Diabetes Healthcare Quality Report Cards: How Accurate Are the Grades?

Leonard Pogach, MD, MBA; Minge Xie, PhD; Yue Shentue, BA; Chin-Lin Tseng, DrPH; Miriam Maney, MA; Mangala Rajan, MBA; Anjali Tiwari, MD; John Kolassa, PhD; Drew Helmer, MD; Stephen Crystal, PhD; and Monika Safford, MD

Objective: To evaluate the accuracy and precision of random sampling in identifying healthcare system outliers in diabetes performance measures.

Study Design: Cross-sectional analysis of 79 Veterans Health Administration facilities serving 250 317 patients with diabetes mellitus between October 1, 1999, and September 30, 2000.

Methods: Primary outcome measures were poor glycosylated hemoglobin (A1C) control and good low-density lipoprotein cholesterol (LDL-C) and blood pressure (BP) control. Facility performance for each measure was calculated using 150 separate random samples and was compared with results using the bootstrap method as the criterion standard for determining outlier status (defined as a ≥ 5% difference from the mean, within the 10th or 90th percentile, or ≥ 2 SDs from the mean).

Results: The study population was largely male (97.4%), with 54.0% of subjects being 65 years or older. The facility-level mean performances were 22.8% for poor A1C control, 53.1% for good LDL-C control, and 55.3% for good BP control. Comparing the random sampling method with the bootstrap method, the sensitivity ranged between 0.64 and 0.83 for the 3 outcome measures, positive predictive values ranged between 0.55 and 0.88, and specificity and negative predictive values ranged between 0.88 and 0.99.

Conclusions: The specificity and negative predictive value of the random sampling method in identifying nonoutliers in performance were generally high, while its sensitivity and positive predictive value were moderate. The use of random sampling to determine performance for individual outcome measures may be most appropriate for internal quality improvement rather than for public reporting.

(Am J Manag Care. 2005;11:797-804)

The Institute of Medicine[1] suggested that a key feature of information dissemination to stakeholders is the use of industry benchmarks based on the "evaluability principle."[2] In this conceptual model, information is most likely to be used by all stakeholders if it can distinguish between best and worst performers. Furthermore, it was recommended that findings would be most useful if their robustness could be evaluated by using a wide range of assumptions or methods. Despite these recommendations, hospitals, managed care health plans, states, and federal agencies are increasingly labeled as best or worst performing units based on a variety of publicly reported performance measures, despite concerns about the accuracy of "report cards."[3-6]

Random sampling strategies are used to generate data for report cards and are dependent on the comparisons that are to be made.[1] For example, if the objective of a report card is to provide a broad overview of quality, simple random sampling might suffice, whereas stratified random sampling might be necessary to compare quality of care among multiple patient subgroups.

The National Committee for Quality Assurance (NCQA) is a leader in measuring the quality of ambulatory care in the United States.[7] Managed care health plans are evaluated on sets of composite measures to determine comparative scoring on a variety of general prevention and disease-specific indicators as part of a rigorous accreditation process, with results made available on a public basis as part of the NCQA Quality Compass report. The NCQA recommends a simple random sampling strategy for data collection within separate product lines of healthcare plans. However, in contrast to physician report cards,[8,9] the statistical certainty with which managed care health plans can be identified as outliers on performance of individual ambulatory care measures is not as well studied. We reasoned that the single random sampling method used to determine the outlier status of a plan might not necessarily reflect the overall composition of the population.

From the Center for Healthcare Knowledge Management, VA New Jersey Health Care System, East Orange (LP, CLT, MM, MR, AT, DH), University of Medicine and Dentistry of New Jersey–New Jersey Medical School, Newark (LP, CLT, DH), and Rutgers University, New Brunswick (MX, YS, JK, SC); and the Deep South Center on Effectiveness at the Birmingham VA Medical Center and University of Alabama at Birmingham, Birmingham (MS).

This study was supported by grant IIR 00-072-1 from the Department of Veterans Affairs Health Services Research and Development Service, Washington, DC (LP); and by grant NSF SES-0241859 from the National Science Foundation, Arlington, Va (MX). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs and NSF.

Address correspondence to: Leonard Pogach, MD, MBA, Center for Healthcare Knowledge Management, VA New Jersey Health Care System, 385 Tremont Avenue (111-Medicine), East Orange, NJ 07018. E-mail: leonard.pogach@med.va.gov.

