DOI: 10.1111/j.1475-6773.2008.00834.x
METHODS ARTICLE

Evaluating Health Care Programs by Combining Cost with Quality of Life Measures: A Case Study Comparing Capitation and Fee for Service

Richard Grieve, Jasjeet S. Sekhon, Teh-wei Hu, and Joan R. Bloom

Objective. To demonstrate cost-effectiveness analysis (CEA) for evaluating different reimbursement models.
Data Sources/Study Setting. The CEA used an observational study comparing fee for service (FFS) versus capitation for Medicaid cases with severe mental illness ($n = 522$). Under capitation, services were provided either directly (direct capitation [DC]) by not-for-profit community mental health centers (CMHC), or in a joint venture between CMHCs and a for-profit managed behavioral health organization (MBHO).
Study Design. A nonparametric matching method (genetic matching) was used to identify those cases that minimized baseline differences across the groups. Quality-adjusted life years (QALYs) were reported for each group. Incremental QALYs were valued at different thresholds for a QALY gained, and combined with cost estimates to plot cost-effectiveness acceptability curves.
Principal Findings. QALYs were similar across reimbursement models. Compared with FFS, the MBHO model had incremental costs of $-$ \$1,991 and the probability that this model was cost-effective exceeded 0.90. The DC model had incremental costs of \$4,694; the probability that this model was cost-effective compared with FFS was <0.10.
Conclusions. A capitation model with a for-profit element was more cost-effective for Medicaid patients with severe mental illness than not-for-profit capitation or FFS models.
Key Words. Cost-effectiveness analysis, managed care organizations, QALY

BACKGROUND

Many European countries, Australia, and Canada use cost-effectiveness analysis (CEA) to decide which health technologies to provide (Hutton and Maynard 2000). However, in the United States policy makers do not routinely use

CEA to set health care priorities (Neuman 2004). Commentators have suggested that methodological flaws in published economic evaluations may impede their use in decision making (Rennie and Luft 2000). One concern is that many published studies still use the cost-consequence approach and report costs and effectiveness separately (OHE 2005). These partial evaluations do not provide decision makers with information on any trade-offs between costs and outcomes. Methods are available that provide cost-effectiveness estimates appropriate for use in policy making (NICE 2004).

We illustrate how appropriate CEA techniques can be applied to evaluate a health service intervention using a case study comparing reimbursement models for mental health care. The paper uses a new technique, genetic matching, to adjust for baseline differences in patient mix across the intervention groups (Sekhon 2008). Genetic matching is more appropriate than alternatives such as model-based adjustment, as it does not rely on parametric assumptions that are implausible in this context. The paper reports outcomes using quality-adjusted life years (QALYs), as these can recognize the effect of the reimbursement model on both length and quality of life (QOL). The paper uses recommended methods for dealing with statistical uncertainty in CEA by presenting results using cost-effectiveness acceptability curves (CEACs). Using these methods, the paper demonstrates how CEA can evaluate a "health system" intervention.

Evaluations of "system-level interventions" for mental health services (Hu and Jerrell 1991; Alegria, Frank, and McGuire 2005) and, in particular, different reimbursement models (Manning et al. 1984; Wells, Manning, and Valdez 1990; Dickey 1997; Manning et al. 1999; Bloom et al. 2002; Cuffel et al. 2002; Ray, Daugherty, and Meador 2003) have used the cost-consequence model. Some of these studies reported that reimbursement by capitation was associated with lower costs compared with fee for service (FFS) (Manning et al. 1984; Christiansen et al. 1995; Dickey 1997; Bloom et al. 2002) and no statistically significant differences in outcomes (Wells, Manning, and Valdez 1990; Cuffel et al. 2002); other studies found that capitation was associated with worse quality of care (Manning et al. 1999;

Address correspondence to Richard Grieve, Ph.D., Lecturer, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel St., London WC1E 7HT, U.K. E-mail: Richard.Grieve@lshtm.ac.ok. Jasjeet S. Sekhon, Ph.D., Associate Professor, is with the Travers Department of Political Science, Survey Research Center, University of California, Berkeley, CA. Teh-wei Hu, Ph.D., Professor of Health Economics, Emeritus, is with the School of Public Health, University of California, Berkeley, CA. Joan R. Bloom, Ph.D., Professor, is with the School of Public Health, University of California, Berkeley, CA.

Ray, Daugherty, and Meador 2003). None of these studies combined costs and effectiveness in a formal CEA (Gold et al. 1996; Drummond et al. 2005).[1] None of these studies used a choice-based outcome measure such as the QALY, or appropriate methods for reporting the statistical uncertainty surrounding the cost-effectiveness results.

