Insurance Experiment. Santa Monica, CA: The Rand Corporation, 1985 (Publ. no. R-3326-HHS).

57. Leibowitz A, Manning WG, Newhouse JP. The demand for prescription drugs as a function of cost sharing. Soc Sci Med 1985;21:1063.

58. Leibowitz A, Manning WG, Newhouse JP. The demand for prescription drugs as a function of cost sharing. Santa Monica, CA: The Rand Corporation, 1985 (Publ. no. N-2278-HHS).

59. Shapiro M, Ware JE Jr, Sherbourne CD. Effects of cost sharing on seeking care for serious and minor symptoms: results of a randomized controlled trial. Ann Int Med 1986;104:246.

60. Bailit HL, Newhouse J, Brook R, et al. Dental insurance and the oral health of preschool children. J Am Dent Assoc 1986;256:1930.

61. Ware JE Jr, Brook RH, Rogers WH, et al. How do health outcomes at an HMO compare with those of FFS care? An experimental comparison. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. R-3153-HHS).

62. Manning WG, Wells KB, Ware JE Jr, et al. How cost sharing affects the use of ambulatory mental health services. JAMA 1986;256:1930.

63. Foxman B, Valdez RB, Lohr KN, et al. The effect of free care on the use of antibiotics: results from a population-based controlled trial. J Chronic Dis, in press.

64. Wells KB, Manning WG, Duan N, et al. Sociodemographic factors and the use of outpatient mental health services. Med Care 1986;24:75.

65. Lohr KN, Brook RH, Kamberg CJ, et al. Use of medical care in the Rand health insurance experiment: diagnosis- and service-specific analyses in a randomized controlled trial. Santa Monica, CA: The Rand Corporation (Publ. no. R-2369-HHS), 1986.

66. Hay JW, Leu R, Rohrer P. Ordinary least squares and sample selection models of health care demand: a monte carlo comparison. Stanford: Hoover Institution, 1985 (Working Papers in Economics E-85-29).

67. Hay JW, Olsen RJ. Let them eat cake: a note on comparing alternative models of the demand for medical care. Journal of Business and Economic Statistics 1984;2: 279.

68. Maddala GS. A survey of the literature on selectivity bias as it pertains to health care markets. Advances in Health Economics and Health Services Research 1985;6:3.

69. Duan N, Manning WG, Morris CN, et al. Choosing between the sample-selection model and the multipart model. Journal of Business and Economic Statistics 1984;2:283.

70. Duan N, Manning WG, Morris CN, et al. Comments on selectivity bias. Advances in Health Economics and Health Services Research 1985;6:19.

71. Duan N, Manning WG, Morris CN, et al. A comparison of alternative models of the demand for medical care. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2754-HHS).

72. Duan N, Manning WG, Morris CN, et al. A comparison of alternative models of the demand for medical care. Journal of Business and Economic Statistics 1983;1: 115.

73. Cramer H. Mathematical methods of statistics. Princeton, NJ: Princeton University Press, 1946.

74. Stein C. Efficient nonparametric testing and es-

timation, proc. third Berkeley Symp. Math. Stat. Prob. University of California Press 1956;1:187.

75. Bickel PJ. On adaptive estimation. Annals of Statistics 1980;10:647.

76. Maddala GS. Further comments on selectivity bias. Advances in Health Economics and Health Services Research 1985;6:25.

77. Lee LF. Tests for the bivariate normal distribution in econometric models with selectivity. Econometrica 1984;52:843.

78. Cosslett SR. Distribution-free estimator of a regression model with sample selectivity. Discussion Paper, Center for Econometrics and Decision Sciences, University of Florida, 1984.

79. Duan N, Li KC. Regression analysis with nonrandomly missing outcome data: distribution-free and link-free estimation of the sample selection model. Santa Monica, CA: The Rand Corporation, 1986 (unpublished manuscript).

80. Hay JW, Leu R, Rohrer P. Ordinary least squares and sample selection models of health care demand: a monte carlo comparison. Stanford: Hoover Institution, 1985 (Working Papers in Economics E-85-29).

81. Manning W, Duan N, Rogers W. Monte Carlo evidence on the choice between sample selection and two-part models. Santa Monica, CA: The Rand Corporation, 1986 (Publ. no. WD-3028-HHS).

82. Manning WG, Newhouse JP, Duan N, et al. Health insurance and the demand for medical care: evidence from a randomized experiment. Santa Monica, CA: The Rand Corporation, forthcoming (Publ. no. R-3476-HHS).

83. Newhouse JP, Manning WG, Morris CN, et al. Some interim results from a controlled trial of cost sharing in health insurance. N Engl J Med 1981;305:1501.

