The Design of the RAND Health Insurance Experiment: A Retrospective

Joseph P. Newhouse

Harvard Kennedy School

Department of Health Care Policy, Harvard Medical School

Department of Health Policy and Management, Harvard T.H. Chan School of Public Health

National Bureau of Economic Research

Acknowledgment: I would like to thank Larry Orr for his comments on a draft of this paper, as well as his assistance and counsel in designing and implementing the Experiment. Many other persons deserve credit as well, as evidenced not only by the many co-authors of Newhouse and the Insurance Experiment Group (1993) but also the lengthy list of acknowledgments in that book.

Abstract

This paper, prepared as part of a special issue on multi-armed experiments, describes the design of the RAND Health Insurance Experiment, paying particular attention to the choice of arms. It also describes how the results of the Experiment were used in a simulation model and, looking back, how the design might have differed and how the results apply today, four decades after the Experiment was conducted.
The RAND Health Insurance Experiment (hereafter the Experiment) is a well-known example of a multi-arm experiment (Newhouse and the Insurance Experiment Group 1993; Gruber 2006; Aron-Dine, et al. 2013). Its primary aim was to estimate the effects from varying the amount of cost sharing in health insurance plans, in particular varying coinsurance rates or the proportion of medical spending paid by the patient when using care. I was the Principal Investigator for the Experiment, the field work for which was carried out between 1974 and 1982.

Any experiment raises many issues of how well the results generalize to actual policy. I touch on some specific issues that the Experiment raised below, but there is an enormous literature on the general subject. Papers an interested reader might consult include (Deaton 2010; Heckman 2010; Imbens 2010; Banerjee and Duflo 2011; Card, et al. 2011; List 2011; Ludwig, et al. 2011; Ravallion 2012; Rosenzweig 2012).

Despite the time that has passed since the Experiment was completed and the natural issue of the degree to which its findings remain applicable, its results on demand elasticities are still widely used. In 2006 the Congressional Budget Office said, “Even though the RAND data were gathered several decades ago, the study’s findings remain relevant and are widely relied on by analysts…” (Congressional Budget Office 2006). Two years later it called the findings “the best available evidence about the effects of cost sharing…” (Congressional Budget Office 2008).

In what follows I describe the rationale for several of the Experiment’s design decisions; additional material is in (Newhouse 1974), which describes the initial design. (Newhouse and the Insurance Experiment Group 1993) describes the final design, which was close to the initial design but incorporated some mid-course corrections described below. I also describe below one use of the Experiment’s results, a simulation model to extrapolate results from the Experiment’s insurance plans to other insurance plans. I conclude with some thoughts on what in retrospect I would have done differently and how one might think about the applicability of the Experiment’s results from four decades ago to today’s health care financing and delivery. Before turning to issues around the design, however, I briefly describe how the Experiment came about.

The Genesis of the RAND Health Insurance Experiment

In August 1969 the Nixon Administration proposed the Family Assistance Program, a form of guaranteed income for poor families, with amount of government provided assistance falling with earnings. This form of income support, or a negative income tax, was favored by economists across the political spectrum as an anti-poverty policy (Friedman 1962; Tobin, et al. 1967). In hearings on the proposal Senator Russell Long (D-LA), the chairman of the Senate Finance Committee, asked about how the Administration’s proposal might affect the Medicaid notch, whereby individuals just below an income or asset threshold were eligible for health insurance from Medicaid but those just above it were not. The Nixon Administration took seriously the issue of how Medicaid should integrate with its Family Assistance Program proposal and formed a task force to address its policy toward Medicaid.

Larry Orr, an economist from the Office of Economic Opportunity, the agency created by the Johnson Administration to implement its War on Poverty, was assigned as a staff member to the Deputy...
Under Secretary for Policy, who was a member of the task force. Although the Medicaid statute, Title XIX of the Social Security Act, did not allow for any cost sharing, Orr raised the issue of whether some cost sharing should be considered. Based on the initial discussions of the task force, Orr concluded that little was known about the effect of cost sharing on the utilization of medical care.

At this time the Office of Economic Opportunity was sponsoring a series of randomized experiments in income support or negative income taxes (Cain and Watts 1973; Rivlin and Wiener 1988). Although these experiments had not concluded, enough experience had been gained that the Office of Economic Opportunity viewed them as a promising method to develop evidence-based social policy.

In 1969 I had submitted a grant to the agency that was a forerunner of today’s Agency for Healthcare Research and Quality to study the effect of cost sharing on the demand for medical care using observational data. My grant application found its way to Orr, and he came to visit me to explore the possibility of an experiment in health insurance and the effects of cost sharing for low income persons. I had not contemplated an experiment – if nothing else the necessary budget would have greatly exceeded monies available through the usual grant mechanisms – but Orr and I agreed that an experiment could be useful and so the Office of Economic Opportunity gave me a small grant to design such an experiment. With the help of colleagues at RAND and economists at the Office of Economic Opportunity I did so. Based on that design, the Office of Economic Opportunity subsequently decided to hold a competition to implement the experiment, which our group won.

Although the original design was intended to ascertain effects of cost sharing only in a poor and near-poor population, consistent with the Medicaid focus and the mission of the Office of Economic Opportunity, the Nixon Administration in 1971 proposed a national health insurance plan that contemplated near universal coverage - and in fact bore many similarities to the Affordable Care Act (Obamacare) that was enacted four decades later (Blumenthal and Morone 2009; Altman and Shactman 2011). The Nixon proposal for national health insurance included cost sharing for services. The principal Democratic alternative to the Nixon proposal had been introduced by Senator Edward Kennedy (D-MA) and Representative James Corman (D-CA). It called for universal coverage with no cost sharing, i.e., free medical care. These national health insurance proposals for universal coverage with their different stances on cost sharing naturally led to interest in the effects of cost sharing in a general population, and the target population for the Experiment was duly modified to include a general population as I describe below.

