Summary measures of quality of diabetes care: comparison of continuous weighted performance measurement and dichotomous thresholds

DAVID ARON1, MANGALA RAJAN2 AND LEONARD M. POGACH2

1VA HSR&D Center for Quality Improvement Research, Louis Stokes Cleveland Department of Veterans Affairs Medical Center, Cleveland, OH, and 2VA HSR&D Center for Healthcare Knowledge Management Research, VA New Jersey Healthcare System, East Orange, NJ, USA

Abstract

Background. The Institute of Medicine has suggested that related individual measures of quality be combined into summary measures. Averages of adherence on dichotomous measures for intermediate outcomes have shortcomings because control of individual risk factors differs in their health benefit. Therefore, a common metric is necessary to weight measures appropriately.

Objective. Compare health care system performance using continuous measures weighted based on quality adjusted life years saved (QALYsS) versus dichotomous threshold measures.

Research design. Retrospective cross-sectional analysis of 2000–01 chart abstraction data of diabetic patients from 141 Veterans Health Administration medical centers. Outcome variables included correlation of individual level and facility level adherence to and rankings by continuous weighted individual and summary dichotomous measures for glycemic control (<8% A1c), blood pressure (<140/80 mm/Hg), and low-density lipoprotein-cholesterol (LDL-C) <130 mg/dl.

Results. The 141 facilities had a range of 163–740 (mean 263) subjects. The population (n = 37,142) was largely male (86.1%) and older (mean age 65.9 years, SD ±11.4 years), with mean overall A1c of 7.58%, systolic blood pressure of 137.2 mm/Hg, and LDL-C 104.8 mg/dl. There was excellent correlation between QALYsS and dichotomous outcomes for A1c (r = 0.86), blood pressure (r = 0.94), LDL-C (r = 0.95), and the summary measure (r = 0.92), but poor correlation among the risk factors (r = 0.19–0.36). There was considerable difference in rankings between the dichotomous and the continuous weighted measures; only 46% of facilities remained within the same or within 1 decile.

Conclusion. Continuous weighted measures for the major risk factors for diabetes-related complications have high correlation with dichotomous measures. We propose that a continuous QALYs-weighted summary measure could function as a global measure for the quality of diabetes care.

Keywords: quality measurement, quality of care, diabetes

Performance measurement is an increasingly important aspect of health care from the perspectives of consumers, payers, clinicians, managers, and regulators, and the development of measures entails relevance, soundness, and feasibility [1]. However, responding to the proliferation of measures from diverse organizations and government agencies, the Institute of Medicine (IOM) has recommended the use of summary measures of quality to simplify data presentation [2].

The disease diabetes mellitus illustrates the need for, and difficulty in developing, summary performance measures. Affecting >20 million individuals in the United States alone, it is a major cause of the complications of micro- and macrovascular diseases: chronic kidney disease, visual loss, lower extremity amputations, cardiovascular disease, and death. The societal and payer costs attributable to diabetes are significant and growing [3]. The wide variety of process and intermediate outcome measures has played a role in health care improvement and permitted comparison of care across systems [4].

Glycemic control (A1c), systolic blood pressure (SBP), and low-density lipoprotein-cholesterol (LDL-C) control are intermediate outcomes for diabetes, which are closely linked to morbidity and mortality [5]. Typically, dichotomous measures are used for accountability, i.e. there is a threshold for ‘acceptable’ performance such as percentage of patients with SBP >140. However, dichotomous thresholds are inconsistent with the epidemiological principles and evidence from
diabetes clinical trials that have demonstrated the efficacy of A1c, blood pressure, and cholesterol lowering in persons. For example, whereas relative risk reduction is linear, absolute risk reduction is log-linear; greater absolute risk reduction and benefit accrue from treatment initiated at higher than lower values [6]. In addition, A1c, SBP, and LDL-C control not only differ in their impact on morbidity and mortality [7], but their impacts over a lifetime vary by age [7,8]. Consequently, equally valuing A1c <8%, SBP <140, and LDL-C <130 mg/dl in everyone regardless of age would inaccurately reflect the impact of risk factor control on population health.