84. Newhouse JP, Manning WG, Morris CN, et al. Some interim results from a controlled trial of cost sharing in health insurance. Santa Monica, CA: The Rand Corporation, 1982 (Publ. no. R-2847-HHS).

85. Manning WG, Leibowitz A, Goldberg G, et al. A controlled trial of the effect of a prepaid group practice on use of services. Santa Monica, CA: The Rand Corporation, 1984 (Publ. no. R-3029-HHS).

86. Manning WG, Leibowitz A, Goldberg G, et al. A controlled trial of the effect of a prepaid group practice on use of services. N Engl J Med 1984;310:1505.

87. Manning WG, Morris CN, Newhouse JP, et al. A two-part model of the demand for medical care: preliminary results from the health insurance study. New York: North-Holland, 1981.

Appendix

Choosing Between the Sample Selection and the Multipart Models

Welch et al. repeat some earlier criticisms by Hay and Olsen[67] and Maddala[68] on our choice of the multipart model instead of the sample selection model. They do not address our responses to those criticisms. We therefore repeat part of our response.[69,70]

Most of our analyses were based on another version of the multipart model, the four-part model. However, the two-part model has been the focus of the earlier exchange because it is analogous to the sample selection model. A version of the multipart model, the two-part model, is as follows:[69–72]

$$I_i = x_i\delta_1 + \eta_{1i}, \qquad \eta_{1i} \sim N(0, 1) \tag{1.1}$$

$$\log(MED_i) = x_i\delta_2 + \eta_{2i}, \tag{1.2}$$

$$(\eta_{2i} | I_i > 0) \sim F(0, \sigma^2), \tag{1.3}$$

$$(\eta_{2i} | I_i \leq 0) \equiv -\infty \; (MED_i = 0), \tag{1.4}$$

where

$$MED_i > 0 \quad \text{if} \quad I_i >, 0 \;\; 0 \text{ otherwise.}$$

We need not impose normality in *1.3.*

The appropriate version of the sample selection model, the adjusted Tobit model, is

$$I_i = x_i\delta_1 + \eta_{1i}, \tag{1.5}$$

$$w_i = x_i\beta_2 + \epsilon_{2i}, \tag{1.6}$$

$$\log(MED_i) = w_i \quad \text{if} \quad I_i > 0,$$

$$= -\infty \quad \text{if} \quad I_i \leq 0, \tag{1.7}$$

$$(\eta_{1i}, \epsilon_{2i})' \sim N(0, \Sigma), \tag{1.8}$$

where

$$\Sigma = \begin{pmatrix} 1 & \rho\tau \\ \rho\tau & \tau^2 \end{pmatrix}.$$

We noted:[4]

> ". . . all parties to this debate accept the validity of the sample mean as an estimate of the population mean for the experimental data. Our procedure simply decomposes the sample mean by using the following identity:
>
> $$E(X) = P(X > 0)*E(X|X > 0). \tag{1}$$
>
> "The issue, then, is whether any problem arises from estimating each part of the right-hand side of this identity separately and then using those two parts to arrive at an estimate of E(X). Is the problem bias or inconsistency? Is the problem efficiency? Maddala does not tell us. The problem that he and H-O (Hay-Olsen) stumble over is possible correlation between the error terms of the two equations used to estimate the right-hand side.
>
> "H-O argue that because of this possible correlation our method is nested within the sample selection model; we have showed—and Maddala agrees—that this is not the case (Duan et al., 1984). Maddala and we also agree that we have formulated this model in a way that the correlation, if any, cannot be estimated. Maddala believes that this is a mistake: It is advisable not to formulate the model in a way that the correlation cannot ever be estimated.
>
> "We do not agree with this statement in either the general case or the case at hand. At a general level, there is ample precedent for formulating models in a way that some parameters are not estimated. . . .
>
> "In the specific case of the health insurance data one does not need an estimate of ρ to estimate mean expenditure. For example, the two-part (or multipart) model with only the intercept on the right-hand side and with the smearing retransformation (Duan, 1983) identically reproduces the sample mean irrespective of the value of ρ and is consistent as long as the sample mean is consistent (which is by design in this case). On the other hand, the estimates of the adjusted Tobit model are consistent estimates of the mean only if the distributional assumption is correct."

In the appendix, Welch et al. commented on the joint error distribution given in Duan et al.,[69] and alleged that this joint error distribution violates a regularity condition for maximum likelihood estimation and might not satisfy mathematical axioms related to the measurement of probabilities. The apparent reason for both alleged problems is that this joint error distribution places a nonzero probability mass at minus infinity.