The reader should consult (Newhouse and the Insurance Experiment Group 1993) for a summary of the Experiment’s results, but it may help in reading what follows to know that the two extremes of cost sharing included in the Experiment, no cost sharing and a large income-related deductible, showed around a 25-30 percent difference in use and that for most persons there were no detectable effects on health outcomes from this difference in use of medical care. Confidence intervals were generally small, so large effects on health outcomes could be ruled out. The sick poor, approximately 6 percent of the Experimental population, were an exception; those with hypertension (high blood pressure) were less well controlled if they were assigned to a cost sharing plan and as a result were predicted to have a 16 percent higher likelihood of dying at a given time.
The Design of the Experiment

Of course, the design needed to specify the variation in cost sharing among the experimental insurance plans, but there were also many other design decisions. What medical services should the plans cover and at what rates should hospitals and physicians be reimbursed? Beyond those basic decisions about the details of the insurance plans, how could refusal and attrition be minimized? What was the population to be sampled, and should the sampling be proportional? How many persons should participate, and how should they be allocated among plans? How many sites should be included, and how should specific sites be chosen? How might methods effects be measured?

As in any experiment, in answering these design questions it was crucial to keep two questions firmly in mind:

- What policy questions were we trying to answer?
- How would we analyze the data the Experiment would produce?

Although an initial aim of the Experiment was to estimate demand responses or price elasticities, it was also a goal to say something about effects on health outcomes of any variation in services. There had been a decades-long, heavily ideological debate about whether cost sharing deterred necessary or unnecessary services, or, for those who preferred grey to black and white, the mix of effects on both. In standard welfare economics simply estimating demand and cost functions suffices to estimate the degree of inefficiency (deadweight loss), but it is highly problematic to attach any normative significance to the demand curve in the medical care context because of potential rents in provider reimbursement, asymmetric information, agency issues, and the many behavioral biases persons bring to dealing with both uncertainty and with health care (Bernheim and Rangel 2004; Beshears, et al. 2008; Kahneman 2011; Ericson and Sydnor 2017). Therefore in addition to estimating how use responded to cost sharing, a key aim of the Experiment was to assess the effects of cost sharing on health outcomes, including both self-reported outcomes and physiologic measures such as blood pressure, as well as effects on quality of care and patient satisfaction. This, however, implied a major investment in measure development, since existing measures were, to put it mildly, rudimentary.

The Range of Cost Sharing

What the cost sharing arrangements should be in the Experimental plans depended on the policy questions for which answers were sought. In addition to the Administration proposal and the Kennedy-Corman bill, legislation sponsored by Senators Russell Long (D-LA) and Abraham Ribicoff (D-CT) called for a catastrophic insurance plan with a large deductible, or a considerably higher degree of cost sharing than the Administration proposal envisioned. Given these three proposals, it was clear that the Experiment’s cost sharing should span the range from no cost sharing to catastrophic insurance.

Where should plans be located within this range? An economist’s first instincts might be to estimate a continuous function over this range, which might, for example, assign persons coinsurance rates that varied in quasi-continuous fashion over the 0 percent (no cost sharing) to 100 percent (no insurance benefits) range. The design group rejected this notion because it was clear that the interest
was greater in certain ranges of the coinsurance space than others, namely coinsurance rates that would roughly correspond to the three legislative proposals just described.

The Administration proposal mandated that all employers provide subsidized plans to their workers. It would have replaced Medicaid by a plan that covered all poor families with children with graduated cost sharing tied to income. Other individuals would be eligible for state-based insurance pools. At that time employment-based insurance plans typically had an initial deductible, followed by a coinsurance rate that was usually 20 or 25 percent. As a result, the Experiment’s design called for a plan at 25 percent coinsurance, as well as at 0 (free care or no cost sharing), like the Kennedy-Corman proposal, and a plan with an initial 100 percent coinsurance rate, like the Long-Ribicoff proposal. The desire for some information on responses at coinsurance rates between 25 and 100 percent also led us to include a 50 percent coinsurance rate plan.

All the plans with cost sharing had an annual cap on out-of-pocket spending that was scaled down for low-income individuals and families. The 100 percent coinsurance plan with a cap was thus a large income-related deductible. After the first year in the first site, the 100 percent rate was changed to 95 percent to give families on that plan who did not expect to exceed the deductible a financial incentive to file a claim.

The cap on out-of-pocket spending was included both because of its policy relevance and because Kenneth Arrow’s seminal work on the economics of medical care had shown that, provided there was no behavioral response to insurance, i.e., no moral hazard, an optimal insurance plan was full insurance above a deductible, meaning a cap on out-of-pocket expenses (Arrow 1963). There was also a critical practical reason for including such a limit, namely limiting adverse selection in enrollment. I describe this further below.

Having experimental plans that roughly mimicked the cost sharing in prominent legislative proposals as well as in policies already in the market was not only policy relevant, but also may have made it easier to obtain consent to enroll. Scam artists sometimes posed as insurance sales persons, especially in low-income neighborhoods, so potential enrollees might have harbored doubts about the bona fides of a group that asked them to enroll in a plan with 18 percent or 81 percent coinsurance or some other rate that was not found in any actual policy.

Although common in actual plans of the day, we did not include an initial relatively small deductible in the plans with coinsurance because at the time we did not have methods to extrapolate results from plans in which unit price varied with total expenditure. Although the upper limits in all the cost sharing plans meant this feature was present in the Experimental plans, which posed an analytical issue, adding an initial, rather modest deductible would only have compounded the problem. After the Experiment began, however, we developed methods to analyze the behavior of persons with such plans, as we describe below when discussing the simulation model.

Although a principal Experimental objective was to estimate how varying cost sharing affected the use of medical services and total medical spending, or in technical terms an own-price elasticity, there was an additional debate about differential coverage of certain medical services, especially the
coverage of inpatient services relative to physician office visits. In the early 1970’s coverage was much more complete for inpatient than for outpatient services. Many policies in fact excluded physician office visits from coverage. A common argument was that the lack of coverage of outpatient services led persons to defer seeking care until they became sufficiently sick to require more expensive hospital treatment. That argument in turn led some to conclude that extending coverage to outpatient services would both save money as well as improve health outcomes (Roemer, et al. 1975). In economic terms, this debate was about whether inpatient and outpatient services were substitutes or complements, or about the sign of cross-price elasticities.

There was a similar argument around coverage of mental health services, which were generally excluded from coverage. If they were covered at all, the cost sharing was generally higher than for other services on the assumption that demand for them was more responsive to insurance coverage, i.e., more elastic. For example, the coinsurance rate for mental health services might be 50 percent, whereas for other covered services it might be 20 or 25 percent.