METHODS

Study Design in CEA

CEA *compares* two or more health care programs, and assesses the *incremental* cost-effectiveness for the decision context of interest. Usually each individual is only observed receiving *one* of the interventions in question. To address this causal inference problem, CEA may be conducted alongside an RCT. However, for CEA of area-level interventions RCT data may be unavailable and the only data may be from a nonrandomized study. As selection is nonrandom, the cases in each group are not drawn from the same population and so cost-effectiveness estimates may reflect preexisting differences between the groups rather than the effect of the intervention itself. Methods are therefore required that minimize differences across the groups, so it is as if the samples are drawn from the same population. We describe principles followed at both the design and analysis stages for minimizing differences across the groups.

A key issue in CEA is to combine costs and outcomes in a way that can appropriately recognize the statistical uncertainty surrounding estimates of cost-effectiveness (Willan and Briggs 2006). In this paper we estimate incremental net benefits (INB) by valuing differences in outcomes across the health care programs by λ, the willingness to pay for a QALY gained. CEACs are then derived by reestimating the INB varying λ between \$0 and \$200,000 per QALY gained and plotting the probability that each program is cost-effective at each value of λ (Fenwick, O'Brien, and Briggs 2004).

Overview of the Case Study

To illustrate how CEA can be applied in health services research, the techniques described were used to compare the cost-effectiveness of different reimbursement mechanisms. The scope of the study was limited to those cases who were already Medicaid beneficiaries, had used a mental health service before, and were diagnosed as having a severe mental illness. This CEA compares FFS with two different capitation models using a study conducted alongside the Colorado Medicaid Capitation Pilot Program (Bloom et al.

2002). In the first capitation, model services were provided directly (direct capitation or DC) by not-for-profit community mental health centers (CMHC). In the second capitation, model services were provided in a joint venture between CMHCs and a for-profit managed behavioral health organization (MBHO). The MBHO was an organization that covered several states, and had previous experience in implementing managed care, but not in the context of Medicaid services. In the remaining regions services continued to be provided by FFS.

Selection of Areas for Each Reimbursement Model

The Colorado Pilot program was implemented in selected parts of the state in August and September 1995 and required all Medicaid beneficiaries to participate, and providers were not allowed to select cases. Hence, unless cases moved or enrolled into Medicaid programs as a consequence of the pilot scheme, selection by patients or providers did not arise. The selection of areas for each reimbursement method was by the state that invited bids for capitation contracts to any entity that had the capacity to receive, manage, and execute a capitation contract for Medicaid mental health services. During the bidding process, existing mental health service providers (CMHCs) grouped together to form Mental Health Service Agencies. The state assessed how ready each entity would be to deliver capitated mental health services for the Medicaid program. In four areas, the state perceived that existing CMHCs were ready for capitation and awarded them direct contracts with the state (DC model). In three areas, the state judged that existing CMHCs were "not ready" for capitation and awarded the contract to a for-profit managed behavioral firm. The state encouraged this firm to form an alliance with existing CMHCs, which the study termed the MBHO. In three areas that the state judged inappropriate for capitation, reimbursement by FFS was maintained (Bloom et al. 1998). The key concern for the CEA was that the selection of areas for capitation was nonrandom and according to the perceived readiness of the organizations concerned for capitation. The state assessed "readiness for capitation" according to criteria such as whether there was an appropriate management information system, whether there was a suitable financial system in place for costing services, and whether there were appropriate strategies for utilization review (UR) (see Bloom, Devers, and Wallace 2000). The DC areas scored highest on these readiness criteria. The for-profit managed behavioral firm had no previous experience of administering capitation services for Medicaid. As the DC group was perceived to be "most ready," it was

anticipated that the nonrandom selection would exaggerate any cost reductions observed in this group.

Sampling Strategy

The purpose of the study was to compare the relative cost-effectiveness across three different reimbursement models; hence, it was important to minimize differences in area and patient characteristics across all three groups. The study used a matched group design, which aimed to include similar areas across the three groups. The study used 1990 U.S. census data on the proportion of the population in each area in poverty, the degree of rurality, and the industrial base as it was anticipated that these variables could be associated with costs and outcomes (Bloom et al. 1998). The study then selected those counties that had similar area-level characteristics (see Supplementary Material Appendix SA1).

From those areas included, the study took a random sample of those cases who were already Medicaid beneficiaries, had used a mental health service before, and were diagnosed as having a severe mental illness (diagnoses of schizophrenia, bipolar affective disorder, or at least one 24-hour inpatient stay with a primary mental health [DSM-IV] diagnosis). A total of 522 cases were available for the CEA.

Measurement of Cost and Utilization

Cost and outcome data were collected for 1 year precapitation when all regions were reimbursed by FFS and 2 years postcapitation. In the period immediately following capitation, the first 3 months were viewed as an implementation period, and were excluded from the CEA as were the corresponding periods in the second period postcapitation and precapitation. This gave cost and outcome data for three 9-month periods (one pre-, and two postcapitation).

The cost measurement took a Medicaid perspective and excluded costs borne by other payers. Costs included were those in the capitation rates that covered all Medicaid-eligible individuals for psychiatric inpatient care, specialty mental health outpatient services, and mental health services in nursing homes, but excluded the cost of pharmaceuticals. Costs for all three groups before capitation and for the FFS group for all three time periods (1995–1998) were taken from Medicaid claims databases. Cost data were not available from the Medicaid claims database for the capitation group following capitation; these data were recorded from the state's shadow billing system. The shadow

billing system required the capitated providers to report identical cost information to claims data.

