Persistence of HMO Performance Measures

*Shailender Swaminathan, Michael Chernew, and
Dennis P. Scanlon*

Objective. To simplify the decision-making process, we propose and implement an approach to assess the stability of health plan performance over time when multiple indicators of performance exist.

Data Source. National Committee for Quality Assurance Health Care Effectiveness Data and Information Set data for childhood immunization for both publicly and nonpublicly reporting health plans between 1998 and 2002.

Data/Study Design. We use longitudinal data to examine whether plan quality ratings are stable from year to year. We estimate a parametric Multiple Indicator Multiple Cause Model, a model which allows us to aggregate the multiple measures of performance. The model controls for observed characteristics of the plan and market, allowing for unmeasured heterogeneity.

Principal Findings. We find moderate persistence in plan performance over time. A plan in the upper tier of performance in the year 1999 has only a 0.47 probability of remaining in the upper tier in the year 2001. Multiple years of good performance increase the probability of good performance in the future. For example, from the subset of plans in the upper tier of performance in 1999, 63 percent continued to perform in the upper tier in 2000. However, from the subset of plans in the upper tier in both 1998 and 1999, about three-fourths of the plans continued to perform in the upper tier in the year 2000. Finally, better performance in the more recent past is more indicative of better performance in the future than better performance in the more distant past.

Conclusions. Although there is some persistence in health plan ratings over time, it is not uncommon for ratings of plans to change between when the data are generated and when actions based on that data, such as employers' contracting decisions or consumers' enrollment decisions, may take effect. Decision makers should be cognizant of this issue and methods should be developed to mitigate its consequences.

Key Words. HMOs, health plans, quality improvement, HEDIS

Efforts to measure the performance of health plans on clinical dimensions have expanded rapidly over the past 10–15 years. Clinical performance measures can support plan quality improvement efforts, employer/government

insurance contracting, and consumer plan enrollment decisions. The foundation of many of the plan measurement initiatives is the Health Care Effectiveness Data and Information Set (HEDIS), maintained by the National Committee for Quality Assurance (NCQA).

Typically, reports based on HEDIS utilize data collected before enrollment or contracting decisions are made, and the decisions themselves generally do not take effect for some time after the decision date. For example, measures of health plan performance reported for a company's open health insurance enrollment period in (say) the fall of 2005 are based on data generated by the plan in the year 2004 or earlier, and decisions regarding plan choice made in the fall of 2005 do not take effect until 2006. Thus, there is typically at least a 2-year lag (e.g., 2004–2006) between plan performance and when any action based on that performance would take effect. For this reason understanding the stability of measures of performance over time is crucial to understanding how well these measures can be used to support decision making.

Although internal quality improvement initiatives use individual HEDIS or CAHPS measures (Scanlon et al. 2001), decision making is often easier using aggregate measures. For example, from a consumers' viewpoint, having multiple measures of provider quality can be confusing and at times contradictory (Scanlon et al. 1998). For this reason it is important to develop aggregate measures of plan performance.

Several studies have developed methods to estimate provider quality when multiple indicators exist (Landrum, Bronskill, and Normand 2000; Staiger and McClellan 2000; Roy and Lin 2002), with a subset of these also estimating the determinants of health plan quality (Scanlon et al. 2005). However, none of the aforementioned papers evaluates the extent to which provider or health plan performance is stable.

If we consider only a single indicator of quality per provider, then a simple way to assess stability in health plan performance is to find the one-step transition probability. For example, we could calculate the fraction of plans that had performance scores above/below a certain threshold in (say) 2005 given that they were above/below the same threshold in (say) 2004. Because

Address correspondence to Shailender Swaminathan, Ph.D., Center for Gerontology and Health Care Research, Brown University, 121 South Main Street, Providence, RI 02912; e-mail: Shailender_Swaminathan@brown.edu. Michael Chernew, Ph.D., is with the Harvard Medical School, Harvard University, Boston, MA. Dennis P. Scanlon, Ph.D., is with the Pennsylvania State University Health Policy and Administration, University Park, PA.

this estimate is derived without making any assumptions about the underlying distribution that generates the data, this is also called a nonparametric estimate of the transition probability. Despite its great simplicity, this approach is difficult to implement when multiple indicators of quality exist because there would be separate one-step transition probabilities for each individual measure.