The Experiment’s design addressed these differential coverage issues in two ways. With respect to better coverage of inpatient services, most Experimental plans covered all services at the same coinsurance rate, but in one Experimental plan with 100 percent coinsurance (later changed to 95 percent) the cost sharing applied to outpatient services only and inpatient services were free, thus emulating those policies of the time that had better coverage of inpatient services. In that plan it seemed preferable to make the annual out-of-pocket spending limit a flat $150 per person (to a maximum of $450 for a family) for reasons of administrative simplicity, whereas in the other cost sharing plans the out-of-pocket limit was scaled to household income.\(^5\)

With respect to higher coinsurance rates for mental health services, the design included a plan that covered mental health services at 50 percent coinsurance and all other services at 25 percent coinsurance.\(^6\) This set up a contrast with plans that covered all services at either 25 percent or at 50 percent.

An important decision was where to set the out-of-pocket limit. Because there was no strong consensus around where a national plan would set such a limit or how any such limit might change over time, it was randomly varied between 5, 10, and 15 percent of household income, subject to a $1,000 maximum. The $1,000 would be somewhat over $4,000 in 2018 dollars if using the all-items Consumer Price Index to adjust and around $17,000 using the increase in medical spending per capita to adjust. The reason for choosing $1,000 we come to below (in addition to its being a round integer).

Another further question was whether there should be a control group. There were two arguments for a control group. One was simply optics; a control group permitted a short response to those who mistakenly thought a valid experiment had to have a control group as opposed to randomly assigned comparison groups. A substantive reason for a control group was that if any national health insurance legislation were to be enacted, the insurance of a control group would be directly affected by that plan, whereas the Experiment participants could be held out and used as a classic control group. Although potentially attractive, a pilot study in the first site showed a control group was not feasible;
many hospitals and physicians were simply unwilling to fill out two claim forms, one for the control group member’s actual insurance policy and a second for the Experiment. As a result, it proved impossible to obtain comparable data on use among a control group, and a control group was simply dropped from the design.

We enrolled entire family units rather than randomly selected family members. Although this decision decreased the statistical efficiency of the estimated results because of intra-family correlation in use, we deemed it impractical to enroll only selected family members. Moreover, any national plan with an annual out-of-pocket limit might relate that limit to family income, as indeed the Nixon Administration plan did.

In 1970 the Nixon Administration proposed including Health Maintenance Organizations (HMO’s) in Medicare and Medicaid, which led to the HMO Act of 1973 (Blumenthal and Morone 2009). Although HMO’s, in particular the Kaiser Permanente health plan, were positively viewed by many reformers, there was little hard evidence on their effects. For that reason there was also an HMO arm of the Experiment.

Finally, the Experiment did not include an uninsured arm. Although it was of great policy and intellectual interest to contrast effects of being uninsured with having insurance of varying comprehensiveness, it was both ethically and practically impossible to randomize families who had health insurance to a no insurance arm.7

What Services to Cover

The Experimental plans covered almost all personal health care services because failure to cover a service would have meant no claims data for that service would be available to analyze.8 Not only medical but also dental services were covered. Dental insurance, however, was then and even today remains less generous than medical insurance.9 For that reason the 50 percent coinsurance rate that applied to mental health services in the split 25 percent/50 percent coinsurance plan also applied to dental services. In the other plans dental care was covered at the same coinsurance rate as medical care.

Although the Experiment’s coverage of medical services was broad, there were certain limits and exclusions, especially for services that a national plan was unlikely to cover and where one was unlikely to observe steady state demand because satisfying an accumulated stock of demand could last several years or even the entire duration of the Experiment. Multiple-sessions-per-week of psychoanalysis was an example, so psychotherapy was limited to 52 visits per year, despite a plea from the psychoanalysts’ professional organization to drop the limit. Cosmetic surgery other than for trauma-related incidents that occurred during the Experiment was another example, as was orthodontia for pre-existing conditions. Vision benefits were limited to new lenses and examinations not more frequently than every year and new frames not more frequently than every two years. Although the primary reasons for these exclusions were the lack of policy interest and the likely inability to observe steady state demand in a finite period, the exclusions did reduce cost somewhat.
Reimbursement Rates

To carry out the Experiment required setting up and operating a small insurance company, and therefore RAND contracted with an existing third party administrator to process claims. With respect to reimbursing providers, the administrator followed its usual practice of paying billed charges and negotiating a lower charge if it deemed the fee excessive. Although no record of such negotiations was kept, this happened rarely, in well under one percent of the cases. Prior authorization was also much less common than today, and the Experiment only used it for relatively expensive dental services.

If one were conducting the Experiment today, one would not pay billed charges and would employ prior authorization for some medical services. At the time of the Experiment, however, health insurers were relatively passive, usually reimbursing medical bills subject to limits on unit prices. Today’s networks and tiered formularies had not been developed. Apart from a few HMO’s, an insured patient could seek care from any physician and pay approximately the same amount out-of-pocket. In short, the passive reimbursement policy of the Experiment, which largely mimicked the health insurance practice of the time, would not be relevant in today’s world of networks, prior authorization, and utilization management. I return to this difference from today’s insurance in the concluding section.

Minimizing Refusal and Attrition: The Participation Incentive and Completion Bonus

The cap on annual out-of-pocket spending in all policies with cost sharing made it feasible to calculate a side payment such that no family could be financially worse off from participating. The side payment, called the Participation Incentive, thereby minimized adverse selection against enrolling in the cost sharing plans among the great majority of families who had health insurance at the time of enrollment. The amount of the side payment equaled the worst case outcome for the family under its Experimental plan relative to its existing insurance plan. For example, if the cap were $1,000, the Participation Incentive equaled $1,000 less the amount the family would have paid out-of-pocket under its existing insurance plan in the worst case scenario. This amount, divided by 12, was paid monthly.

Paying the Participation Incentive monthly was another design choice. The Incentive could have been paid at the end of a year of participation, but that might have weakened the credibility of the enrollment offer and led to a higher refusal rate. It could also have been paid at the beginning of the year, but that risked greater attrition from a family pocketing the money and returning to its prior plan. (For ethical reasons and potentially to increase enrollment, families had the right to return to their prior plan at any time even though it was never in their financial interest to do so.)