The study measured the total costs of each episode of care for each user, including inpatient stays (state and local hospitals) and outpatient care (individual or group therapy, case management, and day treatment programs). These total costs per episode were used to derive measures of unit cost and utilization such as the proportion of cases using inpatient or outpatient services during each period.

Measurement and Valuation of Health Outcomes

The CEA reported health outcomes using QALY, which required that the vital status of each case was noted, and for the decedents, information on the date of death was obtained from death certificates to record survival duration. To estimate health-related quality of life (HRQOL), trained investigators administered the SF-36 health survey at 6 monthly intervals throughout the study. The algorithm developed by Brazier, Roberts, and Deverill (2002) was chosen to value the health states described by the SF-6D, a subsample of the SF-36 health states. For each case, HRQOL at each time point was multiplied by the corresponding survival time to give QALYs for each 9-month period.

Matching at the Analysis Stage

Randomizing a sufficiently large number of cases to each reimbursement model would ensure that there were no baseline differences in patient or center characteristics across the intervention groups. This nonrandomized study recorded patient characteristics before the introduction of capitation, and despite the attempts to match areas with similar characteristics at the design stage, there were differences between the patient groups at baseline (see Table 1). For example, mean costs before capitation were significantly higher in the MBHO ($6,822) than the FFS group ($4,820) ($t$-test $p = .02$). These differences in baseline costs partly reflect differences in patient mix, for example the mean costs for men were higher than those for women, and the MBHO group had the highest proportion of men. However, the MBHO model clearly had higher baseline costs even after allowing for differences in patient factors. Hence, it is important to match on baseline cost as well as case-mix variables. By adjusting the samples according to baseline cost, the analysis recognizes differences in baseline cost that arise according to the areas concerned.

Table 1: Baseline Costs ($), QALYs, Client Characteristics, and Utilization: Before and after Matching*

	FFS versus DC			*FFS versus MBHO*		
	FFS	*DC*	*p-value*†	*FFS*	*MBHO*	*p-value*
Mean costs						
Before matching	4,820	4,524	.11	4,820	6,822	.02
After matching	4,820	4,805	.32	4,820	4,580	.42
Mean QALYs‡						
Before matching	0.475	0.485	.10	0.475	0.482	.29
After matching	0.475	0.476	.78	0.474	0.474	.33
% Schizophrenia						
Before matching	72.2	61.9	.05	72.2	65.1	.16
After matching	72.2	70.8	.68	72.2	72.2	1.00
% Bipolar						
Before matching	21.2	30.7	.05	21.2	25.6	.33
After matching	21.2	22.5	.48	21.2	21.9	.56
Mean age						
Before matching	43.4	42.3	.58	43.4	45.1	.32
After matching	43.4	43.4	.84	43.4	43.7	.86
% Men						
Before matching	44.3	47.7	.54	44.3	49.7	.32
After matching	44.3	43.7	.78	44.3	45.0	.70
% Previous high cost client						
Before matching	37.1	36.4	.89	37.1	31.8	.31
After matching	37.1	37.1	1.00	37.1	36.4	.32
% Using any service						
Before matching	89.4	93.8	0.16	89.4	90.3	.80
After matching	89.4	89.4	1.00	89.4	89.4	1.00

*Before matching: $n = 522$, FFS ($n = 151$), DC ($n = 176$), MBHO ($n = 195$); after matching: $n = 453$ ($n = 151$ in each group).

†The tests conducted are nonparametric bootstrap Kolomogorov–Smirnov distributional tests.

‡Note that QALY data were not available for eight cases before, and four cases after matching.

FFS, fee for service; DC, direct capitation; MBHO, managed behavioral health organization; QALY, quality-adjusted life year.

Where there are large imbalances in baseline covariates as in this case study, using a parametric model to adjust for differences is problematic; the results are generally sensitive to the choice of model specification (Rubin 2006). The previous cost analysis of the same data used a parametric model, the two-part model, to try and adjust for baseline differences between the groups (Bloom et al. 2002). A problem with this approach is that it only allows for *mean* differences across the groups, and therefore ignores differences elsewhere in the distribution.

To allow causal inferences to be made when parametric adjustment is problematic, matching methods are recommended (Rubin 2006). This study employs a nonparametric matching method, genetic matching, which is a generalization of propensity score and Mahalanobis distance matching (Raessler and Rubin 2005; Morgan and Harding 2006). The method has been shown to outperform more commonly used matching methods (such as propensity scores) and has been applied in a wide range of areas (see, e.g., Raessler and Rubin 2005; Morgan and Harding 2006; Herron and Wand 2007). The method does not require the analyst to make parametric assumptions, which is important in this context given that cost data generally have highly irregular distributions. The method uses a genetic algorithm (Sekhon and Mebane 1998) to identify those matches that achieve the best possible covariate balance (Diamond and Sekhon 2006; Sekhon 2008).