In this paper, we estimate a statistical model of the determinants of plan quality that allows for multiple indicators of performance in each period and for multiple periods. Health plan performance will reflect the degree of health plan management (e.g., disease or care management), the physician network, and characteristics of the patient population, including compliance with clinical advice. There are a variety of reasons why plan performance might not be persistent. This lack of persistence might stem from true variation in performance because outcomes are stochastic, underlying plan processes or market environments may change, and measurement error may exist. Finally, longitudinal changes in performance can result from plan quality improvement initiatives that vary over time and that may have lagged effects.

We take a “parametric approach,” assuming performance is determined by a multivariate stochastic model. In particular, we estimate the parameters of a model that assumes that the multiple indicators in each period are linked via a latent variable that represents true performance. We further allow for measurement error in the individual performance indicators. True performance is assumed to be determined by a set of measured covariates that includes plan and market characteristics (e.g., profit status of the plan, HMO competition, etc.). A part of the variance in true performance is assumed to be driven by unmeasured heterogeneity. We use the estimated parameters to compute the degree to which plan performance is stable. The model is similar to those used in the literature to examine stability across poverty states among families and individuals (Lillard and Willis 1978; Gottschalk 1982).

The use of a parametric model, as opposed to simply reporting the stability of a single measure over time, has several advantages. First, because there are multiple measures, estimates of the stability of each individual measure would need to be aggregated to provide a summary measure of stability. A model that allows this aggregation in a structured way is consistent with the statistics literature on aggregation of performance measures (Landrum, Bronskill, and Normand 2000). However, we note that aggregation of this sort only makes sense when the measures themselves are highly correlated both within and across periods. In cases where there is a low correlation among measures, we lose a lot of information that might be specific to a particular measure.

Second, the use of a model allows us to address the small sample problem that would arise because we observe an unbalanced panel over a relatively short time frame. By parameterizing the model, we increase our ability to make inferences about the stability of performance over longer periods of time. A further advantage of the parametric model is that we can evaluate how the transition probabilities change as plan and market-level characteristics (i.e., profit status, market competition) change over time. The parameterized model allows us to adjust for changes in observed traits and to disaggregate changes in performance into transitory and persistent components.

We apply the parametric model to HMO performance using six childhood immunization measures from NCQA's HEDIS data for the period 1998–2002. Childhood immunizations are crucial to maintaining public health and represent a cost-effective method for delivering preventative health services to children (National Institute of Health (NIH) 2003). According to data from the National Immunization Survey 2006, only about 78 percent of the children receive the recommended schedule of immunizations by age two. In 1998, the National Vaccine Advisory Committee published a report that outlined strategies to sustain success in childhood immunizations (NVAC 1999). Their primary recommendation was for health plans to guarantee complete immunization of its eligible children using the schedule endorsed by the American Academy of Pediatrics.

Luft and Romano (1993) provide the only other study assessing the stability of health performance measures. They examined the stability of hospital-level measures of mortality from cardiac surgery. That paper did not use a parametric model or attempt to aggregate multiple measures or relate stability to hospital traits. However, it did illustrate the many patterns of instability over time that might exist. No such work exists for health plan performance.

DATA AND SAMPLE

The primary data sources used to derive the analytic sample were NCQA's HEDIS data and the Interstudy Corporation's *MSA Profiler* and *Competitive Edge* (calendar year 1998–2002 data). NCQA data included observations from all plans reporting HEDIS data in NCQA's publicly available data product—*Quality Compass 1999* through *Quality Compass 2003* (i.e., publicly reporting plans), as well as plans that reported data to NCQA but requested that the information not be included in *Quality Compass* (i.e., nonpublicly reporting

plans). The HEDIS data reflect member health care encounters and survey responses occurring during the prior calendar year (e.g., 1999 for *Quality Compass 2000*). We use data from all 457 plans that reported data in at least 1 year. However, the number of plans in the sample varied by year (Table 1), ranging from 357 in 1998 to 290 in 2002. About 120 plans reported data in all 5 years.

