Chapter 8

Evaluation Issues

Contents

Box

Chapter 8
Evaluation Issues

INTRODUCTION

Different observers can, and do, see Oregon's demonstration proposal in very different research contexts. The most obvious context is as a straightforward health services research experiment. Indeed, Oregon's justification for requesting Federal funding for its proposed new Medicaid demonstration program is that the program would provide useful information to the Federal Government. This knowledge would presumably be used to improve other State Medicaid programs and inform Federal health policy decisionmaking.

Because different States operate with very different Medicaid systems, the usefulness of Oregon's proposal in this context depends at least in part on the ability to dissociate the different components of the demonstration and assess their separate effects. For example, States may wish to implement Oregon's prioritized list as a Medicaid benefit package without necessarily implementing the other components that Oregon proposes to demonstrate (e.g., eligibility expansion, managed care implementation). This chapter discusses some of the basic issues likely to arise in evaluating the demonstration on this level.

In addition, however, Oregon's proposal is seen as a potential experiment of two very different questions. First, the proposal can be viewed as a simple experiment designed to answer the question: Is it possible, using the combination of mechanisms Oregon would implement, to provide acceptable health care coverage to the uninsured poor population without significantly raising costs to the taxpayer and to the health care system? A second question is even further from the traditional bounds of health services research: Is health care coverage based on prioritization of health care services, with public input, politically sustainable? These two questions are addressed briefly in the final section of this chapter.

OREGON'S PROPOSAL AS A HEALTH SERVICES RESEARCH PROJECT

Conceptually, the proposed demonstration is an experiment in which two separate populations (the uninsured poor and the Medicaid-eligible population) undergo a number of different, simultaneously administered interventions.

For the uninsured poor population, these interventions are relatively simple: they consist of a package of covered services and a new delivery system (i.e., managed rather than ad *hoc* charity care). The theoretical questions to be answered for this target group are:

1. Does the existence of health insurance coverage (specifically, coverage for services in condition-treatment (CT) pairs 1 through 587), delivered through a managed care system (as Oregon has designed it), increase health access to the uninsured poor? Does it improve health status and satisfaction with care?
2. If it does, at what cost (or savings) to the State, providers, employers, the new beneficiaries, and other groups of interest?

For the population currently eligible for Medicaid, the hypotheses being tested are more complex. This population would undergo a number of changes, including changes in benefits, eligibility, and source and type of care. Although the outcomes of interest still revolve around health care access and cost, the questions are more specific and more complex because they involve comparisons with an existing program. They would include, for example:

1. Does simplifying eligibility rules increase program participation? Who gains and who loses—and how much—through changes in income calculation, elimination of retroactive coverage, and change in the minimum length of eligibility?

2. Do the changes in benefits lead to overall changes in access to services, health status, and satisfaction with care? Do they affect different subgroups of the population differently (i.e., are there ''winners' and 'losers' ')? Do they affect program costs?
3. Does the expansion to statewide managed care affect health access and satisfaction, and is the effect uniform across the population? Does it affect program costs?

Because provider participation is (presumably) critical to health care access, the third set of questions encompasses others: for example, does changing the method of payment affect participation?

From the Federal perspective, it would be important to consider the different components of the experiment separately in the evaluation. Although it is certainly possible that other States (or the Federal Government) would want to duplicate the entire package, it is much more likely that they would choose to adopt only a few components. For instance, another State might consider implementing the prioritized list and simplified eligibility rules for only the existing Medicaid population. To gain information from the Oregon experiment that would be useful to a State entertaining such an option, the two populations affected and the various interventions applied would all need to be evaluated separately, and the outcomes measured would need to be appropriately linked with the intervention(s) that caused them.

Identifying causal effects-i. e., the link between intervention and outcome and the direction of that link-is the crux of any type of applied research. Determining that the intervention being studied caused a particular outcome is especially difficult in social science research, where the intervention is often hard to apply reliably and many environmental factors that may affect the outcome are out of the control of the researchers.

The ability to draw conclusions about cause is enhanced by incorporating evaluation considerations into the design of an experiment and specifying clearly the hypotheses and outcomes of interest before the experiment takes place (129). Oregon's waiver application makes clear that it considers evaluation of the demonstration to be the responsibility of the Health Care Financing Administration (HCFA), not the State. However, it does present a starting evaluation plan, including some possible hypotheses to be tested, data sources, and some suggested methods of analysis using these data sources.

Even with impeccable theory and planning, however, determining causal connections in Oregon's proposed demonstration, as with any research project, might be difficult. Campbell and Stanley (1963) and Cook and Campbell (1979) have described a framework for identifying the research problems that make drawing conclusions about the effects of an intervention difficult (box 8-A). Three problems are especially relevant to the proposed demonstration and deserve mention here.

Selection of Adequate Controls

To help rule out threats to statistical validity (see box 8-A), experiments often randomize the test population to intervention and nonintervention ("control") groups. Where randomization is not attempted, as in Oregon's proposed demonstration program, the control population may be historical (i.e., the test population before the intervention was applied) or matched (e.g., another State's Medicaid population). Oregon's outlined evaluation plan suggests that both types of controls be used.