In this case study, genetic matching was used to identify cases in each capitation group to match to cases in the FFS group. The matching algorithm used the same covariates as the previous parametric model, which were baseline measures for demography (age, gender, ethnicity), diagnosis (schizophrenia, bipolar affective disorder, other), precapitation utilization, QALYs, and cost. The algorithm selected cases using the results of *t*-tests and nonparametric Kolmogorov–Smirnov (KS) tests that compared the distribution of these covariates across the groups. The KS test is a nonparametric test of the equality of two empirical cumulative distributions. This test is distribution free; hence, it does not rely on the assumption of normality, which is important given the highly skewed and kurtotic distribution of cost data. When the KS test is bootstrapped, it is consistent even when variables do not have a continuous distribution (Abadie 2002). For example, in this dataset the distribution of the cost variable has a point mass at zero and it is certainly not normally distributed.

After applying the matching algorithms, no significant differences remained between the groups (Table 1). All subsequent analyses were conducted using the matched dataset.

Cost and CEA

Costs and QALYs were reported for each patient for each observation period (9 months precapitation, and two 9-month periods postcapitation). Costs and QALYs in the second follow-up period were discounted at the recommended rate of 3 percent (Gold et al. 1996). Total costs and QALYs were calculated by summing costs and QALYs across the two follow-up periods. Given the

skewed nature of the cost data, the analysis did not assume that the data were drawn from a normal distribution, and instead used the bootstrap KS test, and the nonparametric bootstrap (bias corrected) to report 95 percent CIs around incremental costs and QALYs (Thompson and Barber 2000). CEACs were derived by using the bootstrap replications to plot the probability that each capitation model was cost-effective at different values for λ.

The CEA was repeated for different patient subgroups, for example those cases with schizophrenia as opposed to bipolar affective disorder. Sensitivity analysis applied parametric models to adjust for remaining differences in patient and area-level characteristics across the groups. This analysis used a two-part model to estimate incremental costs (Mullahy 1998), and a multiple linear regression model to estimate incremental effectiveness.

RESULTS

For these previous users of mental health services, service utilization fell in all three groups over the study's observation periods. For inpatient services, the reduction in service use was similar across the groups (Table 2). These overall changes may reflect reversion to the mean; however, the key finding is that there were differences in the reduction in outpatient services according to the reimbursement model. The reduction in outpatient utilization was largest in the MBHO group, where there was a 22 percent reduction by the end of the second follow-up period (post 2nd) compared with a 7 percent reduction in the FFS group ($p = .04$). The corresponding reduction in outpatient utilization in the DC group (12 percent) was not significantly different to the FFS group ($p = .29$) (Table 2). The mean cost for service users was lower postcapitation in the MBHO group but higher in the DC group, compared with FFS. The net effect of these changes in utilization and cost was that postcapitation, the mean costs per case were higher for the DC model than for the FFS model, whereas the MBHO model had lower mean costs per case (Table 2).

A total of 373 (82 percent) cases completed SF-36 surveys at each time point; the mean HRQOL was 0.63 for each group at baseline (Table 2). The mean HRQOL was higher in the MBHO group at follow-up, and so this group had higher mean QALYs.

Compared with FFS, the MBHO model had negative incremental costs (−$1,991). Although the bootstrapped 95 percent CIs around this estimate of incremental costs included zero (Table 3), the p-value for the bootstrapped KS test was .01. This nonparametric KS test is more appropriate given the highly

Table 2: Utilization of Services (%), Mean Costs ($), HRQOL, and QALYs; Pre- and Postcapitation

Time period	*FFS*	*DC*	*MBHO*
Utilization of services (n = 453)			
Inpatient			
Pre	15.2	15.2	14.6
Post (1st)	10.6	3.3	9.9
Post (2nd)	8.6	8.6	9.9
Outpatient			
Pre	89.4	89.4	87.4
Post (1st)	88.7	82.8	75.5
Post (2nd)	83.4	78.8	68.2
Costs (n = 453)			
Cost per user*			
Pre	5,391	5,375	5,123
Post (1st)	4,888	7,116	3,837
Post (2nd)	4,794	9,002	4,714
Cost per case*			
Pre	4,820	4,805	4,580
Post (1st)	4,338	5,938	2,989
Post (2nd)	4,000	7,094	3,359
Outcomes (n = 373)			
HRQOL			
Pre	0.63	0.63	0.63
Post (1st)	0.64	0.62	0.64
Post (2nd)	0.63	0.61	0.65
QALY			
Post 1st+Post 2nd	0.934	0.919	0.954

*Note that all cases in the sample used service before study entry, cost per user gives the cost for those using services in the given period, whereas cost per case reports costs for all those in the sample.