Setting the annual out-of-pocket limit at $1,000 attempted to balance two conflicting arguments. A higher figure would have increased the range of variation in cost sharing, although potentially above the range of policy relevance, but it would also have increased the budget by increasing the worst case for most families and hence the amount of the Participation Incentive payments.

At the time the Participation Incentive payments were criticized on the grounds that they would not be part of a national health insurance program and thus the Experiment would not replicate what
such a program would look like. At the time we responded that these payments would have a negligible
effect on utilization. Our rationale was that the payments represented an income effect on demand,
which, given the income elasticity of medical care and the mostly small percentage increases in
household income that the payments represented, would cause negligible distortion in the results. We
not only tested that prediction directly, since we wanted to adjust for whatever modest effect there
might be, but also built experimental variation into the design to estimate the size of any effect; we
describe that feature below.

Long after the Experiment concluded, however, high deductible plans in conjunction with Health
Savings Accounts and Health Reimbursement Accounts were introduced and became widespread; 28%
of those enrolled in employment-based insurance were in such plans in 2017 (Claxton, et al. 2017).12
These Accounts, when funded by the employer, are similar to the Experiment’s Participation Incentive
in also creating an income effect on use (Chernew and Newhouse 2008). Thus, in a way that was not
anticipated at the time of the design, the combination of the Experiment’s cost sharing plans with side
payments appears more relevant to today’s world than cost sharing plans with no side payments would
have been.

The design also addressed a potential terminal condition problem. Because payments were
made monthly, a family could be worse off from continuing in the Experiment during its final year of
participation. For example, it might have no medical expenses for the first half of that year and then a
large medical bill that its insurance at its place of employment would pay in full. If the family remained
in the Experiment, its remaining Participation Incentive could be less than its out-of-pocket expense for
the medical bill, so it would be to the family’s financial advantage to withdraw from the Experiment and
return to its employment-based insurance plan. To address this issue, the Experiment also paid a
Completion Bonus, an additional amount equal to the family’s worst case that was paid at the end of the
Experiment if the family completed its period of participation.

The annual out-of-pocket limit was reduced from a maximum of $1,000 to $750 for most of the
Experiment in the 25 percent coinsurance plans. This change, which reduced Participation Incentive
payments by reducing the worst case, was made because of budgetary pressure, but the results of the
Experiment were minimally affected.13

**What Population Should Be Sampled?**

As mentioned at the outset, the initial population of interest was the poor and near-poor, but
after the 1971 Nixon Administration proposal for national health insurance, the interest became the
general under 65 population. We excluded those eligible for Medicare because the Nixon
Administration had no intention of changing Medicare and did not want to risk signaling that it did. We
also excluded households with income above $25,000 (1974 $) because the Administration did not want
to make transfer payments to upper income households. The $25,000 figure excluded about 3 percent
of the population in the sites where the Experiment operated, so it was only a minor limitation on
generalizability.
There remained the question of whether sampling should be proportional or whether the low-income group should be oversampled. The policy interest in the low-income group was greater, both because of the question of cost sharing in Medicaid or whatever program would replace it and also because higher income groups tended were more likely to be insured so that effects of a new national health insurance program would be larger among low-income groups. For these reasons in the first site, Dayton, Ohio, low income groups were oversampled.

Before the second site, Seattle Washington, became operational, we made an argument that income was a noisy variable and a that sufficiently high oversampling rate on such a variable could lose statistical efficiency even for the favored group (Morris, et al. 1979). The intuition is that someone with a low income today, such as a graduate student or someone temporarily unemployed, might well not have a low income next year and conversely. Although in the case of the Experiment we estimated that there was some gain in precision for the low income group by modestly oversampling it, such oversampling would complicate all analyses because of the need to include sampling weights for what the RAND group thought was a modest gain in precision for the low-income group (and a modest loss in precision for higher income groups). This argument for proportional sampling carried the day in the second site, but in the subsequent sites, which came on stream later, the government asked that we revert to oversampling low income groups.

**How Long Should the Experiment Run?**

Several considerations went into a decision to randomly split the sample into two groups whose enrollment periods varied. One group, comprising 70 percent of the participants, participated for three years, and the remaining 30 percent participated for five years. Why three and five years? The design generally sought to optimize the amount of information subject to an overall budget constraint. For purposes of estimating effects on demand, two independent individuals participating for one year were more useful than one individual participating for two years because of the positive correlation of utilization across years for the same individual. For the purpose of estimating demand the gain in statistical efficiency from individuals participating for more than two years and especially for more than three years was small. More persons participating for a shorter period of time also meant that the results of the Experiment would be available sooner. But the period of participation had to be long enough to detect a beneficial health effect of lower cost sharing if such an effect existed. Not surprisingly there were no data to estimate how long a period might suffice. We arbitrarily decided that three years was likely to be a sufficiently long period of time for health effects to manifest themselves, and we allocated a greater proportion of the sample to the three year group because of the more precise estimates of standard errors that that allocation permitted and because they were less costly. Having some persons participate for five years, however, gave some protection against a three year period being insufficient for health status effects to appear.

The main argument for an enrollment period longer than three years, however, was the high likelihood of transitory effects on utilization at the beginning and end of the Experiment, which would interfere with the goal of measuring steady state demand. Using two different lengths of participation permitted direct estimation of any transitory effects. In some sites we started both the three year and
five year groups at the same time, so comparing behavior of the three year group in their final year with that of the five year group in their third year provided an estimate of effects from anticipating the end of the Experiment. In other sites we started the five year group two years ahead of the three year group, so any initial transitory effects could be estimated by comparing the five year group in their third year with the three year group in their first year. Fortunately, such transitory effects turned out to be quantitatively modest except for dental services.

**How Many Persons on Each Plan?**

The initial analysis plan for utilization and spending effects envisioned estimating plan means (ANOVA) or, to gain statistical precision, means adjusted by some standard covariates such as age (ANOCOVA). The optimal allocation of a sample when one is estimating means of discrete groups is in proportion to \( \sqrt{w_i/c_i} \), where \( w_i \) is the weight assigned to the mean of the \( i \)th group and \( c_i \) is the marginal cost of enrolling another household in the \( i \)th group. This formula was used to allocate the sample to plans.

