Considering Whether Medicaid Is Worth the Cost: Revisiting the Oregon Health Study

The Oregon Health Study was a groundbreaking experiment in which uninsured participants were randomized to either apply for Medicaid or stay with their current care.

The study showed that Medicaid produced numerous important socioeconomic and health benefits but had no statistically significant impact on hypertension, hypercholesterolemia, or diabetes. Medicaid opponents interpreted the findings to mean that Medicaid is not a worthwhile investment. Medicaid proponents viewed the experiment as statistically underpowered and, irrespective of the laboratory values, suggestive that Medicaid is a good investment.

We tested these competing claims and, using a sensitive joint test and statistical power analysis, confirmed that the Oregon Health Study did not improve laboratory values. However, we also found that Medicaid is a good value, with a cost of just $62 000 per quality-adjusted life-years gained. (*Am J Public Health*. Published online ahead of print March 19, 2015: e1–e5. doi: 10.2105/AJPH.2014.302485)

Peter A. Muennig, MD, MPH, Ryan Quan, BA, Codruta Chiuzan, PhD, and Sherry Glied, PhD

MARK TWAIN IS BELIEVED TO have penned the saying "There are lies, damn lies, and statistics."[1] Even when working with gold standard data—a well-conducted randomized trial of a social policy—there is both an art and a science to the analysis of the data at one's fingertips. Although social science experiments are the most rigorous means of evaluating a social policy, they tend to be logistically messy, requiring post hoc analytic adaptations. It is often the case that less than perfect policy experiments come under fire in the media, particularly when the findings do not align with a particular group's beliefs. One recent example of this mix of science and media politics can be found in the case of the Oregon Health Study (OHS).

In 2008, the state of Oregon randomly provided Medicaid coverage to approximately 10 000 individuals randomly selected from 30 000 names drawn from the 90 000 who were eligible for Medicaid.[2] An expert interdisciplinary research team collected comprehensive survey responses, administrative information, and biomarker data on this subsample of winners and losers of the lottery. They found that Medicaid provided substantive financial protections, increased rates of preventive testing, reduced depression, and improved self-rated health.[2] They also found that those randomized to receive Medicaid did not achieve a statistically significant reduction in blood pressure, serum cholesterol levels, or blood glucose levels.

However, relatively few of those who won the right to enroll in Medicaid actually did, and those who did turned out to be quite healthy to begin with. This greatly reduced the effective sample size of the "treated" group—those participants who were assigned to receive Medicaid. Further complicating matters, Oregon rapidly accelerated enrollment when more funds became available. This greatly shortened the time that the researchers had for data collection.[2] These complications led to a less than perfect experiment.

As a result of these issues, some researchers pointed out that the study was statistically "underpowered,"[3] meaning that the number of participants should have been larger. These researchers based their claims on post hoc statistical analyses of each individual laboratory measure. For example, they showed that there were only 80 participants who might qualify as diabetic, and many more would have been needed to detect a meaningful reduction in diabetes.[4]

Proponents of Medicaid point to these flaws to suggest that conclusions cannot be drawn about the effectiveness of Medicaid in improving these laboratory measures of health. They further note that, even if one sets the laboratory results aside, the other benefits are important, meaningful, and worth the investment.[5] Various opponents of Medicaid, conversely, tended to focus on the null results in the laboratory tests and declared that the study "proved" that Medicaid is a poor policy investment.[6–8] Supporting these claims, the OHS authors objectively note that the joint effect of all the tests combined was also not statistically significant.[2] This suggests that even if one considers the impacts of all the tests together, the OHS still fails to show statistically robust improvements in laboratory measures of health.

The arguments of both Medicaid proponents and opponents are plausible. They rest on concerns about whether (1) the improvements in laboratory values would have been statistically significant had the OHS sample been larger and (2) the nonlaboratory benefits that were realized are meaningful enough to justify further expansion of Medicaid. We have addressed the first concern by conducting a highly sensitive joint test coupled with a post hoc power analysis of this test. If this test, which is more sensitive than is the one the authors originally used, is powerful enough to detect combined differences in laboratory values, it should validate or refute the critique that the sample size was too small. To address the second concern, we performed a cost-effectiveness analysis and a cost–benefit analysis to test the concern that Medicaid is not worth the investment.