Variables

We use the six HEDIS measures for childhood immunization, which include the Diphtheria, Tetanus, and Pertussis (DTP) rate; Measles, Mumps, and Rubella (MMR) rate; oral polio rate; *Haemophilus influenzae* type B (HIB) rate; Hepatitis B rate; and the Chicken Pox (VZV) rate. We note that all measures are highly correlated with each other (in general between 0.6 and 0.9, with the exception of Chicken Pox) both within and across periods. Table 1 presents descriptive statistics on these HEDIS measures illustrating that the mean of these measures shows a slight increase, even though there is still room for improvement on all six measures of childhood immunization. Our parametric models incorporate several control variables capturing characteristics of the health plans and of the markets in which the plans operate. These data come from several sources including NCQA (whether the plan gives permission to allow NCQA to publicly report the data, whether plans use the administrative [versus Hybrid] data collection method), Interstudy (plan age, plan profit status, HMO model type, Metropolitan Statistical Areas [MSA] HMO penetration, and HMO competition based on a Herfindahl index), and the Area Resource File (the percentage of the MSA population that is nonwhite and MSA per capita income). The administrative approach is based solely on claims data. The hybrid approach is based on a review of a sample of medical records and is considered to be more accurate because it supplements administrative data with clinical record review. The Herfindahl index is a widely used measure of competition among firms and is calculated as the sum of the squared market shares of all competitors (Scanlon et al. [2006a, b]). We compute our Herfindahl index for each MSA using HMO market shares based on enrollment. Because HMOs serve multiple MSAs, all MSA-level covariates are based on a weighted average of the values for the MSAs served by the plan, where the weights are based on the share of the plan's enrollment in each MSA (Scanlon et al. [2006a, b]). On average, plans have about 183,000 enrollees (excluding those with Medicare or Medicaid coverage because our HEDIS measures reflect commercial enrollment).

Table 1: Descriptive Statistics for Childhood Immunization Quality Measures

Childhood Immunization Measures	1998 (N = 357)		1999 (N = 334)		2000 (N = 329)		2001 (N = 305)		2002 (N = 290)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Childhood immunizations—DTP rate	75.93	14.93	78.87	12.95	80.79	12.07	81.68	10.77	80.38	11.91
Childhood immunizations—MMR rate	85.59	9.87	87.09	8.70	88.77	6.66	89.57	6.41	90.24	5.91
Childhood immunizations—OPV rate	81.62	14.10	82.83	11.79	84.65	11.22	85.56	9.68	86.35	10.97
Childhood immunizations—HIB rate	78.14	13.77	80.82	11.67	82.95	10.29	83.45	9.84	83.52	10.67
Childhood immunizations—Hepatitis B rate	71.32	17.28	75.34	14.29	78.04	13.81	80.19	11.82	82.17	12.63
Childhood immunizations—VZV rate	52.06	11.72	64.01	10.85	70.76	9.43	75.29	9.21	82.37	7.94

The sample size *N* represents the number of health plans used in analysis in each particular year. There are a total of 457 plans used in estimating the Multiple Indicator Multiple Cause (MIMIC) model. There are varying numbers of plans in each year. In general, we see that the mean of immunization scores increases over time while the variance decreases.

SD, standard deviation; DTP, Diphtheria, Tetanus, and Pertussis; MMR, Measles, Mumps, and Rubella; OPV, oral polio vaccine; HIB, *Haemophilus influenzae* type B; VZV, Chicken Pox.

MODEL

We estimate a longitudinal Multiple Indicator Multiple Cause (MIMIC) model of performance. The observed performance measures are assumed to be linear functions of an unobserved latent variable, and measurement error. We posit that variation in true performance is determined by time-varying and time-invariant covariates representing measured heterogeneity, and by stochastic error components representing unmeasured heterogeneity.

Specifically, our model is formulated as follows. Let Y_{jit} denote the j indicators of childhood immunization for plan i in period t , where $j = 6$ and $t = 5$ in our application. We assume that the indicators are a function of true performance (q) but are also affected by random variation that in part may include measurement error denoted by u . The reported HEDIS scores (Y 's) are derived using either the administrative or hybrid method. Because reporting method (hybrid versus administrative) may affect the values of the indicators (Y), but should not affect true performance (q), we can write the model as:

$$Y_{jit} = \lambda_{0jt} + \gamma \text{Hybrid}_{jit} + \lambda_{1j} q_{it} + u_{jit} \quad (1)$$

The vector λ_{0jt} consists of indicator-period-specific constants that represent means of the indicators in each period $t = 1998, 1999, 2000, 2001$, and 2002 . The vector λ_{1j} comprises HEDIS-specific factor loadings that measure the strength of the relationship between the latent variable and the HEDIS measures. Equation (1) also shows how our empirical model allows us to aggregate individual measures, thereby allowing us to estimate transition probabilities based on the latent variable q . We posit that the variation in the observed indicators (Y) is generated by variation in q and variation in u . If u represents measurement error, then the relevant part of the variation in Y is captured by variation in q . Note that q (unlike Y) is not subscripted by j (individual measures).