The Veterans Health Administration (VA) is a nationwide integrated healthcare system with semiautonomous facilities that function in many ways like managed care health plans.[10] We used the VA's automated clinical, pharmacy, and laboratory data[11] to evaluate the statistical certainty of the current practice of using a single random sample to evaluate managed care health plan outliers in performance, defined as the percentage of the eligible population meeting the measure specifications. We constructed facility-level performance measures for glycemic, lipid, and blood pressure (BP) control for persons with diabetes mellitus using different dichotomous thresholds. To examine the accuracy of the random sampling approach (the ability of random samples to correctly identify the outlier status), we compared performance measures derived from a single random sample with those that were derived among all facility-level patients using a more rigorous statistical technique (the bootstrap method) as the criterion standard. We evaluated 3 different approaches to random sampling (unadjusted, age stratified, and age standardized). To examine the precision of the random sampling approach (the ability of random samples to repeatedly identify the outlier status), we redrew 150 separate random samples and examined the effect of age adjustment and measure thresholds on facility outlier classification.

METHODS

Data Sources and Patient Identification

Inpatient and outpatient utilization data and *International Classification of Diseases, Ninth Revision, Clinical Modification* codes were obtained from the National Patient Clinical Dataset (Austin, Tex), and medication data were provided by the VA's Healthcare Analysis Information Group.[12] We identified veterans as having diabetes mellitus in fiscal year 1999 if they had more than 1 diabetes diagnosis code (codes 250.xx, 357.2, 362.0, and 366.41) associated with separate calendar days or any diabetes-specific medication prescription (insulin, sulfonyureas, biguanides, alpha glucosidase inhibitors, and thiazolidinediones). We obtained utilization, pharmacy, laboratory, and death data on these veterans in fiscal year 2000 (October 1, 1999, through September 30, 2000).[13] To approximate the NCQA criteria for indemnity plan member inclusion in the Health Plan Employer Data and Information Set reporting, we included veterans who had at least 1 face-to-face visit with a clinician during the reporting period and were alive on the last day of the reporting period.[14] The VA New Jersey Health Care System Institutional Review Board approved this study.

Facility-level Inclusion and Exclusion Criteria

To minimize confounding of outlier status by variation in glycosylated hemoglobin (A1C) test results due to laboratory assay method variation, we restricted our analyses to those facilities using assays certified by the National Glycohemoglobin Standardization Program.[15] We identified 125 facilities, with 397 147 veterans who met our criteria for evaluation of diabetes care during the observation period. We eliminated 11 facilities with less than 1000 patients. We then excluded facilities that were more than 4 SDs from the means with respect to the proportion of A1C tests (25 facilities), low-density lipoprotein cholesterol (LDL-C) level (5 facilities), or BP level (5 facilities) to avoid identification of outliers that were likely due to incomplete data collection attributable to the semiautomated extraction programs that were used for administrative data collection, as previously reported.[16-18] These exclusion criteria reduced the number of facilities from 125 to 79. The final study sample comprised 250 317 patients, with a facility-level range of 1015 to 13 598 patients per facility.

Outcome Measures

We used the accountability performance measures for diabetes care, reflecting glycemic, lipid, and BP control based on those developed by the Diabetes Quality Improvement Project[19] and reported by the NCQA in the 2000 Diabetes Comprehensive Care measures.[20] Poor A1C control was defined as the proportion of the population with an A1C test not performed or with an A1C level of 9.5% or higher. Good LDL-C control was defined as the proportion of the population with an LDL-C test performed and with an LDL-C level less than 130 mg/dL (< 3.37 mmol/L). Good BP control was defined as the proportion of the population with BP less than 140/90 mm Hg. We also constructed performance measures using thresholds of 8.5% or higher for A1C level, less than 100 mg/dL (< 2.59 mmol/L) for LDL-C level, and less than 130/80 mm Hg for BP to more closely approximate professional society guideline recommendations during the study period.[21] All 6 performance measures were based on the last recorded values.

Sampling Methods

We followed the Health Plan Employer Data and Information Set guidelines for calculation and sampling to draw the random samples for this study.[21] Based on the binomial theorem, a sample size of 411 subjects provides 90% power for detecting differences when the mean percentage of patients meeting a measure is 50%, which is justified as a tradeoff between statistical certainty and the cost of obtaining data from medical record review. Therefore, we drew 150 separate single

random samples of 411 subjects from each facility among the entire facility population to be used for our primary analyses.

