

Risk-Adjusted Quality of Care Ratings for Diabetes

Ready for prime time?

Interest is surging in publicly available quality-of-care ratings of hospitals, health plans, and individual providers (1,2). High-profile reports have been compiled by the National Committee on Quality Assurance's Health Plan Employer Data and Information Set (HEDIS) program, which evaluates managed care plans and other providers, state initiatives in New York and Pennsylvania that rate coronary artery bypass graft facilities, and *U.S. News and World Report's* annual hospital rankings. Diabetes is emerging as one of the key test cases for scorecard efforts, because the condition is common and costly and requires several readily measurable processes of care that are linked to improved health outcomes (3). The Diabetes Quality Improvement Project (DQIP), a joint venture of the American Diabetes Association and several national partners, created a common measurement set for performance evaluation (4). Many of the DQIP quality indicators are currently being included in HEDIS and the American Medical Accreditation Program (5).

Proponents of publicly available ratings argue that these data provide powerful incentives for providers to improve their quality of care in order to gain the business of larger purchasers and health plans. In addition, public ratings could specify those areas in which providers should target their quality improvement efforts, and report cards might also enable individual patients to make wiser decisions when they choose hospitals and insurance plans. While competition among providers has recently emphasized cost-cutting, the hope has been that improving the quality of care will become a more salient goal. Cultural and technological trends, such as the increased public demand for provider accountability and the ease of information dissemination through the internet, make it likely that public ratings of providers will continue to grow at a rapid pace.

As the push toward public scorecard systems for diabetes care gains momentum, several problems have arisen (6). One

of the most important concerns is the lack of case-mix adjustment among current report card systems. If, for example, one provider or medical center takes care of a particularly complicated group of diabetic patients with advanced disease who are difficult to treat, why should that provider be penalized if its patients' average HbA_{1c} value is higher than that of the typical provider's patients? DQIP attempted to avoid this problem with the HbA_{1c} value by creating a relatively lenient performance cutoff value of 9.5%. This decision has met significant criticism from some elements of the diabetes and provider communities who believe that this cutoff value, which was chosen for population-level performance measurement purposes, may inadvertently give providers the incorrect message that a HbA_{1c} value <9.5% is sufficient for the individual patient. However, without adequate case-mix adjustment, incentives are created for providers to avoid or dump the sickest and most complicated patients to improve their ratings (7). Unfortunately, no widely accepted diabetes case-mix adjustment tool that is applicable for practical policy use currently exists.

Therefore, Zhang et al. (8) make an important and timely contribution to this clinical and policy debate by developing a new case-mix adjustment tool for diabetes by use of a Veterans Administration (VA) database. Using demographic and clinical information available in this administrative dataset, they developed their tool and then analyzed how rankings of the quality of care among 48 VA facilities, based on patients' HbA_{1c} values, would change depending on whether the case-mix adjustment tool was used. For 2 performance measures, the mean HbA_{1c} value and the proportion of HbA_{1c} values exceeding 9.5%, they found that 15 and 25%, respectively, of facilities initially identified as outliers became average performers, or vice-versa, after risk adjustment. It is worth noting that the number of recategorized patients might be artificially high; Zhang et al. (8) defined a relatively low threshold for a facility to be an outlier. For

example, at baseline they labeled 30 of 48 (62.5%) VA facilities as outliers for the high-risk HbA_{1c} performance measure.

What, then, are the implications of this study for diabetes ratings systems? Has the time of adequate case-mix adjustment arrived? Should we scrap DQIP's HbA_{1c} cutoff value of 9.5% and use risk-adjusted absolute values? Are publicly available risk-adjusted scorecard systems for diabetes truly ready for prime time? The development and application of case-mix adjustment tools are extremely difficult methodological areas (7,9), and I urge caution before we claim victory in this area. I will outline the key elements of a diabetes case-mix adjustment instrument, assess where we are in the field, and suggest where we should go regarding both developing risk adjustment tools and using diabetes quality-of-care rating systems.

Theoretically, the fairest comparative report card system would rate providers based on how they cared for the exact same group of patients. Of course, in the real world, providers care for different groups of patients. Therefore, to make comparative ratings among different providers meaningful, ratings should be adjusted for factors affecting HbA_{1c} values beyond the provider's reasonable control.

What are the key elements of the ideal case-mix adjustment tool? First, the tool needs to have clinical face validity. Even if a case-mix model performed well from a statistical standpoint, buy-in from clinicians would be unlikely if they felt the model lacked key variables. Zhang et al.'s conceptualization of 6 fundamental domains (demographics, access to care, health care-seeking behavior, geographic location, duration and severity of diabetes, and comorbidity) is a good start to ensure incorporation of relevant variables. Elements of both clinical and social case-mix models are included. However, it is important to assess whether a given factor is truly legitimate to include in the case-mix model, or if adjusting for a particular factor wrongly pardons a difference in the quality of care

for which the provider should be accountable. For example, Zhang et al. adjust for rural geographic location. If the quality of care is lower in rural areas because of intractable access to care problems, this correction may be reasonable. However, if quality of care is lower because of suboptimal provider behavior in rural areas, then this adjustment is inappropriate. Similarly, we know that older African-Americans with diabetes are not as likely as Caucasians to receive several processes of diabetes care (10). If African-Americans expressed an informed preference for less aggressive care, then part of these differences may be justified (11). However, if this disparity is due to racial discrimination by providers or perhaps differences in how options are presented, then adjustment for race is inappropriate. Analogously, on the supply side, community health centers with relatively few resources that serve predominantly indigent poorly educated diabetic patients provide similar quality of care as wealthier providers who treat more privileged populations (12). Current case-mix adjustment tools do not take into account the limited resources that these community health centers have when establishing the ratings.