We assume that the latent variable is a function of covariates, X_{it} :

$$q_{it} = \beta_1 X_{it} + v_{it} \quad (2)$$

X_{it} includes plan and market characteristics, e.g., age of the plan, profit status, HMO penetration in the market, HMO competition, etc. In addition to the effect of the measured covariates, we posit that several unmeasured factors affect performance. These factors are denoted by v_{it} . For example, several potentially relevant variables such as contracted physicians' performance or the priority the plan places on quality improvement are not available in our

data and represent unmeasured heterogeneity across plans that might affect both the level and the growth in performance.

The unmeasured component v_{it} is specified as

$$v_{it} = \delta_i + (\eta_i \times t) + \varepsilon_{it} \quad (3)$$

The random component δ_i represents unmeasured plan-specific and time-invariant factors that affect levels of performance, while the term $(\eta_i \times t)$ represents unmeasured plan-level heterogeneity in growth rates. We allow for correlation in the components δ_i and η_i , which we assume are jointly normally distributed.

In addition, ε_{it} represents transitory shocks in each period that might have spillover effects in the periods that follow. For example, a new Surgeon General's report in year (t) that highlights the importance of childhood immunization might improve the HMO scores on childhood immunization in year (t) and in subsequent years, but the effect of this "shock" might wane over time. We let ε_{it} denote the shock to performance induced by the warning and assume that it is a normally distributed AR(1) variable.

Given that we assume all of the stochastic components of the model are normally distributed, it follows that q_{it} is also distributed normally. In particular, $q_{it} \sim N(\beta \bar{X}_t, \Sigma_{vv})$, where $\bar{X}_t$ is the mean of vector of variables X in period t and Σ_v is the variance-covariance matrix of v_{it} . The six measures and five time periods provide sufficient information to identify and estimate the coefficients and parameters of the variance-covariance matrix. As noted earlier, we use data on all 457 plans that reported at least 1 year of data. However, identification of the plan-level heterogeneity components and the autocorrelation parameter use information on plans that were in our sample for two or more years. The model parameters, including the β 's and all the components of Σ_{vv} are estimated via maximum likelihood.

Computing the Transition Probabilities

Once we know the mean and variance-covariance matrix of q_{it} we completely define the distribution of q . All the required transition probabilities can be computed from this distribution. Without loss of generality, we classify plans into three different tiers of performance (high, middle, and low). We use the distribution of plans in 2000 to define the two thresholds that determine whether a plan is in the high, middle, or low tier. Specifically, the upper threshold is the mean of plan latent performance (q) in 2000 plus the standard deviation of plan latent performance in 2000. Analogously, the lower threshold is the mean of plan latent performance (q) in 2000 minus the standard deviation of plan latent

performance in 2000. Once we have estimated the parameters of the model, we can compute the probability of whether a plan will be in the upper, middle, or lower tier of performance conditional on any pattern of past performance.

We note that the transition probabilities depend only on the absolute performance of the plan (i.e., own HEDIS score) rather than being dependent on the scores of other plans. An alternative approach that redefined thresholds each year could be developed, but such an approach may be more difficult to interpret by consumers because the ratings would reflect, in part, the performance of other plans in the sample. This could result in a change in the plans' rating despite no change in the plan performance.

Because some of our transition probabilities involve the evaluation of multivariate normal integrals, we use the Geweke, Hajivassiliou, and Keane simulator (Geweke, Keane, and Runkle 1994) to compute the probabilities (Appendix SA2, available on the web). Our model includes both time-invariant and time-varying measured and unmeasured variables in the probability statements, and thus the estimated transition probabilities can vary across plans and over time as well. We outline the procedure we use to compute the transition probabilities, but the details are provided in an appendix to this paper (Appendix SA2).

RESULTS

We begin this section with a description of the results when treating each measure of immunization as a separate indicator of performance. We do so because it is easy to understand the key features of our model and its potential usefulness when the goal is to estimate transition probabilities at the plan level by aggregating across individual performance measures.

For expositional purposes, let us focus on estimating the probability that a plan is in the top tier in 2002 given that it is in the middle tier in 2001. There are at least two ways to calculate this probability. The simpler, nonparametric approach is described first. From our earlier discussion on classification of plans into upper, middle, and lower threshold, we find that the upper threshold for DTP is $80.79 + 12.07 = 92.86$ and the lower threshold is $80.79 - 12.07 = 68.72$. By definition, this value is constant across years. Using these thresholds, the nonparametric estimate of the transition probability is simply the share of plans with HEDIS score on DTP > 92.86 in 2002, conditional only on those plans that had DTP scores between > 68.72 and 92.86 in 2001. We repeat this procedure for each of the six measures of childhood immunization. The transition probabilities (and bootstrapped standard errors) are presented in Table 2 (column 1).