In the first site the 0, 25, 50, and 95 percent coinsurance plans were given equal weight and the Individual Deductible plan, the plan that covered outpatient services less well, a lower weight. After the Experiment had begun in the first site but before it began in the second site, however, a more sophisticated analysis of demand based on episodes of illness was developed, which gave us more confidence in our ability to analyze data from plans in which unit price changed with total spending. This led us to increase the weight on the 95 percent coinsurance and individual deductible plans, the plans where the caps on annual spending were most likely to be binding.

Also after the first site had begun, the group analyzing health outcomes had made substantial progress in developing measures and, in contrast to the group analyzing utilization and spending, wished to compare persons on the free plan with all persons on cost sharing plans taken as a group. Thus, from the health status group’s point of view, the free plan would have comprised approximately half the sample. The final allocation of the sample to plans gave the optimal allocation for the utilization analysis half the weight and the optimal allocation for the health status group the other half. Further details of the rationale for the allocation of sample by plan are given in Appendix B of (Newhouse and the Insurance Experiment Group 1993).

**How Many Sites?**

Adding a site entailed fixed costs. Field offices had to be opened and maintained for the duration of the field work. Local personnel had to be hired. Local medical and political leaders had to be informed. Additional sites reduced between-site variance, but the greater fixed costs meant fewer total participants and so greater within-site variance.

A general design would seek to minimize the sum of between- and within-site variance subject to a budget constraint using the formula...
\[
\frac{\sigma_b^2}{n_s} + \frac{\sigma_w^2}{n_f}, \text{ subject to } C_s n_s + C_w n_f = B, \quad \text{where } \sigma_b^2 \text{ and } \sigma_w^2 \text{ are the between-site and within-site variances respectively, } n_s \text{ is the number of sites, } n_f \text{ is the number of families, } C_s \text{ is the marginal cost of a site, } C_w \text{ is the marginal cost of a family, and } B \text{ is the budget constraint.}^{16} \text{ Although } B \text{ was left unspecified, the Secretary of the Department of Health, Education, and Welfare, in deciding that the Experiment should go forward, stipulated that 2,000 families should be included in the fee-for-service arms of the Experiment. Given that value for } n_f \text{ and our estimates of the variances and costs, the optimal number of sites was four, although the total variance would not have increased very much for any larger number of sites up to nine. The cost of adding two rural sites to nearby metropolitan sites, however, was not large because there was no need for an additional field office, and so we expanded the number of sites from four to six while maintaining the number of families at 2,000.}

Which Sites?

There remained the question of which six actual sites should be chosen. We chose the sites purposefully rather than at random to assure obtaining variation in a number of characteristics that would give some face validity to a claim of generalizability. Because of the geographic variation in the pattern of medical care, we wanted at least one site in each of the four Census regions. Because the sophistication of the medical care delivery system varied with city size, we wanted varying city sizes, including both metropolitan and non-metropolitan areas. At least one site had to have a well-established HMO that was amenable to participating in the Experiment.

Finally, one aim of the Experiment was to understand the nature of any non-price rationing mechanisms that a national health insurance plan might activate. Although there was a wide range of existing estimates of the price elasticities of demand from observational data, most estimates implied that the Kennedy-Corman proposal with no cost sharing and universal coverage would likely increase demand beyond the short-run capacity of the delivery system. It was clear, however, that the Experiment would not stress the delivery system in any site and so would not trigger any rationing mechanisms. Not only would the Experiment’s participants represent a small share of any site’s population, some families would have more cost sharing than their prior insurance while other families would have less. Thus, the net change in demand at any site would be negligible.

To shed some light on the issue of non-price rationing mechanisms, we therefore chose sites with varying degrees of excess demand, in effect a small observational study within the Experiment. The measure of excess demand that we used was the wait time in a site for an appointment with a primary care physician for a non-urgent problem. Before choosing the actual sites we carried out a survey to measure wait times for an appointment in many sites; across five of the six sites that we ultimately chose, wait times varied from 4 to 25 days; the sixth site was a rural site where physicians did not make appointments and operated on a first-come, first-served basis.\(^17\) Despite the wide range of demographic, economic, and social differences among the six sites we chose, the utilization response to the Experimental plans was remarkably uniform across them. The only relationship with wait times to an appointment was greater use of the Emergency Department in sites with longer waits.

Measurement of Methods Effects
Virtually all experiments run the risk that some features of the design will not be replicated in an actual program and so could cause the results to differ from those of an actual program. In the case of the Experiment, the likelihood of transitory demand and the Participation Incentives have already been mentioned as examples of such features. The solution in those two cases was to build variation into the Experiment’s design to allow estimation of the relevant effects. In the case of transitory demand, as already described, the approach was to split the sample and to stagger the start of enrollment across sites. In the case of the Participation Incentive, the approach was to randomly give some families amounts above their worst case, allowing a comparison with families that received only their worst case.

A similar approach was taken with respect to obtaining baseline measures of physiologic biomarkers such as blood pressure and cholesterol levels. It was an ethical imperative to notify the participants of any abnormal results – and if sufficiently abnormal to try to facilitate their getting immediate treatment. Notifying participants, however, could induce demand for medical care. To measure any induced demand we decided to split the sample such that only a random 60 percent of the participants received a baseline screening exam.

This decision was not without cost because baseline measures of health status greatly improve statistical power. Absent medical intervention most biomarkers such as blood pressure and cholesterol levels do not much change over a three to five year period, and the availability of a baseline measure absorbs much of the substantial between-person variation in these measures. Despite the loss of power from not having baseline values for the entire sample, the measurement of the actual outcome measures turned out to be estimated with sufficient precision for policy purposes.¹⁸

Yet another measurement issue was the frequency of data collection directly from the participants. Collecting such data was necessary because not all the information we sought would be available on claims forms and would be difficult to recall if the participant was asked about it well after the fact; examples include data on days lost to activity impairment from illness and time spent seeking medical care. To obtain this information we initially sent a mail questionnaire weekly to some randomly selected families and to others bi-weekly. There seemed to be only a small difference in the quality of the data between the two groups, so after the first year in the first site this questionnaire was sent bi-weekly for the duration of the Experiment. To determine whether this questionnaire itself stimulated use, we did not send it to a random 25 percent of the sample in four of the sites in their first year of participation.¹⁹

Although different in spirit from the type of methods effects just described, another possible distortion was potential underfiling of claims in the 95 percent coinsurance plan. This related to a question of whether the Experiment was simply trying to measure insurer payout or whether it was trying to measure all use of medical care. If the goal was simply measuring insurer payout, underfiling was of no concern, but if the goal was measuring use of the medical care system, it was. The question we deemed more important was use of the medical care system, which implied a need to estimate underfiling.
A preliminary step described above was simply to reduce the 100 percent coinsurance rate plan to a 95 percent rate after the first year in the first site, which gave families with that coinsurance rate a financial incentive to file claims. To directly measure underfiling, however, we conducted an audit study of physician use in two sites. The audit showed that 7 to 9 percent of physician visits had not been billed in the 95 percent coinsurance plan; this amount of underfiling did not change any qualitative conclusions.