We do not know of any mathematical axiom that prohibits placing a nonzero probability mass at minus infinity. Consider, for example, a random variable Y that can assume two values, zero and minus infinity, with probability P for the value zero, and probability $1 - P$ for the value minus

infinity. Given a sample, one can obviously estimate P, e.g., using the maximum likelihood estimate, which estimates P by the proportion of zeroes. (This example is in essence the probit equation in the two-part model, as well as the first equation of the sample selection model.)

The alleged violation of a regularity condition appears to be a misconception by Welch et al. The asymptotic properties of the maximum likelihood estimate usually require certain moment conditions for the score function and its derivatives.[73] These conditions are obviously satisfied for the example above. It is true, though, the the random variable Y does not have a moment. However, the moments of Y are not required for the asymptotic theory of the MLE for P.

The violation of a regularity condition alleged by Welch et al. is equivalent to the following question: Does the standard asymptotic theory for the MLE apply to the two-part model under the joint error distribution specified in reference 3? Strictly speaking, the question is ill-posed because the two-part model need not assume a parametric form for the conditional error distribution *1.3*. In order to make the question meaningful, we must impose a parametric assumption on the conditional error distribution *1.3*, (e.g., assuming F is normal)

$$(1.3') \qquad (\eta_{2i}|I_i > 0) \sim N(0, \sigma^2),$$

so that the two-part model estimate defined above is indeed the MLE.

Under the normality assumption *1.3'*, the answer to the question posed above is yes, under the usual conditions on the design matrix for the asymptotic theory for the probit and the least squares estimates to apply. Note that for the two-part model with this joint error distribution (or any joint error distribution that satisfies *1.1*, *1.3'*, and *1.4*), the conditional distribution of the outcome, MED, given the explanatory variables, x, is identical to the conditional distribution of MED given x under the sample selection model with zero correlation. Therefore, the sampling distribution of the maximum likelihood estimate for the two-part model under the normality assumption *1.3'* is identical to that obtained from the sample selection model with the correlation known to be zero. This follows from the following observation.

Observation

Consider a semiparametric model:

$$MED|x \sim f(MED; x, \theta, G),$$

where θ is the vector of parameters of interest, G is a vector (possibly infinite-dimensional) of nuisance parameters. If the conditional distribution of MED|x does not depend on the nuisance parameter G, i.e.,

$$(A.1) \qquad f(MED; x, \theta, G_1) = f(MED; x, \theta, G_2)$$

for all admissible values G_1 and G_2 of the nuisance parameter, then the maximum likelihood estimate $\hat{\theta}$ does not depend on the nuisance parameter G; furthermore, the sampling distribution of the MLE $\hat{\theta}$ does not depend on G. In other words, $\hat{\theta}$ has the same distribution under the models f(MED;

x, θ, G_1) and f(MED; x, θ, G_2) for all admissible values G_1 and G_2 of the nuisance parameter.

To see how the observation applies to the two-part model, we take

$$\theta = (\delta_1', \delta_2', \sigma^2)',$$

for any joint distribution G of (η_1, η_2). Condition (A.1) is satisfied for any G_1 and G_2 that satisfy *1.1–1.4.* It follows that $\hat{\theta}$ has the same distribution under the following specific joint error distributions:

G_1 = the joint error distribution in Duan et al.,[69]

G_2 = the adjusted Tobit model with the correlation known to be zero.

It follows that the asymptotic theory holds under the normality assumption *1.3'* for the two-part model estimate, as long as the asymptotic theory holds for the MLE for the sample selection model with the correlation known to be zero.

Note that the observation indicates that the MLE $\hat{\theta}$ for the semiparametric model f(MED; x, θ, G) is adaptive in G, i.e., we can estimate θ "as well asymptotically not knowing G as knowing G."[74,75] (Condition (A.1) actually implies Stein's condition for adaptive estimation.) In other words, we do not need to know (or estimate) the joint error distribution in order to estimate θ efficiently. In particular, we do not need to know (or estimate) the correlation between η_1 and η_2 in order to estimate θ efficiently. This is the basis for our earlier assertion that we do not agree with Maddala's belief that "it is advisable not to formulate the model in such a way that the correlation cannot ever be estimated".[68]

Welch et al. repeat Maddala's[68] distinction between a concomitant decision process and a sequential decision process. As we noted,[70] the two-part model does not require a concomitant decision process:

> "Even if one's data did arise from a bivariate distribution, mathematically Maddala's distinction is unnecessary because a bivariate distribution can always be specified either jointly or as a marginal and a conditional distribution, although one method might be more convenient than the other in any given application."