Table 2: Observed and Predicted Transition Probabilities from the Nonparametric and Parametric Models

<i>Middle-Tier in 2001</i> <i>Individual Childhood Immunization Measures</i>	<i>Probability That Plan Is in Top-Tier in 2002</i>	
	<i>1 (Nonparametric Model)</i>	<i>2 (Parametric Model)</i>
Childhood immunizations		
DTP rate	0.015 (0.005–0.040)	0.073 (0.068–0.078)
MMR rate	0.050 (0.023–0.08)	0.102 (0.07–0.138)
OPV rate	0.090 (0.06–0.133)	0.108 (0.103–0.115)
HIB rate	0.052 (0.028–0.085)	0.091 (0.087–0.097)
Hepatitis B rate	0.118 (0.085–0.175)	0.135 (0.128–0.143)
VZV rate	0.608 (0.534–0.682)	0.571 (0.568–0.573)
Transition probability based on latent variable q		
Latent performance	N/A*	0.076 (0.070–0.081)

95% confidence intervals in parenthesis. Estimated standard errors in (parentheses) calculated using the bootstrap method based on 1,000 replications. Standard errors are based on the percentile method and are not based on the normal distribution. Plots of the empirical distribution of the probabilities showed a very long right tail and hence did not follow a normal distribution. The confidence intervals do not reflect uncertainty introduced in the computed probabilities because they are a function of estimated parameters. Thus, e.g., the probabilities are a function of estimated β 's—the effect of the covariates on the outcomes. The standard errors on β are not taken into account in estimating the confidence intervals shown in the table.

*As noted earlier, the main advantage of the parametric model is that it allows us to compute the transition probability based on the single latent variable.

DTP, Diphtheria, Tetanus, and Pertussis; MMR, Measles, Mumps, and Rubella; OPV, Oral Polio Vaccine; HIB, *Haemophilus influenzae* type B; VZV, Chicken Pox.

The parametric approach involves imposing the assumption of joint normality on the distribution of data in 2001 and 2002. For each of the six measures, we estimate the mean, variance, and covariance in the measures between the years 2001 and 2002. Given the assumption of normality and the thresholds defined in the preceding paragraph, we compute the transition probability that plans are in the upper tier in 2002 given that they are in the middle tier in 2001. These estimates (and bootstrapped standard errors) are presented in Table 2 (column 2). Appendix SA2 and SA3 (which are available on the web) provide further details on our methodology used to compute the parametric transition probabilities.

We note that the nonparametric (and parametric) estimates vary across HEDIS measures. For example, the nonparametric estimates suggest that the transition probabilities vary from 0.015 for DTP immunization to 0.61 for VZV for 1999 to 2001. We also note that the differences between the parametric and nonparametric estimates are due to departures from the parametric assumptions such as multivariate normality of the observed measures, AR(1) process on the transitory shocks, etc. The primary value of the parametric model

is that it aggregates results for many measures. A measure of the transition probability based on the latent variable is presented in the last row of Table 2. We note that in the absence of a parametric model to aggregate the measures, the decision maker is left with basing his decision on six transition probabilities, one for each childhood immunization measure. On the other hand, when we use our parametric model, we need to look at only the one transition probability (last row of Table 2). We find that the transition probability of a plan is in the top tier of performance in 2002 given that it is in the middle tier in 2001, and is $= 0.76$. It is evident from Table 2 that the parametric model gives less weight to the one large transition probability (about 0.6 for Chicken Pox). This is because the parametric model that links the observed measures to the latent variable q gives less weight to the measures that are not very highly correlated with the other measures. We note here that the correlation between Chicken Pox and the other measures is roughly between 0.30 and 0.45.

Aggregate Transition Probabilities

Table 3 presents results from both one- and two-step transition probabilities for more complete permutations of rankings. All of these transition probabilities are based on the latent variable q . Because we are interested in the transition probabilities, we do not discuss the specific parameter estimates for the plan and market characteristics included in Equation (2), but those estimates are consistent with prior research (Scanlon et al. 2005). Instead, we focus our discussion on the estimated aggregate transition probabilities from the MIMIC model, which are presented in Table 3.