HRQOL, health-related quality of life; QALY, quality-adjusted life year; FFS, fee for service; DC, direct capitation; MBHO, managed behavioral health organization.

nonnormal distribution of the cost data. The DC model had positive incremental costs of $4,694 compared with FFS (95 percent CI from 302 to 10,170; KS test $p = .08$). The incremental costs of the MBHO model compared with DC were −$6,685 (95 percent CI from −$11,242 to −$1,658). Aside from the significant difference in mean costs, the MBHO model had significantly lower costs as determined by the nonparametric KS test ($p = .002$). Indeed, the MBHO model had lower costs across the entire distribution of costs (empirical QQ plots available upon request).

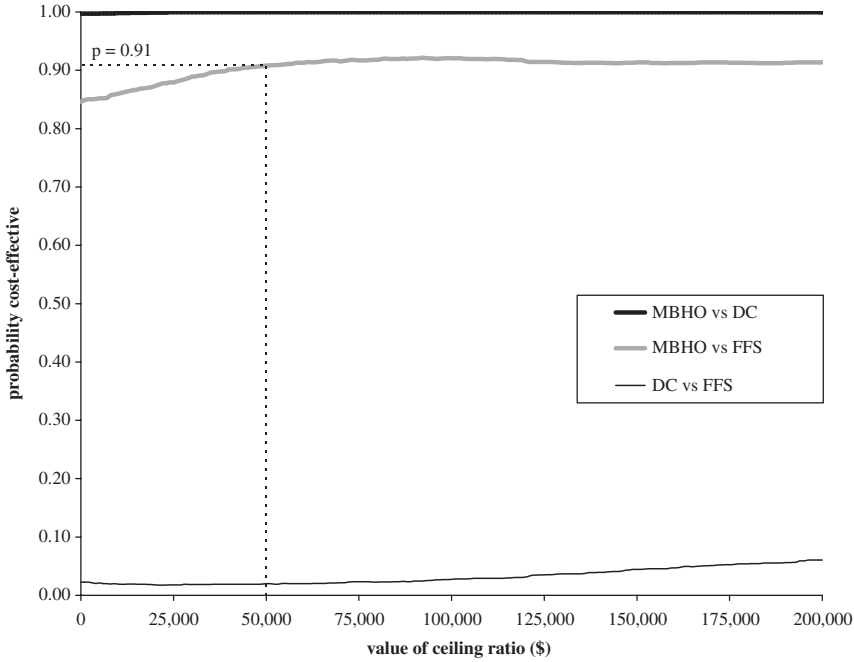
The MBHO model had positive incremental QALYs compared with FFS or DC, although the confidence intervals around the central estimates

Table 3: Incremental Costs ($), Incremental QALYs, and INBs ($) (Mean Estimates [95% CI])

	DC–FFS	*MBHO–FFS*	*MBHO–DC*
Incremental costs	4,694 (302 to 10,170)	− 1,991 (− 5,801 to 1,839)	− 6,685 (− 11,242 to − 1,658)
Incremental QALYS	− 0.016 (− 0.061 to 0.026)	0.019 (− 0.017 to 0.059)	0.035 (− 0.006 to 0.073)
Incremental net benefit (λ = $50,000)	− 5,477 (− 10,832 to − 542)	2,950 (− 1,697 to 7,078)	8,428 (3,338 to 13,297)
Incremental net benefit (λ = $100,000)	− 6,262 (− 12,779 to − 13)	3,908 (− 1,717 to 9,279)	10,169 (3,890 to 16,113)

QALY, quality-adjusted life year; INB, incremental net benefit; FFS, fee for service; DC, direct capitation; MBHO, managed behavioral health organization.

Figure 1: Cost-Effectiveness Acceptability Curves



were wide and included zero. The CEA found that the DC model was not cost-effective compared with the FFS or MBHO models, across a range of values for the cost-effectiveness threshold, λ. For example, when λ was valued at \$50,000 per QALY, the INB for the DC model compared with FFS was $-$\$5,477 (Table 3). The MBHO model was cost-effective compared with either the FFS or DC models. For example, at \$50,000 per QALY, the mean INB of MBHO compared with DC was \$8,428 (95 percent CI from \$3,338 to \$13,297) (Table 3).

The CEACs plot the probability that either capitation model is cost-effective for different levels of λ (Figure 1). The intersection with the *y*-axis shows the probability that "the intervention" is cost-effective when only cost differences are considered. As the value for λ increases, relatively more weight is given to the incremental effectiveness. At all realistic levels of the cost-effectiveness threshold, the probability that the MBHO model is cost-effective compared with either FFS or DC exceeds 0.90. For example at \$50,000 per QALY, the probability that the MBHO model is cost-effective compared with

FFS is 0.91. The CEAC for the MBHO versus FFS comparison does not asymptote to 1 as it is not certain that the MBHO model is more effective than FFS; although the mean incremental QALYs are positive, the CIs surrounding this estimate include zero. The CEACs also show that the probability that the DC model is cost-effective compared with FFS is <0.1 at all values of λ.