Some historical (predemonstration) utilization data exist for hospital inpatient services and for other services provided outside of the existing managed care area. Also, new and existing program participants could be surveyed regarding their health status and satisfaction at the onset of the demonstration.

Both types of historical baseline data are useful, but both also have strong limitations. For example, few data on utilization of capitated services (including physician, laboratory, and x-ray services) exist for beneficiaries enrolled in the current prepaid program. Prepaid plan enrollment is presently mandatory for all persons eligible for Medicaid through Aid to Families with Dependent Children (AFDC) who live in a nine-county area that encompasses most of Oregon's urban areas (and approximately 54 percent of Oregon's AFDC beneficiaries (see ch. 4). Thus, many of the utilization comparisons possible under the demonstration would be restricted either to certain areas (e.g., fee-for-service (FFS) counties), specific services (e.g., hospital inpatient services), or groups of beneficiaries not currently enrolled in prepaid plans (e.g., poverty level medical women and children).

Box 8-A—Attributing Causality in Program Evaluation

Most research has underlying it one basic goal: to test whether the intervention (e.g., a new drug, a new school curriculum, a change in Medicaid rules) causes one or more outcomes. In laboratory and some clinical research, the outcome desired can be clearly specified and measured, and outside influences that might affect that outcome can be rigidly controlled for. In these cases, the researcher's control over external factors raises the likelihood that the researcher can conclude with confidence that the outcome (if it occurs) was caused by the intervention. In other kinds of research, however, including most social science research, the researcher has much less control over the outside factors that might act upon the population of interest. In such cases, the conclusion that the intervention caused a given outcome is strengthened by eliminating various "threats" to its validity.

Threats to validity can be separated into four categories:

- Statistical conclusion validity--Are the intervention and the outcome related on the basis of statistical evidence? For example, does the study have enough statistical power (e.g., a large enough sample size) to detect an effect of the intervention? Are the outcome measures reliable (e.g., if the outcome is a score on a test, is the test itself statistically reliable)? Is the intervention applied uniformly across the population, and if not, can the population heterogeneity be itself measured and analyzed?
- *Internal validity--Given* that an intervention and outcome are statistically linked, how plausible is it that the intervention (and not some outside factor) actually caused the outcome? Threats to internal validity include biased selection (e.g., a difference between test and nontest populations was detected because the test population was predisposed to that difference); diffusion or imitation of the intervention into the control (nontest) group; and ambiguity about the direction of causality (did A cause B, or did B cause A).
- *Construct validity--Do the* measurements representing the intervention and the outcome really stand for the "constructs" they are intended to, or might they accommodate other concepts as well? For example, if a person improves after being given a pill by a physician, is it the the pill's therapeutic effect being measured--or is it some combination of the pill's chemical effect, the physician's helpful concern, and the patient's belief that the pill will be effective? (Such concerns led to the widespread use of "placebo" controls in drug research.) Having several different measures (e.g., length and number of physician visits, waiting time to visits) to represent the "construct" (e.g., access to health care) can reduce threats to construct validity. If the intervention being tested includes many components, which must be separately measured, threats to construct validity maybe more difficult to rule out.
- *External validity--Can the* results of the experiment be inferred to apply outside of the test population? If the setting and the intervention interact, for example (e.g., instilling discipline in boot camp), the intervention may not have the same effect in another setting (e.g., a preschool). Similarly, if the population selected for the experiment differs substantially from the nonexperimental population, the experimental conclusions may not be valid when applied to the broader population.

SOURCES: D. Campbell and J. Stanley, Experimental and *Quasi-Experimental Designs for Research (Chicago, IL: Rand* McNally, 1966) and T. Cook and D. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Boston, MA: Houghton Mifflin, co., 1979).

These historical utilization data, even where available, would apply only to existing Medicaid beneficiaries. Surveys of incoming program participants would be the only mechanism by which to estimate baseline utilization and health status of newly eligible persons. However, such surveys would be expensive to conduct and would have to be implemented very rapidly if the waiver is approved, limiting the sample size of the data and raising the chances that the survey would not be adequately tested before being applied. Also, a survey at the time of enrollment might overestimate the health problems of this population, since many individuals might postpone seeking care if they know they will soon have coverage and would not have to pay out-of-pocket.

Using comparison groups outside the demonstration population as the controls eliminates some problems inherent in the historical controls (e.g., sample size), but this strategy also has limitations. Using data from other State Medicaid programs, for example, introduces confounding factors due to differences in State- and program-specific characteristics (e.g., coverage limitations, general availability of health resources). Similarly, using as the population control another group within Oregon (e.g., persons eligible for the program who did not enroll) introduces confounding factors related to the charac-

teristics of that population and the lack of a systematic method for obtaining utilization and other relevant data from individuals within it.