The results in Table 3 suggest that conditional on being a plan in the upper tier in 1999, approximately 62 percent remained in the upper tier in 2000 while 47 percent of the plans continued to be rated in the upper tier of performance in 2001. Similarly, conditional on being a plan in the middle tier in 1999, about 11 percent of plans are in the upper tier in the year 2000. Multiple years of good prior performance increases the probability of good performance in the future. Approximately three-fourths of the health plans that were in the upper tier in both year $t-1$ and year t (1998 and 1999, respectively) continue to remain in the upper tier in 2000, while about 58 percent of the plans rated in the upper tier 2 years later (in 2001).

As expected, better performance in the more recent past is more indicative of better performance in the future than better performance in the more distant past. In particular, the probability of being in the upper tier in the year 2000 given that a plan is in the upper tier in 1999 and the middle tier in 1998 is

Table 3: Aggregate Transition Probabilities Based on Latent Variable *q*

		<i>Probability of Being in Upper Tier in Year</i>	
		<i>t</i> + 1 (2000)	<i>t</i> + 2 (2001)
Results using only 1 year of past performance			
	Rating in <i>t</i> (1999)		
NA	Upper tier	0.6230 (0.6120–0.6340)	0.4723 (0.4636–0.4813)
NA	Middle tier	0.1097 (0.1030–0.1158)	0.1186 (0.1140–0.1234)
NA	Lower tier	0.0008 (0.0007–0.0009)	0.0078 (0.0075–0.0081)
Results using 2 years of past performance			
Rating in <i>t</i> – 1 (1998)	Rating in <i>t</i>		
Upper tier	Upper tier	0.7507 (0.7252–0.7746)	0.5801 (0.5320–0.6286)
Middle tier	Upper tier	0.6669 (0.6190–0.7263)	0.4302 (0.3860–0.4803)
Lower tier	Upper tier	0.4162 (0.3701–0.4734)	0.4134 (0.3683–0.4600)
Upper tier	Middle tier	0.2319 (0.2190–0.2447)	0.2027 (0.1914–0.2138)
Middle tier	Middle tier	0.1003 (0.0944–0.1071)	0.1129 (0.1083–0.1187)
Lower tier	Middle tier	0.0196 (0.0175–0.0223)	0.0499 (0.0459–0.0545)
Upper tier	Lower tier	0.0020 (0.0017–0.0023)	0.0128 (0.0114–0.0143)
Middle tier	Lower tier	0.0015 (0.0013–0.0017)	0.0101 (0.0091–0.0111)
Lower tier	Lower tier	0.0007 (0.0006–0.0008)	0.0073 (0.0068–0.0077)

95% confidence intervals in parenthesis. Estimated standard errors in (parentheses) calculated using the bootstrap based on 1,000 replications. We use the percentile-based method to calculate the bootstrapped standard errors. Plots of the empirical distribution of the probabilities showed a very long right tail and hence did not follow a normal distribution. Hence, asymptotic standard errors are incorrect, and also give us negative values on confidence intervals because some of the probabilities shown in Table 3 are very close to zero. The confidence intervals do not reflect uncertainty introduced in the computed probabilities because they are a function of estimated parameters. Thus, e.g., the probabilities are a function of estimated β 's—the effect of the covariates on the outcomes. The standard errors on β are not taken into account in estimating the confidence intervals shown in the table.

equal to 0.67. In contrast, the probability of being in the upper tier in the year 2000 given that a plan is in the middle tier in 1999 and the in the upper tier in 1998 is only 0.23. This phenomenon reflects the presence of autocorrelation that results in plan performance in adjacent years to be highly correlated. We note that the confidence intervals on the transition probabilities are fairly tight.

DISCUSSION

In this paper, we estimate the stability of health plan performance using the domain of childhood immunization as an empirical example. Our model aggregates several immunization measures and relates performance and the stability of performance to plan covariates. Aggregation is important because for many purposes aggregated measures are needed to facilitate decision

making. The estimated parameters are used to simulate changes in transition probabilities for specified paths of time-varying covariates.