The sensitivity analysis found that the MBHO model remained the most cost-effective after applying parametric and semi-parametric models to allow for any remaining differences across the groups postmatching. As the data were well matched, the results were not sensitive to the choice of model. The subgroup analysis showed that for patients with schizophrenia (72 percent of cases) the MBHO model was most cost-effective. However, for patients with bipolar affective disorder, both capitation models were associated with increased costs and no gain in QALYs compared with FFS.

DISCUSSION

This paper presented some key methodological features of CEA and illustrated these techniques with a case study. The CEA found that the capitation model with the for-profit component was the most cost-effective at all levels of willingness to pay for a QALY gained. The CEA incorporated any differences in both costs and outcomes across the reimbursement models, and therefore extended previous cost minimization analyses that have focused on the relative costs of managed care compared with FFS (Manning et al. 1984; Dickey 1997; Bloom et al. 2002). The CEA used appropriate techniques to measure and value outcomes, to deal with baseline imbalances across the groups (Morgan and Harding 2006), and to allow for the skewed distribution of the cost data (Abadie 2002). The techniques presented could be used more generally for evaluating different ways of financing and providing health services where there may be differential impacts on costs and outcomes and where RCT data are unavailable.

An earlier paper reporting cost results from the same study found that both the not-for-profit capitation model (DC) and the capitation model with a for-profit element (MBHO model) were associated with cost reductions compared with FFS (Bloom et al. 2002). Our paper finds that the DC model is associated with higher costs, and the MBHO model is associated with lower costs compared with FFS. Under the DC model the costs for service users were higher compared with FFS, whereas in the previous paper these costs were reported as similar in the DC and FFS groups.

The reason for the difference in the cost results across the papers is the approach taken to adjusting for baseline differences across the groups. The previous paper used a parametric model, the two-part model, and only allowed for *mean* differences across the groups at baseline. This is a particular deficiency for a variable such as baseline cost, which is highly skewed; using the mean differences at baseline ignores important differences elsewhere in the distribution. Instead, we used a nonparametric technique, genetic matching, as recommended in the biostatistics literature (Rubin 2006). The two key advantages of genetic matching are as follows: firstly, it did not rely on parametric assumptions such as assuming that the baseline costs were normally distributed; secondly, rather than just adjusting the samples based on mean characteristics, it allowed for baseline differences across the groups right across the distribution. When this method was applied, excellent covariate balance was achieved. Our results are not sensitive to model-based parametric adjustment postmatching.

The study illustrated that CEA can provide clear information on the relative cost-effectiveness of alternative reimbursement methods. Methodological guidance for economic evaluation requires that authors place appropriate limits on the generalizability of their results (Drummond et al. 2005). It is therefore important to recognize that the finding that a capitation model with a for-profit element was more cost-effective than a not-for-profit capitation model may not be transferable to other health care contexts. When capitation was introduced for Colorado Medicaid mental health services, the state took steps to try and maintain service quality. For example, the state specified the services to be delivered in the capitation contract; strict limits were imposed on profits and further investment in mental health services was encouraged. These features may have been important in ensuring that similar health outcomes were maintained across reimbursement models. In other contexts, if capitation schemes are less carefully implemented, they can lead to poorer quality of care (Ray, Daugherty, and Meador 2003), and may be less cost-effective than FFS.

This study was restricted to previous users of mental health services. These patients were relatively costly (average cost of $7,500 per year) and there may have been more scope for reductions in utilization for these users than for other groups, for example patients with less severe mental illness or newly identified patients. The subgroup analysis found that while the for-profit model was most cost-effective for patients with schizophrenia, FFS was more cost-effective for patients with bipolar affective disorder who had lower average costs.

The cost-effectiveness results in the case study were driven by cost differences across the reimbursement models. A potentially important feature of the capitation models was that contracts were retendered every 2 years. In the for-profit areas the contracts moved between health care organizations, whereas in the DC areas the contracts remained with the same CMHCs. Faced with this greater risk coupled with the incentive to make profits, the MBHO group may have been more inclined to adopt processes that reduced costs while maintaining quality. For example, a qualitative investigation of care processes found that in the MBHO areas UR informed the management of each case (Bloom, Devers, and Wallace 2000). By contrast in the DC areas, administrators only employed UR for outlying cases. For patients with severe mental illness, costs are notoriously difficult to predict and using UR for all cases would be more likely to identify those cases with scope for cost reduction. Furthermore, interviews with decision makers in the DC areas suggested that, faced with little incentive to reduce costs, there was more emphasis on expanding services (Bloom, Devers, and Wallace 2000). This strategy appeared to lead to higher costs without improvements in patient outcomes.

General concerns that capitation leads to "cream skimming" are unlikely to apply in this study as health care providers were legally required to maintain access to care for the cases in the study who were all Medicaid enrollees. The state selected for the not-for-profit capitation model those CMHCs judged "ready" for capitation; those CMHCs judged "not ready" were linked with a for-profit MBHO (Bloom, Devers, and Wallace 2000). It was anticipated that this selection process would lead the CEA to overstate the cost-effectiveness of the not-for-profit capitation model. As the study found that the for-profit capitation model was relatively cost-effective, the findings are robust to bias in the selection of centers.