Severe diabetes can lead to higher HbA_{1c} values, but advanced diabetes could also be the result of poor quality care. Thus, issues of causality can be problematic when adjusting for the severity of diabetes. For example, if more severely ill diabetic patients were more likely to have HbA_{1c} values that are difficult to control, then it could be valid to adjust for severity of illness measures, such as hospitalization. However, if higher HbA_{1c} values due to poor quality care by the provider led to more severe diabetes, which in turn would lead to more complications and hospitalizations, then it would be wrong to adjust for severity of illness. A similar directionality problem can occur when operationalizing other elements of the case-mix model. For example, Zhang et al. use total pharmacy costs and the number of outpatient contacts to a VA facility as proxies for health care-seeking behavior. However, these variables could also represent more severe diabetes secondary to poor quality care. More generally, physician-influenced factors in the case-mix model may reflect variation in provider practice patterns as opposed to differences in case-mix severity. For example, the use of insulin, which the authors chose as a proxy for duration of diabetes, partly depends on

providers' thresholds for using this treatment, not necessarily the underlying severity of the disease.

The case-mix adjustment tool must be feasible. It should be simple to use and affordable. While the detailed information available in the medical record and a survey of the patient would be ideal (13), more realistic data elements for widespread use will probably be those found in administrative databases, albeit with datasets that will probably contain less information than the one used by Zhang et al. In addition, the score attained by a rating system or case-mix tool should be reproducible and not dependent on the idiosyncrasies of local charting or coding conventions. The case-mix tool should also be immune to gaming (7), as evidenced after the incorporation of diagnosis-related groups for prospective Medicare payment, in which a "creep" toward higher reimbursing diagnostic codes was seen over time (14).

When assessing if a case-mix tool can be validly applied to a specific target population of diabetic patients, it is critical to determine if the instrument was developed on a similar set of patients. Zhang et al. carefully and correctly note that their findings may apply only to a select group of VA facilities, not the entire VA system. Only 65 of the 173 VA facilities were in their initial dataset. Furthermore, after various exclusions due to factors such as missing data, only ~20% of the diabetic patients in the original cohort of 65 facilities were included in the final case-mix model. If the patients in their ultimate analytical dataset were representative of VA patients, then their results would be generalizable to the VA population. However, the progressive narrowing of the study population raises concerns about selection bias. Moreover, regression models developed on one dataset without validation techniques can be overfitted (15). That is, the case-mix tool might stratify patients extremely accurately on the original dataset, at the expense of making more errors on different groups of patients. Whether the findings in Zhang et al.'s predominantly male VA study sample apply to other patients, such as the broader Medicare population (16,17), is unknown.

Case-mix adjustment tools should be valid, meaning that they should do the required task of adjusting for factors affecting HbA_{1c} value beyond the provider's reasonable control. The problem is that it is difficult to determine how one would validate a case-mix tool for the end point of the

HbA_{1c} value. No gold standard of validity exists for a tool of this type. Nonetheless, a case-mix tool should have at least partial validation before widespread use. For example, the authors use the relatively crude information available in the administrative dataset to operationalize more complex concepts of case-mix severity. More detailed medical records and patient surveys would probably capture the abstract concepts more precisely. In a validation study, one might devise creative ways to compare the data in the administrative dataset with the finer information in the chart or patient survey to get a sense of how accurate the administrative dataset proxy was. With these more detailed chart and survey data sources, investigators could also develop explicit and implicit criteria for case-mix severity along the 6 individual domains described in Zhang et al.'s conceptual model (18,19). Researchers could then determine to what extent the administrative case-mix tool categorized patients into appropriate risk-adjustment strata compared with the more detailed criteria (9). In addition, a case-mix tool developed for one end point, such as HbA_{1c} value, is not necessarily valid for other diabetes-related outcomes, such as amputation. The set of risk factors and their relative importance may vary for different outcomes.

Case-mix adjustment in the diabetes field is still in an early stage. Zhang et al. provide important exploratory data, but I believe that we are not quite ready for prime-time usage of this tool. If the goal is to implement a standard case-mix tool for national performance measurements, it would be premature and potentially dangerous to adopt an instrument that has not been validated. Currently, it is commonly known and intuitively obvious that crude unadjusted quality-of-care rankings have limitations. Risk adjustment helps legitimize quality-of-care ratings. Therefore, if we start complicating report cards with risk adjustment, we must be reasonably confident that the transformed ratings are valid. Moreover, the wider risk adjustment literature notes that the relative ranking of hospitals based on mortality rates is highly dependent on which of several available case-mix adjustment tools is used (20). Therefore, the choice of a national case-mix tool becomes politically contentious, and validation of the instrument becomes essential.