THE PROBLEM WITH SOCIAL SCIENCE EXPERIMENTS

One aspect of the art of social science experiments is deciding the extent to which one might err on the side of a type I error or a type II error[9]—or, more simply, deciding whether one wishes to use an approach that moves the study more toward acceptance of

a false positive or a false negative finding.

Unlike drug trials or biological experiments in which experimental conditions can be more tightly controlled, social experiments have many moving pieces. In a drug trial, for instance, participants can be closely monitored to ensure that they take their medication (be it the placebo or the control). However, the most careful and experienced research teams cannot control a person's choice of whether to actually enroll in a government program if they win the right to do so. They also cannot control the decisions of policymakers, who are much more motivated to act according to politics than science when a policy is implemented. Finally, there also tend to be challenges associated with multiple and overlapping outcome measures, participants who switch between treatment and control arms of the study, and a requirement for long follow-up times to detect important health outcomes, such as mortality.[10]

With respect to the issue of overlapping outcome measures, consider Medicaid, which may both improve health and provide protection against catastrophic medical costs.[2,11] These outcome measures overlap; whereas poor health can impoverish a family, higher income is thought to make poor people healthier,[12] and it is virtually impossible to disentangle these 2 effects.[12]

With the prospect of a pile of messy experimental data that will displease reviewers and readers alike, social scientists tend to take the most tidy, rigorous analytical path. Without these protections, not only is there uncertainty about positive outcomes, but there also exists the potential for the researchers to change model specifications or remove participants with missing or otherwise undesirable values (e.g., outliers) until statistical significance is achieved. As a result, researchers tend to prespecify their statistical approach and to employ intent to treat analysis.

In intent to treat, all participants in the experimental (treatment) group are regarded as treated, whether they actually received the treatment, withdrew from the study, or deviated from the protocol. This generally increases the number of participants that are needed to observe a result and reduces the measured impact (effect size) of the study outcome measures. For instance, in the OHS, only half of the participants who won the right to enroll in Medicaid actually did, but all of them were included as treated. So, if Medicaid actually reduced the chance of depression among treated participants by 30%, the measured benefit would only be half of that, or 15%. On the upside, intent to treat all but eliminates problems associated with selection bias that might produce a false positive, thereby reducing the chances of a false positive result. Moreover, if the sample size is large enough, it is possible to correct the effect size (e.g., effectively multiplying the 15% rate of depression by 2).

Of course, any attempt to avoid a false positive result also increases one's chances of a false negative result. It is notoriously difficult to estimate how many participants will actually choose to be treated, stay in their experimental group, receive alternative forms of community insurance, and so forth. When one applies more rigor to crunching the numbers, more participants are needed. In the worst case, a program that was actually socially beneficial could be found to produce no statistically significant effects at all. This could discourage policymaking and might even discourage future research in the topic at hand—few researchers wish to spend their time and effort evaluating a program that is unlikely to produce results.

So, therein lies the problem. Rigor can be used to reduce one's chances of a false positive, but rigor can also increase one's chances of producing false negative results. A false result is objectively undesirable whether it is a false positive or a false negative. The art of social experiments lies in how the lines are drawn and how the results are presented. In the case of the OHS, the authors chose to generally err on the side of rigor. They presented the OHS as having an impact on important outcomes, such as financial protections, depression, and preventive screening, but not on laboratory values.

It is the latter claims that created a political storm because laboratory values were the primary objective measure of physical health in the study. This raises the key question: if Medicaid does not improve physical health, why are we spending hundreds of billions of dollars a year on this program?

RESPONDING TO THE QUESTIONS THE OREGON HEALTH STUDY RAISES

Although little can be done to circumvent problems that arose during the OHS, it is possible to address the concerns raised by proponents and opponents of Medicaid in this case. For instance, post hoc power analyses can inform the consumer of whether the initial sample size estimate was large enough to ensure sufficient power to draw sound conclusions. It is also possible to conduct more sensitive joint tests across all outcome measures rather than perform separate tests for each individual outcome. This will tell us whether the overall effect of Medicaid on laboratory measures was meaningful in any way across measures. If a sensitive joint test across all laboratory markers proves to be adequate to detect changes in laboratory values but fails to produce statistically significant results, we can be more confident of the results.