Guidelines for CEA recommend that ideally a broad range of costs are included and a lifetime time horizon is taken for the analysis (Luce et al. 1996). Compared with this "gold standard," the case study presented had certain limitations; for example, costs outside the capitation contract including pharmaceuticals were excluded. Another study found that the only difference in pharmaceutical costs was that the DC group used more antipsychotic medication compared with FFS (Wallace et al. 2005). Hence, including these costs would further substantiate the conclusion that the DC model was not cost-effective. Of greater concern is the relatively short time frame adopted. While a follow-up study found that the MBHO and DC models had similar costs after 6 years (Wallace et al. 2006), further research is required to evaluate the long-

term cost-effectiveness of different reimbursement mechanisms using the techniques outlined.

The methods presented are of general use to policy makers aiming to reduce costs without compromising the quality of care. They are particularly relevant for evaluating Medicaid programs where budgetary pressures are perennial (Johnson 2005). CEA highlights trade-offs between costs and outcomes, allowing policy makers with differing views on the relative importance of costs versus outcomes to use the same analysis.

In conclusion, this study illustrates appropriate methods for estimating and valuing health outcomes, adjusting for differences in patient mix across the intervention groups, and representing the sampling uncertainty surrounding the results. The case study found that a capitation model with a for-profit element was more cost-effective than either a not-for-profit capitation or an FFS model for Medicaid patients with severe mental illness. These techniques can be applied to a wide range of contexts in health services research, to help policy makers identify which health care programs to prioritize.

ACKNOWLEDGMENTS

This work was undertaken while R. G. was a visiting scholar at the University of California, Berkeley; we acknowledge the U.K. Medical Research Council for their funding. We also thank Professor Richard Scheffler and the Petris Center (UC Berkeley) for supporting this work. We acknowledge Soo Pak (UC Berkeley) for her database assistance throughout this project. We thank Professor Neal Wallace (Portland State University), Dr. Janet Coffman, and Dr. Anne Morris (UC, San Francisco) for their helpful advice.

Disclosures: None.

Disclaimers: None.

NOTE

1. Here CEA is defined broadly to include studies that report outcomes as utilities.

REFERENCES

Abadie, A. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *Journal of the American Statistical Association* 97 (457): 284–92.

Alegria, M., R. Frank, and T. McGuire. 2005. "Managed Care and Systems Cost-Effectiveness: Treatment for Depression." *Medical Care* 43 (12): 1225–33.

Bloom, J. R., K. Devers, N. T. Wallace, and N. Z. Wilson. 2000. "Implementing Capitation of Medicaid Mental Health Services in Colorado: Is 'Readiness' a Necessary Condition?" *Journal of Behavioral Health Services and Research* 27 (4): 437–45.

Bloom, J. R., T. W. Hu, N. Wallace, B. Cuffel, J. W. Hausman, and R. Scheffler. 1998. "Mental Health Costs and Outcomes under Alternative Capitation Systems in Colorado: Early Results." *Journal of Mental Health Policy and Economics* 1 (I): 1–11.

Bloom, J. R., T. W. Hu, N. Wallace, B. Cuffel, J. W. Hausman, M. L. Sheu, and R. Scheffler. 2002. "Mental Health Costs and Access under Alternative Capitation Systems in Colorado." *Health Services Research* 37 (2): 315–40.

Brazier, J., J. Roberts, and M. Deverill. 2002. "The Estimation of a Preferenced-Based Measure of Health from the SF-36." *Journal of Health Economics* 21: 271–92.

Cuffel, B. J., J. R. Bloom, N. Wallace, J. W. Hausman, and T. W. Hu. 2002. "Two-Year Outcomes of Fee-for-Service and Capitated Medicaid Programs for People with Severe Mental Illness." *Health Services Research* 37 (2): 341–59.

Diamond, A., and J. S. Sekhon "Genetic Matching for Estimating Causal Effects: A General Matching Method for Achieving Balance in Observational Studies." Working Paper. The Society of Political Methodology [accessed on September 19, 2006]. Available at http://sekhon.berkeley.edu/papers/GenMatch.pdf

Dickey, B. 1997. "Assessing Cost and Utilization in Managed Mental Health Care in the United States." *Health Policy* 41 (suppl): S163–74.

Drummond, M. F., M. J. Sculpher, G. W. Torrance, B. J. O'Brien, and G. L. Stoddart. 2005. *Methods for the Economic Evaluation of Health Care Programmes*, 3d Edition. Oxford: Oxford University Press.

Fenwick, E., B. J. O'Brien, and A. Briggs. 2004. "Cost-Effectiveness Acceptability Curves—Facts, Fallacies and Frequently Asked Questions." *Health Economics* 13: 405–15.