A Simulation Model for Spending and Use of Medical Services

Although analysis of mean plan spending was straightforward, extrapolation to other price schedules, including being uninsured, was not. When the Experiment was designed, the assumption was that relatively few families would spend amounts that approached or exceeded the upper limit, so that ignoring the upper limit and treating spending differences among plans as simply a response to coinsurance would approximate an estimate of a pure response to price.

That assumption would not hold for small deductibles, however, so, as mentioned above, the Experimental plans did not replicate a common insurance policy at the time, one with an relatively small deductible, followed by coinsurance, followed by an upper limit on covered spending. What medical care spending might be under such a plan, or for that matter in a plan with an out-of-pocket limit that was substantially smaller than that in the Experiment, was not clear. Furthermore, if a plan covered a different mix of services, accrual of spending toward any deductible or out-of-pocket limit would differ and so price effects would likely differ.

During the design phase of the Experiment we assumed that users of the Experiment’s results could extrapolate from estimates of mean spending by as they best saw fit. Shortly after the Experiment began, however, we developed a theory of behavior when a consumer faced non-constant unit price schedules (Keeler, et al. 1977). That theory implied that the unit of observation to analyze utilization behavior given a non-constant unit price schedule was not annual per person spending but rather episodes of illness. Using this result, the RAND group built a simulation model from the Experimental data based on episodes of illness. Estimates from this model made it unnecessary to extrapolate from plan means in an ad hoc fashion, because the model could be used to estimate effects on spending of plans with varying initial deductibles, coinsurance rates, out-of-pocket limits, and covered services. It could also account for whether deductibles and out-of-pocket limits were specified at the individual or the family level.

The simulation model, which is summarized in (Keeler and Rolph 1988), began by grouping an individual’s claims into one of five different types of episodes, hospitalization, well-care, routine chronic outpatient care, outpatient care to treat acute episodes or flare-ups of chronic disease, and dental. It allowed the response to plan (price elasticity) to differ for each type of episode and the propensity to initiate episodes to differ across persons. The model assumed that the total cost of each type of episode was known at the time the episode began, for example, when a woman made her first visit for a pregnancy she would know the ultimate cost. Episodes of routine care for chronic diseases such as hypertension or diabetes were assumed to begin on the first day of the annual accounting period and
last the entire year. Flare-ups of chronic disease, however, were dated to when they occurred during the year, as were acute episodes, hospitalization, well care, and dental episodes.

The model required estimating how use and spending responded as a person's total spending moved toward a change in unit price, for example, as spending approached the upper limit or moved past it. A rational maximizer would use at higher rates the greater the probability of exceeding an upper limit because of a higher probability that the marginal unit of care would be free. Once over the limit the rational maximizer would treat medical care as being on sale for the remainder of the accounting period and use at an even higher rate than those on the free care plan because at the beginning of the next accounting period medical care would again be expensive. Empirically, however, the average Experimental participant turned out not to be the rational maximizer of economic theory, but instead did not much anticipate the change in unit price, i.e., was myopic before the annual upper limit was exceeded. Once over the limit the average participant spent at roughly similar rates to those on the free plan, that is, while participants recognized that care was now free and increased their use of care, they did not increase it above the rate of those on the free plan, i.e., they did not treat care as temporarily on sale.

Empirically almost all of the actual response to the different insurance plans was driven by variation in the frequency of treated episodes, meaning the less costly was medical care, the greater was the number of episodes of each type. Cost per episode had only a very small response to price. In other words, the main effect of cost sharing was on the consumer’s decision to initiate care; once under treatment, cost per episode (for a given type of episode) was nearly constant.

Although the plan to analyze episode of illness, or more properly episodes of treatment, had not been conceived at the time of the initial design, the ability to analyze the data by episode fortuitously turned out not to require any information beyond that information that was already being collected. In particular, date of service, which was critical for the episode analysis in order to date episodes within the year, was routinely collected as part of the claim form.

The Hindsight of Forty Years

Looking back, what would I have done differently in designing the Experiment? And what do I now make of its results? The answer to the first question is that I would have not materially changed the design, although if I had known at the outset that outcome effects would be concentrated among the sick poor, I would have oversampled that group. At the outset, however, measures of health outcomes for studies such as the Experiment were in a primitive state. As a result, when the Experiment was being designed it wasn’t clear what measures of health status would even be included, let alone what effects they might reveal. Indeed, one of the lasting contributions of the Experiment has been the development of measures of health outcomes and quality of care.22

Of course, I would have made the mid-course changes in the design of the Experiment at the outset to obviate the need for such corrections. For example, I would have started with the 95 percent coinsurance plan rather than changing the 100 percent coinsurance plan to 95 percent after the first year in the first site. Like that change, however, the changes in the design that were made after the
Experiment started were small in magnitude and in my view have not affected any of the major inferences that have been drawn from the results of the Experiment.

From today’s vantage point, however, there are three potential changes in the design that would have warranted greater consideration. First, it would be desirable to know more about what today are called value-based insurance designs (Chernew, et al. 2008; Lee, et al. 2013). These designs are more specific versions of the argument that lowering or exempting some services from cost sharing would lower total spending and/or improve outcomes. While the Experiment’s design did test the effect of varying cost sharing for only outpatient services and only mental health services, today’s value-based proposals are at a finer level of detail, for example exempting from cost sharing specific drugs known to be efficacious for those with a chronic disease such as diabetes in order to encourage compliance. Even if this idea had been proposed at the time, however, it may have been left on the cutting room floor because incorporating it into the design would have required additional treatment arm(s) and thus lowered precision for the arms that were included.