The results suggest moderate stability in plan performance. In particular, from the subset of plans in the upper tier of performance in 1999, about two-thirds continued to perform in the upper tier in 2000. Further, from the subset of plans in the upper tier in both 1998 and 1999, about three-fourths of the plans continued to perform in the upper tier in the year 2000. Thus, transition probabilities depend on past history over multiple periods. The probability of being in the upper tier in year 2000 depends on the state in year 1999, 1998, and possibly earlier states as well. However, because of the presence of autocorrelation, ratings in the more recent past are more important than ratings in the distant past. The presence of autocorrelation introduces stickiness in plan performance, probably attributable to unmeasured attributes of a health plan, such as plan management, physician network, etc., which are persistent or change slowly over time. While there is moderate stability, performance is by no means a permanent feature of health plans. For example, more than 25 percent of the plans cease to classify as upper-tier plans in 2000 despite being in the upper tier in 1998 and 1999. We conjecture a couple of reasons for this phenomenon. First, the HEDIS scores are calculated from data on a sample of health plan members. It is possible that in any given year, a plan faces a particularly bad draw of such members such that immunization rates are low. Second, we compute that more than 15 percent of the variation in performance q is driven by autocorrelation in transitory shocks. The autocorrelation parameter, ρ , that we estimate is < 1 so that transitory shocks to performance have muted effects over time. Thus, there is a mean reversion that could explain the fact that 25 percent of plans cease to remain in the upper tier in the year 2000 despite being in the upper tier in 1998 and 1999.

Taken together, these results suggest that consumers and purchasers should make use of as much past information as they can in making informed decisions about enrollment. If plans are good performers over multiple-time periods, there is a strong likelihood that they will be good performers in the future as well. However, good performance in just the one prior period does not strongly indicate good performance in the future. This suggests that report cards might want to include information on past performance.

Our analysis is subject to several limitations. First, the number of plans reporting performance data on childhood immunization is not constant over time and varies from 357 in 1998 to 290 in 2002. Most of these changes reflect mergers, acquisitions, and to a lesser extent, decisions about reporting to NCQA. Nonrandom decisions to report (e.g., only good-quality plans report HEDIS data to the NCQA) could influence the analysis; however, we consider

this potential to be relatively low because the share of plans reporting to the NCQA was reasonably stable over our study period (81 percent in 1998 and 82 percent in 2002, with small variation in between). In particular, although plans with poor performance can report to NCQA and request that their results are not made available publicly (McCormick et al. 2002), we had access to the results for both publicly reporting plans and nonpublicly reporting plans, and hence our analysis is less likely to be affected by selection bias.

A second limitation is that we only examine childhood immunization. Although important, it is only one of the many dimensions of plan performance measured by current reporting systems. Other work in the area, such as Luft and Romano (1993) use only one domain of performance (and in fact typically use only one measure within that domain, avoiding the aggregation issue our model addresses). Nevertheless, persistence may vary by other clinical areas and thus the conclusions regarding the value of the parametric model may depend on the clinical area chosen. Third, we only used data for 5 years and we focused only on 2-year lags. A longer panel would allow us to examine the role of more experience in predicting future performance. Fourth, all estimates derived in our model are based on the assumption of multivariate normality of the HEDIS indicators. In reality there are departures from normality. For example, a few HEDIS indicators have a long left tail. This departure from normality is one of the reasons for the difference between the observed and the predicted transition probabilities. However, those differences are modest, which bolsters our belief that any departures from normality are not severe enough to cause large biases. Fifth, although we focus on transition probabilities based on absolute performance, we acknowledge that relative performance, where relative is based on current levels of performance, may also be important for decision making. We note that our approach is flexible to accommodate an analysis of transition probabilities based on relative performance as well.

Finally, NCQA gave plans the option of "rotating" these child immunization measures in 2001. Specifically, to reduce the cost of data collection, plans were given the option of reporting the childhood immunization HEDIS scores from the year 2000 again in 2001. About 30 percent of the plans chose to exercise this option. Because of the possibility that the plans that choose to rotate data might have lower growth rates than those that chose to collect new data are the higher growth plans, we estimated the average growth rates (using the time points 1998 and 2002) of plans that chose to rotate data and those that did not rotate data. We did not find statistically significant differences in growth rates between the two groups of health plans. This finding suggests that potential nonrandom decisions to rotate measures do not substantially bias our results.

Despite these limitations, our results are important. The systems for generating ratings and reporting performance data are continuing to evolve, consuming considerable investment at the plan and purchaser levels. However, these systems are weakened by the inherent lags between data generation and plan choice. Our model provides a basis for beginning to think about how big a problem the time lag is, and contributes to approaches that provide more relevant data for consumers, purchasers, and policy makers.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project was supported by a grant from the Agency for Healthcare Research and Quality (AHRQ)—P01-HS10771. Dennis Scanlon acknowledges support from the Robert Wood Johnson Foundation Investigators in Health Policy Research Program. The authors thank Woolton Lee for research assistance. The authors are also grateful for comments received from Greg Pawlson, Michael Morrissey, Sarah Shih, and Sara Scholle. Prior versions of this paper were presented at the 2006 Academy Health Annual Research Meeting in Seattle and at the Health Policy Seminar Series at the University of Alabama, Birmingham.