Furthermore, dichotomous measures do not provide partial credit, i.e., they do not capture the incremental progress that a clinician may have made in an individual patient towards achieving an ‘optimal’ target threshold. Therefore, the use of threshold measures as the basis for comparisons between clinicians can magnify the impact of biases resulting from the duration of diabetes or comorbid conditions that can impact therapeutic efficacy [9]. Notably, recommended target thresholds for SBP and A1c are difficult to achieve even in clinical trials where additional resources devoted to the intervention are available [10]. Because adherence/achievement rates to a dichotomous measure for an intermediate outcome neither truly reflect quality improvement nor accurately assess population health benefits, their use for public reporting may violate core principles of valuing different populations appropriately and assessing performance fairly. Another approach is needed. We investigate the benefits of an alternative approach.

The development of meaningful summary measures of performance requires a common metric. Measures of population health have been developed using this approach [11–13]. We propose the use of quality-adjusted life years saved (QALYsS) as a common metric that would permit the summation of various intermediate measures and age adjustment and recognize incremental system improvements towards ‘optimal’ targets. This is consistent with recommendations of the Health and Human Services Panel on Cost Effectiveness and is possible for A1c, blood pressure, and cholesterol control, given the available evidence [14]. It is also consistent with the evidence that supports varying health benefits of glycemic control and cholesterol control across different age groups [7]. We have previously compared a continuous measure based on QALYsS to an A1c threshold measure in terms of adherence and league table rankings for the Veterans Health Administration (VHA) [15]. We now evaluate the feasibility of creating a summary measure that combines continuous measures (A1c, SBP, and LDL-C) using the common metric of QALYsS.

Methods

Study design

We obtained from the chart abstraction data VHA Office of Quality and Performance, External Peer Review Program (EPRP) for all veterans identified as having diabetes during the chart review process between 1 October 2000 and 30 September 2001. As previously described [15], the data set consisted of 37 142 individuals >24 years of age. The data set included the last recorded A1c, SBP, and LDL-C as continuous variables or noted that the tests had not been performed. Individuals with triglyceride values >400 mg/dl (n = 2980) were removed to permit calculation of LDL-C. This study was approved by Institutional Review Boards at the Veterans Affairs Medical Centers in Cleveland, OH, and East Orange, NJ.

Determination of a weighted continuous A1c performance measure for individual patients

We determined QALYsS for reduction of A1c from 7.9 to 7.0% for different age groups based on published values from the Centers from Disease Control and Prevention (CDC) as previously reported [7,15]. The following age strata were used: 25–34, 35–44, 45–54, 55–64, 65–74, 75–84, and 285 years. We calculated QALYsS resulting from A1c reduction for each individual by assigning a value of zero to A1cs >7.9%. Because relative risk reduction is linear, we used linear interpolation to assign QALYsS between 7.9% and a value between 7.0 and 7.9% for each patient based on the maximal possible QALY reduction for each age group. Values <7% were assigned the maximal number of QALYsS. Consistent with performance measure criteria used by the National Committee for Quality Assurance/Health Plan Employer Data and Information Set (HEDIS) Comprehensive Diabetes Care, a zero score was assigned if an A1c was not obtained. We then calculated the percentage of the maximal QALYsS for each subject within each facility by dividing the extrapolated QALY score for that individual by the maximal QALYsS value that could be achieved by lowering A1c from 7.9 to 7.0% based on that individual’s age. For example, a 55-year-old can achieve a lifetime maximal of 0.127 QALYsS by lowering A1c from 7.9 to 7.0%. If that individual achieved an A1c of 7.5, the actual QALYsS would equal the maximum QALYsS – {[(A1c value – 7.0)/0.9] × maximum QALYsS} or 0.127 – {[(7.5 – 7.0)/0.9] × 0.127} = 0.127 – 0.0706 = 0.056. The achieved percentage of maximal possible QALYsS = actual QALYsS/maximum QALYsS or 0.0506/0.127 = 45%. This process was used for SBP and LDL-C.

For QALYsS for SBP control, the CDC modeled the results of the United Kingdom Prospective Diabetes Study (UKPDS) [16] because the relative risk reduction in SBP reduction has been demonstrated to be linear [17]. We calculated QALYsS based on the SBP differences between the mean SBP of the control and the treatment arms of the study between 154 and 142 mmHg for all subjects in the sample. A zero score was assigned if SBP was not obtained. We assumed for the purposes of this analysis that everyone started with an SBP of 154 systolic. We then calculated the percentage of the maximal QALY gained for that individual by dividing the extrapolated QALY score for that individual by the maximal possible QALY saved value by lowering SBP from 154 to 142 mmHg. In so doing, the achievement of an SBP <142 received maximum credit.