The NCQA requires commercial managed care health plans to collect and report results separately by populations covered by different product lines, including Medicare. Consequently, we used 2 other random sampling strategies that involved age adjustment (age-stratified random sampling and age-standardized random sampling) by randomly sampling within 2 age strata (≥ 65 vs < 65 years old). In the first strategy (age stratified), we sampled from the 2 age strata according to the proportions at the facility level to ensure that the proportions of patients in the age categories were the same as the proportions in that facility.[22] In the second strategy (age standardized), we sampled from the 2 age strata to ensure that the proportions of patients in the age categories were the same as the proportions in the entire study population.

To create a criterion standard that most likely represented a facility's true performance measure, we used the bootstrap method. Bootstrapping is a computer-intensive method that resamples from the original data to create a large number of bootstrap samples to cover various events of random chance. It has been proven to be an effective tool in statistics to estimate parameters and to assess variability.[23] In our context, 1000 bootstrap samples, each consisting of 411 randomly drawn subjects with replacement, were obtained for a given facility. The bootstrap mean was calculated for each of the 6 performance measures. Because bootstrap means are highly accurate assessments of the corresponding true performance measures, they were treated as if they were the true values. We repeated the same bootstrap method for the age-stratified and age-standardized sampling approaches.

Outlier Identification

The outlier status for each facility on each of the performance measures was determined using 3 separate strategies for each of the 150 individual random samples with and without age adjustment, as well as for the bootstrap method. In the first strategy, an individual facility was considered an outlier if its performance measure percentage was different by at least 5% from the performance mean across facilities, which is a strategy recommended by the NCQA.[21] In the second strategy, an individual facility was considered an outlier if its performance measure was within the 10th or 90th percentile of rank order (the best performers were in the 10th decile), which is used by the NCQA Quality Compass and by others.[5,6] In the third strategy, based on determination of the 95% confidence interval, an individual facility was considered an outlier if its performance measure was 2 or more SDs from the mean.

Evaluation of the Random Sampling Method

Within each random sampling strategy, the facility outlier status calculated from each of the 150 individual random samples was compared with the outlier status determined using the bootstrap method for each performance measure. We defined accuracy as the ability of a single random sample to correctly identify the bootstrap outlier status. For each random sampling strategy, we calculated 4 evaluation measures (sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]) for each performance measure for each of the 3 outlier definitions.

Sensitivity was defined as the probability of a facility being identified as an outlier by a random sampling method, given that it is an outlier facility (according to the bootstrap method). Specificity was defined as the probability of a facility being identified as not being an outlier, given that it is not an outlier facility. Positive predictive value was defined as the likelihood that a facility is an outlier using the bootstrapping method given that it was identified by random sampling. Negative predictive value was defined as the likelihood that a facility is not an outlier using the bootstrapping method given that it was not identified by random sampling. We determined the ranges of the sensitivity, specificity, PPV, and NPV of the random sample outlier status compared with the bootstrap outlier status using each of the 3 random sampling approaches and the 3 outlier approaches for each of the 6 performance measures.

We defined precision as the ability of the single random samples to repeatedly identify the outlier status as determined by the bootstrap method. Therefore, we reported the ranges of sensitivity, specificity, PPV, and NPV from the 150 separate samples for each random sampling approach for each measure.

RESULTS

Table 1 gives the distribution of patient characteristics and diabetes care measures by facility inclusion status. The study population was largely male (97.4%), with 54.0% of patients being 65 years or older. The mean ± SD A1C level was 7.6% ± 1.7%, the LDL-C level was 107.4 ± 32.6 mg/dL (2.78 ± 0.84 mmol/L), the systolic BP was 139.0 ± 21.0 mm Hg, and the diastolic BP was 73.7 ± 12.1 mm Hg. The patients from the excluded facilities had similar demographic characteristics and mean values for A1C level, LDL-C level, and BP, but as expected they had lower percentages of A1C and LDL-C tests performed.