Maddala, in further comments, agrees that our methods, i.e., multipart models, are "applicable to experimental data," though they may not be applicable to non-experimental data.[70]

Welch et al. commented on our statement[72] that the sample selection model relies on "untestable assumptions" and repeated Maddala's comment that more recent work[77,78] has gone beyond those assumptions.[79] In the light of those new developments, our statement[72] should be interpreted in its historical context: The tools for going beyond the parametric joint error distributions were not available then. We subsequently noted:[70]

> "We are glad to see the work of Lee (1984) and hear of the work of S. R. Cosslett (cited in Maddala's paper) which test and relax the distributional assumptions in the selectivity models. We would like to point out, though, that the parametric form of the error distribution is not the only key distributional assumption made in selection models; exchangeability is also critical."

Duan and Li discuss further the exchangeability condition.[79]

So far, this discussion has been theoretical. Monte Carlo studies by Hay et al.[80] and Manning et al.[81] indicate that the alleged problems with two-part model raised by Hay and Olsen,[67] Maddala,[68,76] and Welch et al. are of negligible magnitude even *if* the sample selection model is the true state of the world. Hay et al.[80] suggest that the inconsistency in the two-part model is small relative to the sample variance of the selection models.

Manning et al.[81] found that the overall forecast bias from the two-part model (for data generated by a normal theory selection model) is negligible for a wide range of correlations. In fact, the two-part model often outperforms the sample selection model with smaller mean squared forecast error and bias than that obtained from limited information or full information maximum likelihood (LIML and FIML) estimators for the sample selection model, even when the data are generated by a sample selection model. Mean squared error was smaller for two-part than LIML selection models, and on the same (or smaller) order for two-part than FIML estimates. Manning et al.[81] also show that if the two equations share the same set of covariates, as the two-part models typically do, one cannot distinguish sample selection effects from simple nonlinearity (e.g., a quadratic or logarithmic term). Hence, when the sample selection model is true, it is empirically difficult to distinguish the sample selection model from the two-part model.

Given the poor numerical performance of sample selection models in Monte Carlo studies, and their poor fit to the HIE data[69,82] we believe the case for using the multi-part models is very strong.

Pre- versus Per-experimental Results

Welch et al. suggest that we should have estimated the difference between use rates prior to and during the experiment ("pre- and per-experimental" to use their term). This misstates the issue. The issue is not whether people changed their use or not, but whether they would behave differently on different coinsurance or HMO plans.

The simple ANOVA comparisons do not, of course, take account of pre-experimental utilization. One could compare the difference in experimental and pre-experimental use across plans. However, the randomization of insurance coverage means that cross-sectional ("per") comparisons give unbiased estimates of differences in utilization across insurance plans, although we do lose some precision by not accounting for pre-experimental utilization. However, we did control for pre-experimental utilization in the fee-for-service estimates[83,84] and provided the Seattle data on pre-experimental use[85] that Welch et al. cite.

Welch et al. also err in their claim that the differences between pre- and per-experimental rates are not significant, because they use inflated standard errors to reach this conclusion. Specifically, they use the published standard error of the mean level to calculate the standard error of the difference, but these values are not sufficient to do so. From the data provided, they can calculate the standard error for the difference if and only if the pre-

and per-observations are uncorrelated. This condition does not hold, because of the large positive intertemporal correlation (those with high use in the pre-period tend to have high use in the experimental period). Calculating the standard error for the difference as if the pre- and per-observations are independent overstates the true standard error of the difference. Hence, the pre- versus per-experimental results, if these were of interest, would be more significant than they assert.

Specification Searches

Welch et al. suggest that our results are too significant because we engaged in specification searches. This argument is inappropriate for the ANOVA results presented in both Newhouse et al.[83,84] and Manning et al.,[85,86] as well as in the main body of this response. By its very nature, ANOVA does not involve specification searches. The ANOVA results are statistically significant at conventional levels. Because ANOVA is inherently less precise than appropriately estimated multiple regression models, the true precision of the plan response on the Health Insurance Experiment lies somewhere between the ANOVA and regression results reported in Newhouse et al.[83,84] and Manning et al.[85,86] Because the ANOVA and multiple regression estimates are quantitatively similar, our conclusions are not affected by specification searches.

(In the case of the GHC versus fee-for-service analysis[85,86] [Table 2 above], all of the analyses, other than expenditures, are by analysis of variance. In the case of expenditures, we used a four-part model with an age-sex specification developed on other HIE data. Thus, there is no question of reported inference statistics overstating our precision.[85,86] In Table 2, we have also provided the expenditure results estimated by ANOVA.)