Finally, it is important to address the ultimate question that arises from the failure of the OHS to produce improvements in laboratory measures: is Medicaid worth the investment of hundreds of billions of dollars? Economic analyses allow us to answer this question.[13] In the case of the OHS, the outcome measures that grab the headlines—such as the effect of Medicaid on cholesterol or blood pressure levels—are by no means the only or even the best measures of Medicaid's value to society. The presence of other benefits does not mean, however, that it makes sense for the government to pay for them.

To answer that question, a much more comprehensive look at Medicaid is needed. Such an analysis would include estimates of its joint benefits and an estimate of how much it costs to realize these benefits. This type of analysis addresses concerns from Medicaid opponents that the program is not worth taxpayers' investment.

OUR APPROACHES

We revisited the findings of the OHS using a post hoc power analysis, a uniquely sensitive test of joint effects (seemingly unrelated regression),[14] and a basic cost-effectiveness analysis.

Power Analysis

In the OHS, most of the outcomes of interest appeared to be statistically trending in a way that favored the treatment arm, and it has been argued that the statistical power of the experiment was too low to detect meaningful effects.[15] In conjunction with 2 biostatisticians at Columbia University, we used Columbia University's high-performance UNIX computers to determine the power associated with the Hotelling 2-sample T^2 test, a multivariable version of the *t* test, across all the biomarkers in our analysis.[16] For the power calculation, we used the original study data to input the vector of mean differences (treated vs nontreated) associated with each biomarker and the pooled variance–covariance matrix for the 2 groups. We performed these calculations in Power Analysis and Sample Size software (NCSS, Kaysville, UT). Additional information is available as a supplement to the online version of this article at http://www.ajph.org.

Seemingly Unrelated Regression Analyses

We next set out to perform a more sensitive joint test than was employed in the original study. Seemingly unrelated regression uses a single variance–covariance matrix across multiple seemingly unrelated regression equations.[14] In seemingly unrelated regression, multiple equations that appear unrelated (e.g., a model with cholesterol as a dependent variable and another with hypertension as a dependent variable) are actually related via a correlation in their error terms. To be specific, the errors of the linear equations are correlated across equations for a particular individual but are uncorrelated across participants. By estimating multiple equations together, we greatly improve the statistical efficiency of the estimators over fitting the models separately. Seemingly unrelated regression has been used to detect joint effects across similarly modest effect sizes for health outcomes in much smaller studies than the OHS.[17,18]

Cost-Effectiveness Analysis

Our objective in estimating the cost-effectiveness of Medicaid was to simply ask if Medicaid is worth its cost. Recall that Medicaid was associated with financial protections, higher rates of diagnosis and treatment of diseases, a higher chance that a recipient will receive preventive care (such as a colonoscopy or breast cancer screening), and lower rates of clinical depression. We included only the value of the observed, statistically significant reduction in clinical depression relative to the full cost of Medicaid to obtain a lower-bound estimate of the overall cost-effectiveness of Medicaid. Because none of the observed outcomes were harmful and because all costs were included, one can be much more certain that the combined effects of all benefits of Medicaid are cost-effective.

The value of the financial protections afforded by health insurance to families that are faced with paying for catastrophic medical coverage is certainly large, but we did not attempt to measure it. This is in part because it may present a challenge to the validity of the overall cost-effectiveness ratio. Specifically, it is not clear to what extent financial protections reduced depression, and it is not clear to what extent reduced depression improves financial protections. Including this benefit might lead to double counting of important benefits.