Table 1. Characteristics of Patients With Diabetes Mellitus at 79 Veterans Health Administration Facilities, 1999-2000*

Characteristic	Included Facilities	Excluded Facilities
Patient		
Age, y	63.9 ± 11.1	63.8 ± 11.2
≥ 65	54.1	53.9
< 65	45.9	46.1
Male sex	97.4	97.5
Facility		
Patients, No.	250 317	146 830
Facilities, No.	79	46
Patients per facility [range]	3169 ± 1886 [1015-13 598]	3124 ± 2484 [41-10 718]
Glycosylated hemoglobin		
Test performed	88.7	81.6
Level, %	7.6 ± 1.7	7.6 ± 1.7
Low-density lipoprotein cholesterol		
Test performed	68.9	58.5
Level, mg/dL†	107.4 ± 32.6	106.4 ± 32.9
Blood pressure, mm Hg		
Systolic	139.0 ± 21.0	139.2 ± 21.8
< 140	50.9%	41.2%
< 130	31.3%	25.8%
Diastolic	73.7 ± 12.1	73.9 ± 12.3
< 90	88.3%	71.5%
< 80	65.4%	52.5%

*Data are given as means ± SDs or as percentages unless otherwise indicated.
†To convert cholesterol to millimoles per liter, multiply by 0.0259.

Facility-level performance was determined for individual diabetes care measures, presented as the overall facility mean and by percentile of rankings (**Table 2**). The mean facility rates of poor A1C control were 22.8% (using a ≥ 9.5% threshold) and 33.0% (using a ≥ 8.5% threshold). The mean facility rates of good LDL-C control were 53.1% (using a < 130 mg/dL [< 3.37 mmol/L] threshold) and 29.8% (using a <100 mg/dL [<2.59 mmol/L] threshold). The mean facility rates of good BP control were 55.3% (using a <140/90 mm Hg threshold) and 32.7% (using a <130/80 mm Hg threshold). There was substantial variation among the facilities, with 2-fold to almost 3-fold differences between the best and worst performers. For A1C level, the worst performing facility had 34.7% of its patients with diabetes mellitus with A1C levels of 9.5% or higher, and the best performing facility had just 13.2%. For LDL-C level, the worst performing facility had 32.3% of its patients with diabetes mellitus with LDL-C levels less than 130 mg/dL (< 3.37 mmol/L), and the best performing facility had 68.2%. For BP, the worst performing facility had 43.2% of its patients with diabetes mellitus with BPs less than 140/90 mm Hg, and the best performing facility had 67.8%. Variations were similarly large across the lower threshold performance measures. These results demonstrate a difference of 20 and 30 percentage points between the best and worst performing VA facilities for each measure and a difference of 15 to 20 percentage points between the facilities in the lowest and highest quartiles. These differences are indicative of clinically meaningful variation in care provided among outliers.

Based on the bootstrap criterion and using the decile method (whether a facility was within the 10th or the 90th percentile of rank order), there were 16 outlier facilities (20%).Using the outlier identification strategies of 5% or greater difference from the mean and 2 SDs from the mean, respectively, the numbers of outlier facilities were 4 (5%) and 6 (8%) for A1C level of 9.5% or higher, 9 (11%) and 7 (9%) for LDL-C level less than 130 mg/dL (< 3.37 mmol/L), and 29 (37%) and 3 (4%) for BP less than 140/90 mm Hg (Table 2).

The **Figure** shows the means and ranges of the sensitivity, specificity, PPV, and NPV for the random sampling method without age adjustment compared with the bootstrap method for A1C level of 9.5% or higher, LDL-C level less than 130 mg/dL (< 3.37 mmol/L), and BP less than 140/90 mm Hg for each of the outlier status methods (≥ 5% difference from the mean, ≥ 2 SDs from the mean, and the decile method). Overall, random sampling had fairly acceptable accuracy and precision for detecting nonoutlier status. Specificity was fairly high (range, 0.88-0.99 across measures), with a fairly narrow range on 150 repeated samples (usual range, 0.9-1.0 for most performance measures but as low as 0.63 for BP outlier status using ≥ 5% difference from the mean). Similarly, NPVs had uniformly high means

Table 2. Distribution of Facility-level Performance for Individual Diabetes Care Measures*

Measure	Facility Mean	Percentile				
		0 (Minimum)	25th	50 (Median)	75th	100th (Maximum)
Glycosylated hemoglobin level, %						
≥ 9.5	22.8	13.2	18.6	21.7	25.5	34.7
≥ 8.5	33.0	21.9	29.2	32.1	36.3	45.5
Low-density lipoprotein cholesterol level, mg/dL						
< 130	53.1	32.3	48.6	53.8	57.8	68.2
< 100	29.8	13.4	27.4	29.7	33.7	44.6
Blood pressure, mm Hg						
< 140/90	55.3	43.2	51.5	55.7	59.6	67.8
< 130/80	32.7	22.5	29.4	32.5	36.0	45.0

*Data are given as percentages among 250 317 persons with diabetes mellitus.