Second, one might have added a second or even a third HMO. The results from the HMO arm of the Experiment have been widely cited, even though they come from just one staff model HMO. Some indication of their generalizability would certainly have been useful. Unlike value-based insurance, however, this idea was considered at the time. The Experiment began with only one HMO because there was much greater doubt about the feasibility of the HMO arm than the fee-for-service arms. That was because the HMO arm required participants to change providers whereas the fee-for-service arms did not. We therefore began with only one HMO to determine if refusal rates would be so high as to make the results not useful for policy purposes. It turned out that refusal bias at the first (and only) HMO site was small.23 We therefore looked into adding a second HMO and found that it would have been feasible to do so without adding another field office.24 At that point, however, the federal government did not wish to increase the budget to accommodate an additional HMO arm.

A third potential design change would have been to include Medicare beneficiaries. Unlike the option of additional HMO’s, that option was off the table from the outset. As mentioned above, the Nixon Administration did not wish to be perceived as potentially considering changes in Medicare. Over the intervening years, however, cost sharing in traditional Medicare has remained above large employer plans (McArdle, et al. 2012).25 Perhaps as a result, well after the RAND Experiment was completed I have occasionally been asked by federal policy makers about carrying out an analogous Experiment among the elderly, but cost and the political sensitivity of Medicare have always precluded it.

How applicable are the Experiment’s results to contemporary policy? Two important caveats stem from changes in the larger environment. The first is the ongoing technological change in medicine and the resulting enhancement of medical capabilities. A dramatic illustration is the fall in the age-adjusted cardiovascular mortality rate, then and now the leading cause of death. Between 1970 and 2014 that rate fell by a factor of three (National Center for Health Statistics 2016). Although this fall is not all attributable to medical care, advances such as medications that better control blood pressure and cholesterol have been estimated to account for 44 percent of the decrease between 1980 and 2000 (Ford, et al. 2007). The ability to save low-birthweight babies has steadily improved. And better medical
care has not only reduced mortality; disability-free life years among the elderly rose between 1982 and 2011 (Freedman and Spillman 2016), one cause of which has undoubtedly been developments such as artificial hips and knees.

With more on offer from the medical establishment, it is certainly conceivable that the response to price today would differ from four decades ago, although subsequent observational studies of this issue have found approximately the same utilization response to price that the Experiment did (Brot-Goldberg, et al. 2017). A similar response seems plausible given the Experiment’s finding that almost all of the effect of cost sharing is on the consumer’s decision to seek care. Because the consumer may well not know the diagnosis before seeking care and therefore be quite uncertain about the benefit from seeking care, it seems plausible that the initial decision to seek care could be less affected by technological change. And although newer technology would likely affect how a physician would treat a given clinical problem after establishing a diagnosis, the physician’s decisions may continue to be relatively unaffected by the amount of cost sharing in the patient’s insurance plan.

Technological change does mean that outcome effects from seeking more care might differ. The Experiment found little or no effect on outcomes in the general population from the additional use in the free care plan, although it did find better blood pressure control among low income hypertensives. For that group the better blood pressure control mattered as mentioned at the outset. Moreover, we not only found no effect of cost sharing on outcomes in the general population but confidence intervals sufficiently small that more than modest effects could be ruled out. At the time our after-the-fact explanation for the lack of a sizeable effect in the overall population was that poor quality care in a generally healthy population had negative effects that offset the positive effects of better access among the subset of the population that was not in good health (Newhouse and the Insurance Experiment Group 1993). Although there was evidence of poor quality care among the Experimental population that supported this view, in the subsequent years a great deal more evidence of poor quality health care has accumulated, making this explanation of no outcome effects for the average non-elderly person more plausible (Institute of Medicine 1999; Institute of Medicine 2001; McGlynn, et al. 2003). An optimist, however, might believe that quality of care has now improved sufficiently such that outcomes today would be more positive from inducing a general population to seek more care by reducing their out-of-pocket cost.

A second change in the environment has been the role of insurers. As already noted, the Experiment was conceived in an era when health insurers, both public and private, were mostly passive, simply reimbursing patients for a percentage of their medical spending. In contrast to today’s managed care, insurance arrangement of the 1970’s could be termed unmanaged care. Beyond that well known change there has also been a subtle change in the nature of cost sharing. At the time of the Experiment the dominant mode of cost sharing was coinsurance, which meant the patient paid something additional at the margin from using a higher priced provider or drug. The dominant mode of cost sharing today is copayment, meaning a fixed dollar price for a physician visit or a month’s supply of a drug.

The shift from coinsurance to copayment was made possible by insurers intervening directly on the supply side to create networks and formularies, whereby patients paid less to use providers in the
network or drugs on preferred tiers of the formulary. This enabled insurers to bargain directly with providers on unit price rather than rely on the indirect mechanism of patients seeking out lower cost providers with the incentive of a somewhat lower out-of-pocket payment because of coinsurance. Both the Experiment and subsequent work have showed that consumers do not actively shop for lower provider prices (Marquis 1985; Sinaiko and Rosenthal 2016). Since unit price today is set in negotiations between insurers and providers, coinsurance could be changed to copayment without materially changing the consumer’s incentives.27

From an insurer’s point of view a copayment can be set to achieve the same effect on the incentive to seek care as a coinsurance rate. From a policy analyst’s point of view, however, using the simulation model described above to estimate the effect of different levels of copayment on demand requires translating copayment rates into coinsurance rates. That translation necessarily introduces uncertainty because the denominator to estimate the coinsurance rate from the copayment is not known.

In sum, the design of the multi-armed RAND Health Insurance Experiment has stood the test of time well. The strongest evidence of that is the continued use and citation of its results. Nonetheless, changes in the financing and delivery of medical care have naturally introduced greater uncertainty about the contemporary applicability of its results.
REFERENCES


---

1 Unlike the Affordable Care Act of 2010, which had a mandate for employers with more than 50 employees, the Nixon proposal for a mandate applied to all employers.  
2 Although today it is common for health care providers to file claims on behalf of the insured for any use of services, that was not a widespread practice in the 1970's, especially for physicians. At that time many physicians would bill the patient for the full fee, with the patient then submitting a claim to the insurance company to be indemnified, which was the derivation of the term indemnity insurance.  
3 Interestingly Medicare then and now has no such limit, leading to a demand for supplementary policies.  
4 Arrow became a consultant to the Experiment and in that role wrote two papers that relaxed the assumption of no moral hazard that he had made in his classic 1963 paper (Arrow 1974; Arrow 1976).