Gold, M. R., J. E. Siegel, L. B. Russell, and W. C. Weinstein (eds). 1996. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press.

Herron, M. C., and J. Wand. 2007. "Assessing Partisan Bias in Voting Technology: The Case of the 2004 New Hampshire Recount." *Electoral Studies* 26 (2): 247–61.

Hu, T. W., and J. Jerrell. 1991. "Cost-Effectiveness of Alternative Approaches in Treating Severely Mentally Ill in California." *Schizophrenia Bulletin* 17: 461–8.

Hutton, J., and A. Maynard. 2000. "A Nice Challenge for Health Economics." *Health Economics* 9: 89–93.

Johnson, P. 2005. "Medicaid: Medicaid: Provider Reimbursement—2005. End of Year Issue Brief." *Issue Brief Health Policy Tracking Service* Dec 31: 1–11.

Luce, B., W. Manning, J. E. Siegel, and J. Lipscomb. 1996. "Estimating Costs in Cost-Effectiveness Analysis." In *Cost-Effectiveness in Health and Medicine*, edited

by M. R. Gold, J. E. Siegel, L. B. Russell, and M. C. Weinstein, pp. 176–213. New York: Oxford University Press.

Manning, W. G., A. Leibowitz, G. A. Goldberg, W. H. Rogers, and J. P. Newhouse. 1984. "A Controlled Trial of the Effect of a Prepaid Group Practice on Use of Services." *New England Journal of Medicine* 310 (23): 1505–10.

Manning, W., C. F. Liu, T. J. Stoner, D. Z. Gray, N. Lurie, M. Popkin, and J. B. Christianson. 1999. "Outcomes for Medicaid Beneficiaries with Schizophrenia under a Prepaid Mental Health Carve-Out." *Journal of Behavioral Health Services and Research* 26: 442–50.

Mebane, W. R. J., and J. S. Sekhon. 1998. "GENetic Optimization Using Derivatives (GENOUD)" [accessed on September 19, 2006]. Available at http://sekhon.berkeley.edu/rgenoud/

Morgan, S. L., and D. J. Harding. 2006. "Matching Estimators of Causal Effects." *Sociological Methods and Research* 35 (1): 3–60.

Mullahy, J. 1998. "Much Ado about Two: Reconsidering Transformations and the Two-Part Model in Health Econometrics." *Health Economics* 17 (3): 181–247.

National Institute for Clinical Excellence (NICE). 2004. *Guide to the Methods of Technology Appraisal.* London: NICE.

Neuman, P. J. 2004. "Why Don't Americans Use Cost-Effectiveness Analysis?" *American Journal of Managed Care* 10: 308–12.

Office of Health Economics. 2005. *OHE—Health Economic Evalvation Database.* London: OHE.

Raessler, S., and D. B. Rubin. 2005. "Complications When Using Nonrandomized Job Training Data to Draw Causal Inferences." Sidney: Proceedings of the International Statistical Institute.

Ray, W. A., J. R. Daugherty, and K. G. Meador. 2003. "Effect of a Mental Health Carve-Out Program on the Continuity of Antipsychotic Therapy." *New England Journal of Medicine* 348: 1885–94.

Rennie, D., and H. S. Luft. 2000. "Pharmacoeconomic Analyses: Making Them Transparent, Making Them Credible." *Journal of the American Medical Association* 283: 2158–60.

Rubin, D. 2006. *Matched Sampling for Causal Effects.* New York: Cambridge University Press.

Sekhon, J. S. 2008. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for R." *Journal of Statistical Software* (in press). Computer program available at http://sekhon.polisci.berkeley.edu/matching/

Thompson, S. G., and J. A. Barber. 2000. "How Should Cost Data in Pragmatic Trials Be Analysed?" *British Medical Journal* 320: 1197–2000.

U.S. Department of Health and Human Services, Health Resources and Services Administration, and Bureau of Health Professions. 2005. *Area Resource File (ARF).* Rockville, MD.

Wallace, N. T., J. R. Bloom, T. W. Hu, and A. M. Libby. 2005. "Medication Treatment Patterns for Adults with Schizophrenia in Medicaid Managed Care in Colorado." *Psychiatric Services* 56 (11): 1402–8.

Wallace, N., J. Hyun, H. Wang, S. H. Kang, T. W. Hu, and J. Bloom. 2006. "Six Years of Longitudinal Analyses of Capitation in Colorado" [unpublished manuscript].

Wells, K., W. Manning, and R. B. Valdez. 1990. "The Effects of a Prepaid Group Practice on Mental Health Outcomes." *Health Services Research* 25 (4): 615–25.

Willan, A., and A. Briggs. 2006. *Statistical Analysis of Cost-Effectiveness Data.* Chichester, U.K.: Wiley.

SUPPLEMENTARY MATERIAL

The following material is available for this article is available online:

Appendix SA1: Characteristics of the Areas and Centers Included in the Study. Data Presented Are Weighted According to the Cases Included, Post Matching.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1475-6773.2008.00834.x (this link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.