The movement for public ratings of diabetes care continues to gain steam, and

current report cards are being used today as is. Regardless of whether or not report cards are risk-adjusted, it is vital to educate consumers and the media on ratings data, their use, and their strengths and limitations (1,2,7). When we have confirmation that diabetes case-mix tools are at least partially valid, generalizable, and reproducible, the time for risk adjustment in diabetes scorecards will have arrived. Unfortunately, we currently lack much of this information. Therefore, the development and validation of case-mix adjustment tools should be funding priorities.

MARSHALL H. CHIN, MD, MPH

From the Diabetes Research and Training Center, Section of General Internal Medicine, Department of Medicine, Center for Health Administration Studies, University of Chicago, Chicago, Illinois; and the Iowa Foundation for Medical Care, West Des Moines, Iowa.

Address correspondence to Marshall H. Chin, MD, MPH, Section of General Internal Medicine, University of Chicago, 5841 S. Maryland Ave., MC 2007, Chicago, IL 60637. E-mail: mchin@medicine.bsdc.uchicago.edu.

Acknowledgments — This study was supported in part by the Diabetes Research and Training Center (P60 DK20595) of the National Institute of Diabetes and Digestive and Kidney Diseases in Chicago. M.H.C. is a Robert Wood Johnson Foundation Generalist Physician Faculty Scholar.

The author would like to thank Naoko Muramatsu, PhD, Eric C. Schneider, MD, MSc, Willard G. Manning, PhD, and Rodney A. Hayward, MD, for their helpful comments on an earlier draft of this article.

The viewpoints expressed in this article are solely those of the author.

References

1. 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* 283:1866–1874, 2000
2. Epstein AM: Public release of performance data: a progress report from the front. *JAMA* 283:1884–1886, 2000
3. American Diabetes Association: Clinical Practice Recommendations 2000. *Diabetes Care* 23 (Suppl. 1):S1–S116, 2000
4. *Diabetes Quality Improvement Project Initial Measure Set*. Final version. Alexandria, VA, American Diabetes Association, 1998
5. American Medical Accreditation Program. 1999. American Medical Association. 4 April 2000 <<http://www.ama-assn.org/med-sci/amapsite/>>
6. Epstein AM: Rolling down the runway: the challenges ahead for quality report cards. *JAMA* 279:1691–1696, 1998
7. 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* 281:2098–2105, 1999
8. Zhang Q, Safford M, Ottenweller J, Hawley G, Repke D, Burgess JF Jr, Dhar S, Cheng H, Naito H, Pogach L: Performance status of health care facilities changes with risk adjustment of HbA_{1c}. *Diabetes Care* 23: 919–927, 2000
9. Iezzoni LI (Ed.): *Risk Adjustment for Measuring Health Care Outcomes*. 2nd ed. Chicago, Health Administration Press, 1997
10. Chin MH, Zhang JX, Merrell K: Diabetes in the African-American Medicare population: morbidity, quality of care, and resource utilization. *Diabetes Care* 21:1090–1095, 1998
11. Chin MH, Polonsky TS, Thomas VD, Nerney MP: Developing a conceptual framework for understanding illness and attitudes in older, urban African Americans with diabetes. *Diabetes Educ* 26:439–449, 2000
12. Chin MH, Auerbach SB, Cook S, Harrison JF, Koppert J, Jin L, Thiel F, Karrison TG, Harrand AG, Schaefer CT, Takashima HT, Egbert N, Chiu S, McNabb WL: Quality of diabetes care in community health centers. *Am J Public Health* 90:431–434, 2000
13. Hayward RA, Manning WG, Kaplan SH, Wagner EH, Greenfield S: Starting insulin therapy in patients with type 2 diabetes: effectiveness, complications, and resource utilization. *JAMA* 278:1663–1669, 1997
14. Simborg DW: DRG creep: a new hospital-acquired disease. *N Engl J Med* 304:1602–1604, 1981
15. Harrell FE Jr, Lee KL, Mark DB: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 15:361–387, 1996
16. Chin MH, Zhang JX, Merrell K: Specialty differences in the care of older patients with diabetes. *Med Care* 38:131–140, 2000
17. Chin MH, Su AW, Jin L, Nerney MP: Variations in the care of elderly persons with diabetes among endocrinologists, general internists, and geriatricians. *J Gerontol Med Sci*. In press
18. Kahn KL, Rogers WH, Rubenstein LV, Sherwood MJ, Reinisch EJ, Keeler EB, Draper D, Koseoff J, Brook RH: Measuring quality of care with explicit process criteria before and after implementation of the DRG-based prospective payment system. *JAMA* 264: 1969–1973, 1990
19. Smith MA, Atherly AJ, Kane RL, Pacala JT: Peer review of the quality of care: reliability and sources of variability for outcome and process assessments. *JAMA* 278:1573–1578, 1997
20. Iezzoni LI: The risks of risk adjustment. *JAMA* 278:1600–1607, 1997