(range, 0.88-0.99) across performance measures, thresholds, and outlier determination strategies, with fairly narrow ranges on repeated sampling (usual range, 0.9-1.0).

However, the mean sensitivity and the PPV for the 150 random samples were significantly lower, with broader ranges for individual samples. The means for sensitivity varied between 0.73 and 0.80 (range, 0.20-1.00) for A1C level, between 0.64 and 0.83 (range, 0.30-1.00) for LDL-C level, and between 0.64 and 0.75 (range, 0.30-1.00) for BP. Similarly, the mean PPVs varied between 0.55 and 0.67 (range, 0.27-1.00) for A1C level, between 0.68 and 0.88 (range, 0.33-1.00) for LDL-C level, and between 0.77 and 0.86 (range, 0.33-1.00) for BP.

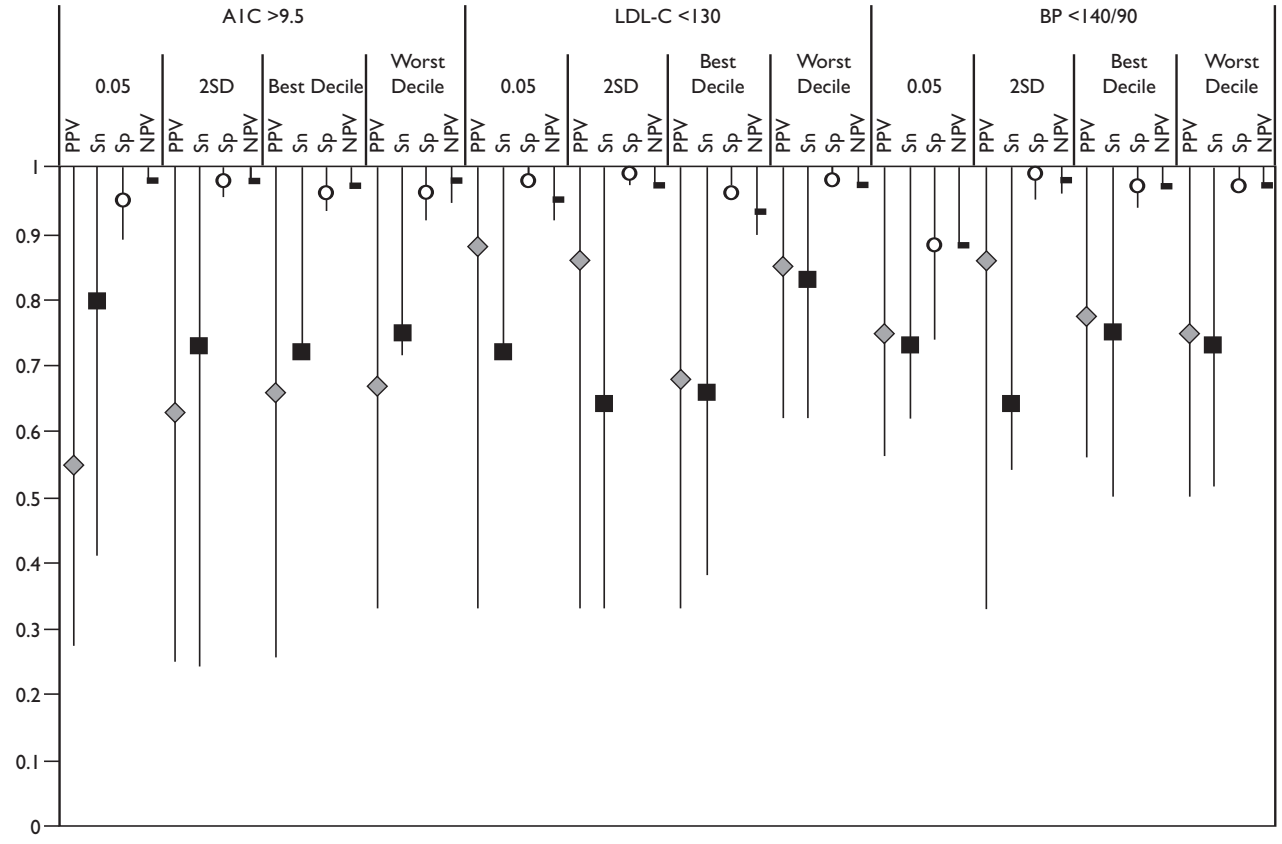
The low sensitivity and PPV were not a result of the method used to determine outlier status, although results varied slightly depending on the strategy used. For example, for A1C level of 9.5% or higher, the 5% or greater difference from the mean approach had the largest difference between sensitivity and PPV, and the best and worst decile method had similar sensitivity, specificity, PPV, and NPV. In contrast, for LDL-C level less than 130 mg/dL (<3.37 mmol/L), the best and worst decile method had different sensitivity, specificity, PPV, and NPV. For BP less than 140/90 mm Hg, the 5% or greater difference from the mean approach to outlier detection was the least favorable of the 3 approaches, and the decile method had similar results. Our results did not differ when age-stratified or age-standardized random sampling strategies were used or when lower performance measurement thresholds (≥ 8.5% for A1C level, <100 mg/dL [<2.59 mmol/L] for LDL-C level, and <130/80 mm Hg for BP) were used (data available from the author).