The issue of specification searches for multiple regression is a serious one because of the danger of overfitting the data and overstating the precision. We were quite concerned with this issue as it applied to the statistical model of the demand for health services. To check our choice of model, we conducted split-sample analyses of alternative models.[69,71,72,82] That study convinced us that we had not made an inappropriate choice of model. As for the specification of the independent variables, it is quite difficult to get the "correct" standard errors in cases where variables are added or dropped based on preliminary analysis. In principle, one could bootstrap or cross-validate the whole process, but that would have been a prohibitively expensive approach. The dearth of such analyses in the literature testifies to the difficulty of carrying them out.

t-Statistic Inflation

Welch et al. allege that

> "there appears to be a substantial systematic inflation of the t-values on which the comparisons are based for which we see no obvious explanation . . . depending upon insurance plan, the t-values for the one-part and two-part models uniformly exceed Cochran's t . . ."

Welch et al.'s assertion requires a set of modeling assumptions that obviously do not apply to our analyses:[69,71,72,82–86] linear models, with untransformed dependent variables and zero correlation between plan and other explanatory variables.

When analysis of variance is used, the SE for the contrast of two groups can be obtained by taking the square root of the sum of the squared SE for the means of the two groups. However, the same calculation does not generally provide the correct SE of the contrast when analysis of covariance (ANOCOVA) or other multiple regression methods are used. For ANOCOVA, the sum of the variance formula is incorrect if the non-plan independent variables are not perfectly orthogonal to (uncorrelated with) the indicator variables for the plan groups, or if the dependent variable is transformed (e.g., if one estimates the log of dollars, rather than dollars directly).

In the Health Insurance Experiment, the correlation between plan and other variables is not exactly zero for two reasons. First, in Seattle, the GHC controls differ on a number of characteristics (e.g., sex and age) from the GHC and fee-for-service experimentals. Second, the HIE randomized subjects within sites, but the mix of plans varied by site. With a different mix of non-plan characteristics in different sites, this implies some small imbalance across plans. But as we have shown above, correcting for site imbalance yields only a modest change in results.

A simple ANOCOVA example may help illustrate these points. First, consider a simple model where we regress raw (untransformed) visits on age and a dummy for plans with copayment:

$$\text{visits} = \alpha_0 + \alpha_1 \cdot \text{pay} + \alpha_2 \cdot \text{age} + \epsilon$$

where ϵ is an i.i.d. error term. In this case, the SE for the contrast of the free and pay plans is the SE of the estimated coefficient α_1:

$$\text{var}(\alpha_1) = \sigma_\epsilon^2 / [\text{n} \cdot \sigma_{\text{pay}}^2 \cdot (1 - \rho^2)]$$

where ρ is the correlation between pay and age. In general, the sum of the variances differs from the variance of the contrast by

$$2\sigma_\epsilon^2 \cdot \rho^2 / \text{n} \cdot (1 - \rho^2)$$

The variance of the contrast can be different from the sum of the variances for the two predictions if a transformed dependent variable is used in the regression and a retransformation is made for predictions. To illustrate this, let us assume that the dependent variable is the square root of visits, rather than visits. Then the prediction for the free plan is

$$\text{free mean} = \text{mean}[(\alpha_0 + \alpha_2 \cdot \text{age})^2] + \sigma_\epsilon^2$$

while that for the pay plan is

$$\text{pay mean} = \text{mean}[(\alpha_0 + \alpha_1 + \alpha_2 \cdot \text{age})^2] + \sigma_\epsilon^2$$

The difference in the two plan means is

$$\text{mean(pay} - \text{free)} = \text{mean}[(\alpha_0 + \alpha_1 + \alpha_2 \cdot \text{age})^2] - \text{mean}[(\alpha_0 + \alpha_2 \cdot \text{age})^2]$$

The variance of the difference is less than the sum of variances of the mean for two reasons. First, it does not include terms for the variance of the estimated variance of the error term. Second, it does not include the terms for the cross-product $\alpha_1 \cdot (\alpha_0 + \alpha_2 \cdot \text{age})$ introduced by the retransformation from square root visits to visits.

Using a log transformation (as we did) leads to a similar result.

Thus, the Welch et al. assertion that our t-statistics are inflated is based on a set of modeling assumptions that do not apply to our analyses: linear models, with untransformed dependent variables and zero correlation between the plan variables and all other independent variables. In fact, our plan variables are not, and could never be, perfectly orthogonal to all other variables, and our methods involve nonlinear methods and transformations. As we have shown,[69,71,72,87] these methods provide us with more robust and precise estimates of the plan response than can be achieved by ANOVA, ANOCOVA, or several other models we considered.