5 One might ask what $150 would be in 2018 dollars. Depending on one’s purpose, it could be better to adjust the $150 value using a medical care price index or an all-items price index, but in practice the official medical care price indices have serious conceptual problems, especially for years before the mid-1990’s (Berndt, et al. 2000). Using the all-items Consumer Price Index, the $150 value would be around $700 today. If instead one asked how much money would be required today to buy what $150 would have bought in medical services in the mid-1970’s,
one could use the increase in personal health care spending per person as a rough indicator; it has increased by approximately a factor of 17.

6 As described below in this plan dental services were also covered at 50 percent coinsurance.

7 An experimental comparison between being insured and uninsured would have to wait three more decades, at which time an initially uninsured population was randomized between a Medicaid policy and remaining uninsured (Finkelstein, et al., 2012, Baicker, et al., 2013).

8 We could have fallen back on self reporting of use of uncovered services, but such information would not have been comparable.

9 In addition to being less generous, dental insurance was then and is now almost always provided under a separate insurance policy, but we covered dental services as part of the medical insurance policy. That meant medical and dental services both counted toward the single annual out-of-pocket spending limit. This decision was a tradeoff between realism and simplicity. Although we could have included a separate dental insurance policy, we opted for the simplicity of a single policy.

10 As a condition of enrollment the families assigned the benefits of any existing policy to the Experiment and during the Experiment we filed claims with their insurance company to recover what the insurer would otherwise have paid. This both ensured the integrity of the experimental treatment (families could not double collect) and helped defray the cost of the Experiment. If the family was uninsured at the time of enrollment, we purchased a policy for it on the individual market and had the benefits assigned to us; at the end of the Experiment the family had the option of paying the premium to continue this policy, an arrangement that meant the family could not become uninsurable during the course of the Experiment.

11 For example, if a family on the 50 percent coinsurance plan had a hospital bill of $2,000 and its annual limit on out-of-pocket spending was $1,000, under the 50 percent coinsurance plan it would have paid $1,000 out-of-pocket. If its prior plan had a deductible of $100 and full coverage above the deductible, the family would have received $900/12 or $75 per month. Any spending above $2,000 was fully covered by the Experimental plan and so was not relevant to a worst case calculation.

12 Unspent amounts in Health Savings Accounts roll over and are portable; they are thus closer in spirit to the Experiment’s Participation Incentive than Health Reimbursement Accounts, in which unspent amounts do not roll over nor are they portable. Both arrangements, however, can ease cash flow or liquidity problems in financing medical care bills, as could the Participation Incentive.

13 The results of the Experiment showed that on average 15 percent of families in the upper half of the income distribution on the 25 percent coinsurance plan exceeded the limit in a year compared with 21 percent on the 50 percent coinsurance plans (Manning, et al. 1988). For the purposes of this argument the upper half of the income distribution is more relevant, because many families in the lower half of the income distribution had a lower limit than $750, either 5, 10, or 15 percent of their income. To exceed a $1,000 limit in the 25 percent coinsurance plan, a family had to spend $4,000 on medical care; to exceed the reduced $750 limit it would have to spend $3,000. To exceed the limit in the 50 percent coinsurance plan, it had to spend $2,000. The results of the Experiment showed that families did not anticipate exceeding the limit, although once having exceeded it, they did spend at a higher rate (Keeler, et al., 1988). Because anticipation effects were unimportant, the 15 and 21 percent values just cited show that reducing the limit from $1,000 to $750 in the 25 percent coinsurance plan affected behavior in only a small number of families.

14 As explained below, the budget constraint was set in terms of the number of families, but many design decisions could be and were made as if it had been specified in dollars. For example, in the 70-30 split here, the exact dollar value of the budget constraint would have affected the total number of families enrolled, but not the 70-30 split.

15 There was, in effect, a soft budget constraint that included claims expenses. As a result, an unexpectedly large claim expenses could reduce the research budget. Although it made little sense from a social welfare point of view for the federal government to purchase a reinsurance policy, we determined after a search that no insurer would sell the Experiment such a policy. Because of the difficulty of adjusting the research budget, other budget adjustments were made to accommodate higher than anticipated expenses, for example the reduction of the annual Participation Incentives in the 25 percent coinsurance rate plans to $750.

16 Although a dollar budget constraint for the lifetime of the Experiment was not specified at this stage – one could not know what annual medical spending trends would be – we could estimate the initial marginal cost of an
additional enrolled family and the marginal cost of another site. We estimated the between- and within-site variances from survey data.

17 The six sites were Dayton, OH, Seattle WA, Charleston SC, Fitchburg MA, Georgetown County SC, and Franklin County MA. The site where physicians did not make appointments was Georgetown County.

18 It turned out that the induced demand from the baseline screening exam was negligible.

19 The mail questionnaire did not stimulate use.

20 We only audited physician bills because we expected that we would receive all hospital bills both because of their size and the amount of money at stake for the hospital.

21 An additional piece of evidence that underfiling did not impair the qualitative conclusion that coinsurance affected use was that the major effect of cost sharing was between free care and 25 percent coinsurance, and in both plans there was a large financial incentive to file a claim.

22 For example, the widely used SF-36 and various other “SF” measures of health status are shortened versions of the Experiment’s General Health Index. SF is an acronym for Short Form.

23 The refusal rate for the HMO arm in Seattle, the one site with an HMO, was only 2 percentage points higher than the refusal rate for all the cost sharing arms in Seattle.

24 A third HMO was also willing to participate, although that would have required opening a new site.

25 This leads most participants in Traditional Medicare to obtain supplementary plans that cover much of the costs sharing or to opt for a Medicare Advantage plan.

26 There were two positive other outcome effects in the free plan, a very slight improvement in corrected vision from persons getting eyeglasses or updating their prescriptions and a considerable increase in dental fillings.

27 Copayment is more attractive for consumers than coinsurance because price is a known amount rather than a function of an uncertain bill.