DISCUSSION

The specificity and NPV of the random sampling method in identifying nonoutliers in performance were generally higher, with narrow ranges, than its sensitivity and PPV. Therefore, the random sampling method was more accurate and precise in identifying facilities that were not outliers than in identifying facilities that were outliers with respect to performance on glycemic, cholesterol, and BP control. Neither different random sampling approaches that took age into account, nor lower thresholds for glycemic, cholesterol, and BP control, consistently improved the sensitivity or the PPV of the random sampling method in the identification of outliers. These findings were consistent across 3 different methods of outlier detection—identification of at least a 5% difference from the mean, at least 2 SDs from the mean, and ranking in the 10th or 90th percentile of facilities. The inconsistency of the random sampling approach in identifying outlier status compared with a criterion standard derived from the entire population contributes to a growing body of literature raising concerns about the increasingly widespread practice of ranking hospitals, health plans, and physician groups.

Our findings can perhaps be attributed to the possibility that several factors beyond the control of the healthcare plan are not randomly distributed among

Figure. Summary of Accuracy and Precision of Random Sampling Method



A1C indicates glycosolated hemoglobin; LDL-C, low-density lipoprotein cholesterol; BP, blood pressure.
Means are depicted by symbols. Each outlier method is presented for each measure. 0.05 indicates at least a 5% difference from the mean; 2SD, at least 2 SDs from the mean; and best and worst decile, facilities above the 10th percentile and below the 90th percentile of rank order, respectively. The mean and range (bar) of sensitivity (Sn ■), specificity (Sp O), positive predictive value (PPV ◆), and negative predictive value (NPV ▬) were derived from 150 random samples of 411 subjects and were compared with values derived using the bootstrap method as the criterion standard.

facilities and could result in biases among random samples using administrative data. For example, veteran clinical users may have enrollment in multiple systems of care,[24] accounting for discrepancies between medical record abstraction and administrative data.[17] However, care in multiple systems would not bias the findings that we observed with BP measurements, because all BP readings were based solely on VA data. Although age and race biases in the availability of administrative data have been reported in the private sector,[25] the automated laboratory databases used by the VA at the facility level render this an unlikely explanation for our findings. Furthermore, our findings were similar when we adjusted for age effects by using stratified sampling and age standardization. Patient-level factors can also contribute to facility variation in performance. For example, factors under provider management, such as shared decision making, can affect glycemic, lipid, and BP control.[26] However, factors extrinsic to a healthcare system, including patient age, duration of diabetes mellitus, and medical comorbidities account for as much as 10% of the variation in A1C levels[18,27,28] and may vary among facilities.

Therefore, it is possible that risk adjustment could improve the accuracy and precision of the random sampling method. However, risk-mix adjustment to facilitate comparison among healthcare plans has not yet been implemented for use in public reporting,[19,29] although it is recommended in the health services literature[30] and is used in a voluntary physician recognition program in which qualifying physicians may receive a bonus for each patient with diabetes mellitus covered by a participating employer or health plan.[31] The use of composite measures may also improve the PPV of random samples, but if there is a small difference among outliers, variations in patient mix among the random samples may still result in

large changes in outlier status based on an "all or none" threshold.

The effect of the accuracy and precision in outlier identification will relate to the context in which the information is to be used and in the consequences that could result from its use.[1] For example, if the objective is to evaluate nonoutliers with reasonable certainty for the purpose of focusing internal quality improvement strategies by identifying possible best practices,[32] then our results suggest that the use of single random samples is an appropriate statistical method based on its high specificity and NPV.