Statistical Power To Detect Effects

Even when an effect occurs, a test population may not always be large enough to detect it within the traditional limits of statistical confidence. Small predicted effects require large sample sizes to detect their occurrence. This problem would place limits on some of the outcomes that an evaluation of a demonstration such as Oregon's could expect to identify. Changes in population mortality that might result from changes in covered services, for example, are unlikely to be detectable in a population of a few hundred thousand persons over a 5-year period. Some more specific health outcomes that one might wish to detect are also unlikely to surface; for example, the measurable benefits of many preventive services are not apparent for many years after the service is used.

Low power to detect effects is especially likely to limit the ability of evaluators to determine that specific intervention components caused particular outcomes (e.g., that implementing the prioritized list reduced costs). Separating the effect of the new benefit package from the effect of prepaid managed care, for example, requires either detailed data from the prepaid sector before the new benefits take place or comparative data during the demonstration between prepaid and *FFS* managed care. In both cases, data would be limited. As noted above, only a few baseline utilization data are available for current prepaid plan enrollees. Although the State has recently begun requiring such data from prepaid plans, there would be less than 1 year's worth if the demonstration were to begin in mid-1992. Furthermore, data currently collected from prepaid plans reflect only very broad categories of service (e.g., physician visits) and would thus be of limited usefulness in linking outcomes to the condition-specific coverage exclusions of the prioritized list (see ch. 4). In addition, the populations receiving prepaid and FFS care during the demonstration would differ by virtue of location (the latter would be mostly rural populations), and again population-specific factors may confound interpretation of the data.[1]

Monitoring or surveying particular subpopulations likely to lose or gain from the change in benefits (e.g., those with chronic conditions below the line; those with terminal conditions who might use hospice care; adults newly eligible for preventive care) does offer one opportunity to evaluate directly the effect of the prioritized list. In many of these cases, the size of the expected effect on the specific population is large enough to be detectable. Choosing appropriate subpopulations to study in depth would thus be an important component of an evaluation plan.

Difficulty Ensuring That the Intervention Is Applied Consistently

The list itself gives no specific guidance regarding how to assign patients to CT pairs, so no two providers are likely to apply the list in the same way to their patients. Differences in how the list is applied would probably be the greatest between FFS and prepaid care providers. Even between two providers under the same payment system, however, the ambiguity of the list is likely to lead to greatly different interpretations of what is covered and what is not. The addition of mental health and chemical dependency services to the prioritized list could further confound this problem.

Some of this ambiguity could be resolved over time through greater provider education and instructions, but it is not clear that these instructions could be sufficiently developed by the time the program begins (assuming a startup date of July 1992) (see ch. 4).[2] And even with clearer instructions for using the list, providers might violate those instructions in their own interests or the interests of their patients (see chs. 3, 4, and 5). The State may be unable to prevent this from happening, or even to detect that it occurs.

[1] At least some of any differences found are likely to be caused by factors such as geographic barriers to access, rural provider shortages, and differences in population characteristics and health care preferences, rather than solely by differences in FFS vs. prepaid care (U.S. Congress, OTA, September 1990). Since the detailed effects of such population-specific and geographic differences are not generally well-described quantitatively, they cannot be easily adjusted for in a statistical analysis.

[2] Note that the original July 1992 startup date has been postponed on a month-to-month basis pending HCFA approval of the waiver (see ch. 4).

THE PROPOSAL AS A BROADER POLITICAL EXPERIMENT

In contrast to the traditional health services research demonstration (as outlined in the waiver proposal), Oregon's plan can also be seen as a chance to test the question of whether a novel idea to cover the uninsured poor can work without substantially increasing costs. Indeed, many people who are skeptical of some of the specifics of the proposed program nonetheless view it as a chance to test a novel health care reform strategy. In this context, the Oregon demonstration would really be a test of a comprehensive package of interventions, in which separating out the effects of various components is unnecessary. The ''research' question in this case is simply: Can the plan successfully extend coverage to uninsured people without substantially raising long-term program and social costs?

Evaluating this question in the aggregate would not require nearly as detailed a level of data analysis as would evaluating the separate effects of the various components of the proposed program. The crucial parameters to measure would be the level of access to care (for which the level of benefits might even be accepted as a proxy) and the difference between actual demonstration program costs, projected Medicaid program costs if the poor uninsured population were not covered, and perhaps estimated costs of some alternative way of providing coverage to uninsured persons. The danger of such an approach is that as an experiment, its results could only be appropriately extrapolated in the aggregate. Other States could apply the results only if they, too, were willing to implement the total package that Oregon has proposed.

Finally, Oregon's proposal presents a larger political feasibility experiment: Can the State keep the structure and dynamic of the program intact? If, for example, program costs were higher than expected, would the legislature actually be willing and able to reduce benefits or increase revenues to fund it? Or would the plan evolve over time into simply another version of the current system, in which neither eliminating specific treatments nor raising taxes is politically feasible, and the State must resort once again to limiting eligibility and provider payment?

In fact, some Oregonians have speculated that the program's design, in which funding can in theory affect only the level of benefits, may actually serve to increase the public's willingness to fund Medicaid by highlighting the treatments that would be cut if funds were unavailable. Thus, the demonstration may be of political interest to some policymakers despite its potential drawbacks as a health services research project.