The expert US Task Force on Clinical Preventive Services estimates that preventive measures such as mammography do save lives,[19] but there is no experimental evidence that proves this. Likewise, higher rates of screening and treating for disease should be lifesaving,[20] but again, this has not been proven, and so these benefits are not included. By excluding these potential benefits while including all their costs, we can be more certain that the cost-effectiveness ratio is a conservative estimate of overall benefits. We also did not count nonstatistically significant benefits to blood pressure, cholesterol, or glycemic index. We adopted the societal perspective, and we have presented the costs in 2013 dollars, which we adjusted using the Consumer Price Index.[21]

The health-related quality of life measure used in the OHS, the SF-8, cannot be used to calculate quality-adjusted life-years (QALYs), which is the primary effectiveness outcome in cost-effectiveness analyses.[22] We therefore turned to the literature to obtain the EQ-5D score change associated remission of moderate depression.[4] Finally, to conduct a cost–benefit analysis, we simply monetized QALYs using 1 commonly used value ($100 000/QALY)[23] adjusted to 2013 dollars ($130 000) using the Consumer Price Index.[21]

When considering costs, it is important to be mindful that Medicaid is paid for by taking money from taxpayers and giving it to those who qualify for Medicaid. Costs are only meaningful to the extent that these funds change medical utilization or produce administrative costs. After adjustment for inflation, the marginal increase in medical costs was $1213.[2] To this, we added the 7.00% administrative cost associated with running Medicaid, for a cost of $1298.[24] We divided these costs by the product of EQ-5D score change (0.23)[25] and the difference in positive depression screens (9.15%).[2] The assumptions of this analysis are listed in Table 1.

OUR FINDINGS

There were 6315 treated participants and 5769 control participants in the final study after removing those with missing values. This produced a statistical power of 69% to detect an effect size of 0.06 equal to the differences of the group means of the 6 response variables under study adjusted by the variance–covariance matrix. Roughly 7500 participants would have been needed in each group to achieve a statistical power of 0.8 at this α. All calculations considered a type I error of 0.05.

The joint test indicated that the laboratory tests for all biomarkers were not significantly different between the 2 groups ($P=.7$). This was true whether the analysis was limited to those older than 50 years ($P=.13$) or those with preexisting conditions ($P=.56$).

Under the assumptions that the provision of preventive medical care saves no lives and treating clinical depression saves no lives, the incremental cost-effectiveness ratio of providing Medicaid in terms of depression alone was $62 000 per QALY gained. Using the common valuation of $130 000 per QALY,[23] this suggests a net savings of at least $68 000. Revealed preference and contingent valuation approaches (alternative, theory-grounded methods for valuing QALYs) put the savings significantly higher—as much as a half a million dollars.[26] The results are summarized in Table 2.

TABLE 1—Assumptions Used in the Cost-Effectiveness Analysis

Assumption	Explanation	Influence
All costs should be included, but only reductions in depression should be included as benefits.	By only including the most certain benefits, we reduce the chance of double counting benefits or including benefits that were not experimentally tested.	This ensures that the ICER value represents a maximum value.
One year of Medicaid investment produces 1 year of benefits.	It is likely that depression rates would return to baseline rates were Medicaid withdrawn.	Because benefits other than depression treatment may be longer lasting (e.g., colonoscopy screening), this assumption also ensures that the ICER represents a maximum value.
Secondary data sources accurately reflect measured reductions in depression in this experiment.	It is necessary to use a specialized instrument to calculate quality-adjusted life-years, and this instrument was not included in the original experiment.	Unknown.

Note. ICER = incremental cost-effectiveness value.

CONCLUSIONS

The OHS produced clear benefits for the recipients with respect to financial protection (what many would argue health insurance is meant for) and reduced depression, increased diagnosis and treatment of diabetes, and increased preventive medical care.[2] These overall impacts were larger than were those found in an older randomized trial of private health insurance.[11]

At $62 000 per QALY gained (or a net return of at least $68 000 per enrollee to society as a whole), Medicaid seems worth the investment—even when we ignore the significant benefits associated with increased prevention, increased diagnosis and treatment of disease, and financial protections for the family. This incremental cost-effectiveness ratio is well below the cost of most medical and social policy investments that we as a society choose to make, such as treating disabling spinal problems[27] or placing smoke detectors in homes.[28] These findings are remarkably similar to earlier assessments of the cost-effectiveness of private health insurance and Medicare that were made using Oaxaca decompositions of prospective survey data entered into decision analysis models.[29,30]

Whether it produced meaningful physical health benefits is still an open question. Our power estimates show that Medicaid had little impact on laboratory tests, even when the benefits of all the laboratory tests are considered together. This in turn raises questions that go far beyond the benefits of Medicaid. Conventional wisdom holds that clinical services (e.g., cholesterol screening) improve the clinical laboratory biomarkers tested in this experiment (e.g., cholesterol).[31,32] Although the OHS was clearly inadequately powered to detect a statistically significant difference in any single test, a lack of benefit across all these tests together raises some questions about the real-world validity of some of these recommendations.