However, if the objective of report cards is to identify best and worst performing healthcare plans or large physician groups that could affect direct marketing to the public and influence consumer choice, then accrediting and payer systems may consider evaluating the accuracy and precision for the extreme outliers separately, providing confidence intervals for the results. Although we recognize the need for a balance between public reporting[32,33] and statistical uncertainty, recent public sector[34] and private sector[35] Web site reports do not present sufficient information to evaluate the statistical confidence with which outliers are ascertained.

Strengths of this study include the VA's standardized national data repositories for laboratory and clinical results, which permitted the construction of performance measures that identified "actual" outliers. Furthermore, the data sets and measures used herein have been studied for reliability[13,16,18,30] and had already been used at the time of the study.[19] We also used only National Glycohemoglobin Standardization Program–certified A1C test results to minimize variation attributable to A1C laboratory assays.[15]

We also recognize several limitations, including the facts that the VA population is homogeneous and that variations among VA facilities may be less than those among private sector plans.[36] However, greater variation in outlier determination may well be seen in settings with more heterogeneity.

In conclusion, the specificity and NPV of the random sampling method in identifying nonoutliers in performance on individual diabetes care measures were generally high and precise, while its sensitivity and PPV in identifying outliers were moderate, with less precision. We suggest that random sampling may not meet the level of certainty necessary for the objective of external reporting (to consumers and payers) as opposed to internal feedback (to plans and physicians) for individual diabetes care measures. Our findings add to the increasing debate about the effect of public reporting on consumer, physician, and health plan behavior.[37-39] Therefore, we recommend that issuers of report cards provide greater transparency in the certainty of identifying best and worst performing plans (or providers), consistent with recommendations from the Institute of Medicine.[9]

Acknowledgment

We thank Christina Croft, BA, for assistance in the preparation of the manuscript.

REFERENCES

1. Hurtado MP, Swift EK, Corrigan JM, eds. Designing the National Healthcare Quality Report. In: *Envisioning the National Healthcare Quality Report.* Washington, DC: Institute of Medicine, National Academy Press; 2001.

2. Hibbard J. Use of outcomes data by purchasers and consumers: new strategies and new dilemmas. *Int J Qual Health Care.* 1998;10:503-508.

3. Bentley JM, Nash DB. How Pennsylvania hospitals have responded to publicly released reports on coronary artery bypass graft surgery. *Jt Comm J Qual Improv.* 1998;24:40-49.

4. Jencks SF. Clinical performance. *JAMA.* 2000;283:2015-2016.

5. Lee TH, Meyer GS, Brennan TA. A middle ground on public accountability. *N Engl J Med.* 2004;350:2409-2412.

6. Marshall MN, Shekelle PG, Leatherman S, Brook RH. The public release of performance data: what do we expect to gain? a review of the evidence. *JAMA.* 2000;283:1866-1874.

7. Schneider EC, Riehl V, Courte-Wienecke S, Eddy DM, Sennett C. Enhancing performance measurement: NCQA's road map for a health information framework. *JAMA.* 1999;282:1184-1190.

8. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *JAMA.* 1999;281:2098-2105.

9. Hannan E, Sui AL, Kumar D, Kilburn H Jr, Chassin MR. The decline in coronary artery bypass graft surgery mortality in New York State: the role of surgeon volume. *JAMA.* 1995;273:209-213.

10. Kizer KW. The "new VA": a national laboratory for health care quality management. *Am J Med Qual.* 1999;14:3-20.

11. Krein SL, Vijan S, Pogach LM, Hogan M, Kerr EA. Aspirin use and counseling about aspirin among patients with diabetes. *Diabetes Care.* 2002;25:965-970.

12. Weinstock RS, Hawley G, Repke D, Feuerstein BL, Sawin CT, Pogach LM. Pharmacy costs and glycemic control in the Department of Veterans Affairs. *Diabetes Care.* 2004;27(suppl 2):B74-B81.

13. Miller DR, Safford MM, Pogach LM. Who has diabetes? Best estimates of diabetes prevalence in the Veterans Health Administration based on computerized patient data. *Diabetes Care.* 2004;27(suppl 2):B10-B21.

14. Jha AK, Perlin JB, Kizer KW, Dudley RA. Effect of the transformation of the Veterans Affairs Health Care System on the quality of care. *N Engl J Med.* 2003;348:2218-2227.