This research was supported by funding from the Lister Hill Center for Health Policy and by a grant from the Agency for Healthcare Research and Quality (AHRQ) (grant # P01-HS10771). Earlier versions of this paper were presented at the Health Policy session of the American Statistical Association (Chicago, November 2003), the Econometrics Workshop at the University of Southern California (April 2005), and the Academy Health Annual Meeting (Seattle, June 2006). We thank the NCQA for providing the data and for thoughtful comments and suggestions. We also thank Woolton Lee for research assistance.

Disclosures: No conflicts of interests.

Disclaimers: None noted.

REFERENCES

- Geweke, J., M. Keane, and D. Runkle. 1994. "Alternative Computational Approaches to Inference in the Multinomial Probit Model." *Review of Economics and Statistics* 76 (4): 609–32.
- Gottschalk, P. 1982. "Earnings Mobility: Permanent Change or Transitory Fluctuations?" *Review of Economics and Statistics* 64 (3): 450–6.

- Landrum, M. B., S. E. Bronskill, and S. L. Normand. 2000. "Analytical Methods for Constructing Cross-Sectional Profiles of Health Care Providers." *Health Services Research and Outcomes Methodology* 1: 23–48.
- Lillard, L., and R. Willis. 1978. "Dynamic Aspects of Earnings Mobility." *Econometrica* 46 (5): 985–1012.
- Luft, H. S., and P. S. Romano. 1993. "Chance Continuity and Changes in Hospital Mortality Rates." *Journal of the American Medical Association* 270 (3): 331–7.
- McCormick, D., D. Himmelstein, S. Woolhandler, S. Wolfe, and D. Bor. 2002. "Relationship between Low Quality-of-Care Scores and HMOs Subsequent Public Disclosure of Quality-of-Care Scores." *Journal of the American Medical Association* 288 (12): 1484–90.
- National Institute of Health (NIH) 2003. "Understanding Vaccines: What They Are, How They Work." U.S. Department of Health and Human Services, National Institute of Allergy and Infectious Diseases, National Institute of Health (NIH) Publication # 03-4219.
- National Vaccine Advisory Committee (NVAC) 1999. "Strategies to Sustain Success in Childhood Immunizations." *Journal of the American Medical Association* 281 (4): 363–70.
- Roy, J., and X. Lin. 2002. "Nonignorable Dropouts and Missing Covariates: Changes in Methadone Treatment Practices." *Journal of the American Statistical Association* 97 (457): 40–52.
- Scanlon, D. P., M. Chernew, S. Sheffler, and A. M. Fendrick. 1998. "Health Plan Report Cards: Exploring Differences in Plan Ratings." *Journal on Quality Improvement* 24 (1): 5–20.
- Scanlon, D. P., C. Darby, E. Rolph, and H. E. Doty. 2001. "The Role of Performance Measures for Improving Quality in Managed Care Organizations." *Health Services Research* 36 (3): 619–41.
- Scanlon, D. P., S. Swaminathan, M. Chernew, J. Bost, and J. Shevock. 2005. "Health Plan Performance: Evidence from Managed Care Insurance Markets." *Medical Care* 43 (4): 1–9.
- Scanlon, D. P., S. Swaminathan, M. Chernew, and W. Lee. 2006a. "Competition in Health Insurance Markets: Limitations of Current Measures for Policy Analysis." *Medical Care Research and Review* 63 (6, suppl): 37–55S.
- . 2006b. "Market and Plan Characteristics Related to HMO Quality and Improvement." *Medical Care Research and Review* 12, 63: 56–89S.
- Staiger, D., and M. McClellan. 2000. "Comparing the Quality of Healthcare Providers." In *Frontiers in Health Policy Research*, Vol. 3, edited by A. Garber, pp. 113–36. Cambridge, MA: The MIT Press.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.

Appendix SA2: Estimation of the Transition Probabilities and the Associated Standard Errors Using the Geweke, Hajivassiliou, and Keane (GHK) Simulator.

Appendix SA3: Procedure to Calculate Model-Based Transition Probabilities at the Indicator Level.

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