This may be because recommendations tend to be partly derived from clinical studies conducted under tightly controlled conditions. In the real world, those who are prescribed drugs may not take them, or they may take them in combination with a wide array of other medications. Participants in clinical trials might also be very different from Medicaid recipients with respect to behavioral risk factors and the use of other substances.

However, the additional information about the OHS that we gleaned from extensive qualitative analyses of participants years after the study was completed does suggest that some clinical benefits were realized.[33] Many interviewees felt that their health had improved substantially (in line with the finding that self-rated health improved by 25% in the original study). Moreover, many of these interviewees felt that their benefits came years, rather than months, after enrollment. The interviews suggest that the limited follow-up of the study may have produced an undercounting of benefits. The potential psychological impact of being eligible for

TABLE 2—Summary of Our Findings

Outcome	Finding
Joint test	The joint analysis implemented through seemingly unrelated regression models overall indicated that the laboratory tests were not significantly different between the "treated" (Medicaid participants) and the "untreated" groups.
Power analysis	Our joint test was able to detect differences in the combined impact of Medicaid on laboratory tests at 70% power. Therefore, we concluded that treated participants did not realize improvements in laboratory values.
Cost-effectiveness analysis	When considering all health and economic benefits, Medicaid comes at a better value than do many other social policy and health investments ($62 000 per quality-adjusted life-year gained).

Medicaid on well-being should also not be discounted.

We find that Medicaid might not do much for high cholesterol, blood pressure, or blood sugar, but it is still a very good investment. ■

About the Authors

Peter A. Muennig and Ryan Quan are with the Department of Health Policy and Management, Mailman School of Public Health, Columbia University, New York, NY. Codruta Chiuzan is with the Department of Biostatistics, Mailman School of Public Health, Columbia University. Sherry Glied is with the Wagner School, New York University, New York.

Correspondence should be sent to Peter Muennig, MD, MPH, Associate Professor, Mailman School of Public Health, Columbia University, MSPH Box 14, 600 West 168th Street, 6th Floor, New York, NY 10032 (e-mail: pm124@columbia.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted November 13, 2014.

Contributors

P. A. Muennig conceptualized the analysis, guided the statistical modeling, and wrote the article. R. Quan and C. Chiuzan made significant contributions to the conceptualization of the statistical analyses, conducted the statistical analyses, and contributed to article development. S. Glied made significant contributions to the conceptualization of the analysis and article development.

Acknowledgments

We would like to thank Heidi Allen and Kate Baicker for their help in framing the study and providing data and for making significant edits to the article.

References

1. Twain M. *My Autobiography: "Chapters" From the North American Review.* New York, NY: Dover; 1999.

2. Baicker K, Taubman SL, Allen HL, et al. The Oregon Experiment—effects of Medicaid on clinical outcomes. *N Engl J Med.* 2013;368(18):1713–1722.

3. The Incidental Economist. The Oregon Medicaid Study and the Framingham Risk Score. Available at: http://theincidentaleconomist.com/wordpress/the-oregon-medicaid-study-and-the-framingham-risk-score. Accessed August 4, 2014.

4. The Incidental Economist. Power calculations for the Oregon Medicaid Study. 2013. Available at: http://theincidentaleconomist.com/wordpress/power-calculations-for-the-oregon-medicaid-study. Accessed November 11, 2014.

5. Drum K. No, the Oregon Medicaid study did not show "no effect." *Mother Jones.* May 11, 2013.

6. Pickert K. 5 things the Oregon Medicaid Study tells us about American health care. *Time.* May 2, 2013.

7. Conover C. Does the Oregon Health Study show that people are better off with only catastrophic coverage? *Forbes.* May 7, 2013.

8. Sargent G. A war over Medicaid. *Washington Post.* May 2, 2013. Available at: http://www.washingtonpost.com/blogs/plum-line/wp/2013/05/02/a-war-over-medicaid. Accessed March 11, 2015.