15. Little RR, Rohlfing CL, Wiedmeyer HM, Myers GL, Sacks DB, Goldstein DE; NGSP Steering Committee. The National Glycohemoglobin Standardization Program: a five-year progress report. *Clin Chem.* 2001;47:1985-1992.

16. Krein SL, Hofer TP, Kerr EA, Hayward RA. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. *Health Serv Res.* 2002;37:1159-1180.

17. Kerr EA, Smith DM, Hogan MM, et al. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *Jt Comm J Qual Improv.* 2002;28:555-565.

18. Safford M, Eaton L, Hawley G, et al. Disparities in use of lipid-lowering medications among people with type 2 diabetes mellitus. *Arch Intern Med.* 2003;163:922-928.

19. Fleming BB, Greenfield S, Engelgau MM, Pogach LM, Clauser SB, Parrott MA. The Diabetes Quality Improvement Project: moving science into health policy to gain an edge on the diabetes epidemic. *Diabetes Care.* 2001;24:1815-1820.

20. National Committee for Quality Assurance. *HEDIS 2001.* Vol 1. Washington, DC: National Committee for Quality Assurance; 2000.

21. American Diabetes Association. Standards of medical care for patients with diabetes mellitus. *Diabetes Care.* 2002;25(suppl 1):S33-S49.

22. Gordis L. *Epidemiology.* 2nd ed. Philadelphia, Pa: WB Saunders Co; 2000.

23. Efron B, Tibshirani R. *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall; 1993.

24. Shen Y, Hendricks A, Zhang S, Kazis LE. VHA enrollees' health care coverage and use of care. *Med Care Res Rev.* 2003;60:253-267.

25. Keating NL, Landrum MB, Landon BE, Ayanian JZ, Borbas C, Guadagnoli E. Measuring the quality of diabetes care using administrative data: is there bias? *Health Serv Res.* 2003;38:1529-1545.

26. Heisler M, Bouknight RR, Hayward RA, Smith DM, Kerr EA. The relative importance of physician communication, participatory decision making, and patient understanding in diabetes self-management. *J Gen Intern Med.* 2002;17:243-252.

27. Harris MI. Racial and ethnic differences in health care access and health outcomes for adults with type 2 diabetes. *Diabetes Care.* 2001;24:454-459.

28. Zhang Q, Safford M, Ottenweller J, et al. Performance status of health care facilities changes with risk adjustment of HbA_{1c}. *Diabetes Care.* 2000;23: 919-927.

29. Chin MH. Risk-adjusted quality of care rating for diabetes: ready for prime time? *Diabetes Care.* 2000;23:884-886.

30. Kerr EA, Gerzoff RB, Krein SL, et al. Diabetes care quality in the Veterans Affairs Health Care System and commercial managed care: the TRIAD study. *Ann Intern Med.* 2004;141:272-278.

31. National Committee for Quality Assurance Web site. Bridges to excellence: rewarding quality across the healthcare system. Available at: http://www.ncqa.org/Programs/bridgestoexcellence. Accessed July 16, 2005.

32. Landon BE, Normand SL, Blumenthal D, Daley J. Physician clinical performance assessment: prospects and barriers. *JAMA.* 2003;290:1183-1189.

33. Galvin RS, McGlynn EA. Using performance measurement to drive improvement: a road map for change. *Med Care.* 2003;41:I48-I60.

34. New Jersey Commercial Health Maintenance Organizations. A comprehensive performance report: year 2003. Available at: http://www.state.nj.us/health/hcsa/2003com_report.pdf. Accessed March 24, 2005.

35. National Committee for Quality Assurance Web site. NCQA report cards. Available at: http://hprc.ncqa.org/menu.asp. Accessed March 24, 2005.

36. Greenfield S, Kaplan SH. Creating a culture of quality: the remarkable transformation of the Department of Veterans Affairs Health Care System. *Ann Intern Med.* 2004;141:316-318.

37. Krumholz HM, Rathore SS, Chen J, Wang Y, Radford MJ. Evaluation of a consumer-oriented Internet health care report card: the risk of quality ratings based on mortality data. *JAMA.* 2002;287:1277-1287.

38. Dranove D, Kessler D, McClellan M, Satterthwaite M. Is more information better? the effects of "report cards" on health care providers. *J Polit Econ.* 2003;111:555-588.

39. McCormick D, Himmelstein DU, Woolhandler S, Wolfe SM, Bor DH. Relationship between low quality-of-care scores and HMOs' subsequent public disclosure of quality-of-care scores. *JAMA.* 2002;2882:1484-1490.