9. Gujarati DN, Porter DC. *Essentials of Econometrics.* 3rd ed. New York, NY: McGraw-Hill Irwin; 1999.

10. Greenburg D, Shroder M. *The Digest of Social Experiments.* 3rd ed. Washington, DC: Urban Institute Press; 2004.

11. Newhouse JP. *Rand Corporation. Insurance Experiment Group. Free for All?: Lessons From the Rand Health Insurance Experiment.* Cambridge, MA: Harvard University Press; 1993.

12. Muennig P. Health selection vs. causation in the income gradient: what can we learn from graphical trends? *J Health Care Poor Underserved.* 2008; 19(2):574–579.

13. Muennig P. *Cost-Effectiveness Analysis in Health, a Practical Approach.* San Francisco, CA: Jossey-Bass; 2007.

14. Zellner A. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J Am Stat Assoc.* 1962;57(298): 348–368.

15. Carroll A, Frakt A. Oregon Medicaid: power problems are important. 2013. Available at: http://theincidentaleconomist.com/wordpress/oregon-medicaid-power-problems-are-important. Accessed May 25, 2014.

16. Kariya T. A robustness property of Hotelling's T2-test. *Ann Stat.* 1981;9(1): 211–214.

17. Muennig P, Schweinhart L, Montie J, Neidell M. Effects of a prekindergarten educational intervention on adult health: 37-year follow-up results of a randomized controlled trial. *Am J Public Health.* 2009;99(8):1431–1437.

18. Muennig P, Robertson D, Johnson G, Campbell F, Pungello EP, Neidell M. The effect of an early education program on adult health: the Carolina Abecedarian Project randomized controlled trial. *Am J Public Health.* 2011;101(3):512–516.

19. Agency for Health Research and Quality. United States Preventive Services Task Force. Available at: http://www.ahrq.gov/clinic/uspstfix.htm. Accessed September 19, 2010.

20. Cutler DM, Rosen AB, Vijan S. The value of medical spending in the United States, 1960–2000. *N Engl J Med.* 2006;355(9):920–927.

21. Bureau of Labor Statistics. Consumer price index. Available at: http://www.bls.gov/cpi/home.htm. Accessed September 1, 2014.

22. Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. *N Engl J Med.* 1977;296(13):716–721.

23. Kenkel D. WTP- and QALY-based approaches to valuing health for policy: common ground and disputed territory. *Environ Resour Econ.* 2006;34(3): 419–437.

24. Medicaid. Oregon. Available at: http://www.medicaid.gov/Medicaid-CHIP-Program-Information/By-State/oregon.html. Accessed September 9, 2014.

25. Sobocki P, Ekman M, Ågren H, et al. Health-related quality of life measured with EQ-5D in patients treated for depression in primary care. *Value Health.* 2007;10(2):153–160.

26. Hirth RA, Chernew ME, Miller E, Fendrick AM, Weissert WG. Willingness to pay for a quality-adjusted life year: in search of a standard. *Med Decis Making.* 2000;20(3):332–342.

27. Tufts Medical Center. The CEA registry. Available at: https://research.tufts-nemc.org/cear4/Resources/LeagueTable.aspx. Accessed October 22, 2013.

28. Tengs TO, Adams ME, Pliskin JS, et al. Five-hundred life-saving interventions and their cost-effectiveness. *Risk Anal.* 1995;15(3):369–390.

29. Franks P, Muennig P, Gold M. Is expanding Medicare coverage cost-effective? *BMC Health Serv Res.* 2005;5(1):23.

30. Muennig P, Franks P, Gold M. The cost effectiveness of health insurance. *Am J Prev Med.* 2005;28(1):59–64.

31. Kawachi I, Adler NE, Dow WH. Money, schooling, and health: mechanisms and causal evidence. *Ann N Y Acad Sci.* 2010;1186:56–68.

32. Truman BI, Smith-Akin CK, Hinman AR, et al.; Task Force on Community Preventive Services. Developing the Guide to Community Preventive Services—overview and rationale. *Am J Prev Med.* 2000;18(1, suppl):18–26.

33. Allen H, Wright BJ, Baicker K. New Medicaid enrollees in Oregon report health care successes and challenges. *Health Aff (Millwood).* 2014;33(2): 292–299.