Finally, QALYsS for cholesterol reduction in the CDC model were obtained from the West of Scotland Coronary Prevention Study, a randomized controlled trial comparing
pravastatin with placebo in individuals without a history of coronary heart disease [18]. The risk reduction achieved by pravastatin was independent of diabetes. Whereas the CDC used cholesterol levels in their analysis, we used mean LDL-C levels of 192 mg/dl for the control group and 142 mg/dl for the treatment group to assign QALYsS, assuming linear equivalence to the cholesterol levels. Similar to A1c and SBP, individuals without an LDL-C were assigned a zero score. Analogous to the approach for SBP, we assumed that everyone started with an LDL-C of 192 mg/dl, and full credit was given for LDL-C levels <142 mg/dl, recognizing that the contemporaneous (2000–01) performance measure threshold for LDL-C was 130 mg/dl.

We created summary measures for each subject by summing the QALYsS for each risk factor, because they are a common unit of measure. We recognize that this approach may overestimate actual QALYsS, but the issue of whether the health benefits of controlling each risk factor are independent is still under investigation in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) Trial. Furthermore, this approach avoids the need for the more complex calculations and more extensive data requirements of lifetime modeling engines, e.g., UKPDS. Finally, we calculated a summary score of the percentage of maximal QALYsS achieved for each facility by aggregating individual summary scores within age groups and for the total population as previously described [15].

**Determination of dichotomous thresholds for individual patients**

We used dichotomous thresholds of A1c, SBP, and LDL-C based on the American Diabetes Association standards that were contemporaneous with our data set (2000–01). We constructed three dichotomous ‘performance measures’ for our main analyses: A1c <8%, SBP <140, and LDL-C <130. This A1c value corresponds to the upper limit of the mean A1c level achieved in the control group in the UKPDS [19]. SBP <140 was also the performance measure for HEDIS, as well as being comparable with the lower level of mean blood pressure achieved in the UKPDS blood pressure substudy and was the VA’s accountability measure [16]. LDL-C <130 was the threshold for the initiation of statin therapy in the 2000 National Cholesterol Education Program Guidelines as well as the HEDIS performance measure for lipid control [20]. An individual subject met adherence/achievement criteria to an individual measure if they achieved an A1c, blood pressure, and LDL-C level that was below the dichotomous threshold for each risk factor. Non-performance of the test was considered as not meeting the criterion. We determined the average percentage adherence for subjects within each age group achieving the measure as well as an average percent adherence for the entire population. We created a summary dichotomous measure by calculating the arithmetic mean (adding the percentage adherence for A1c, blood pressure, and cholesterol and dividing by three), an approach that has been used in other studies to present an average score of adherence to unrelated measures [21].

**Facility level analyses.** At the facility level, the percentage of maximal QALYsS was obtained by dividing the QALY's achieved in an age strata by the total maximal QALYsS possible, which were calculated by multiplying the number of people in each age category by the QALY weight assigned to that age category. A similar approach was used for SBP, LDL-C, and the summary measure.

We then performed bivariate correlations (Pearson’s correlation coefficient) between percentage QALYsS achieved and percentage adherence to each dichotomous threshold for each risk factor and between risk factors for the entire population at each facility.

Because the use of the 90th percentile and the 10th percentile of performance is one industry standard for identifying best- and worst-performing plans [22], we ranked facilities into deciles by the two methodologies (adherence to each dichotomous threshold and percentage of maximal QALYsS) for both the main analyses (A1c <8%, SBP <140/90 mm/Hg, and LDL-C <130 mg/dl).

**Results**

The study population consisted of 37 142 individuals and was largely male (86.1%) and had a mean age (±1 SD) of 65.9 (±11.4); 61% were ≥65 years of age (Table 1). Mean overall A1c, SBP, and LDL-C were 7.58%, 137.2 mm/Hg, and 104.8 mg/dl, respectively. The 141 facilities had a mean (±1 SD) of 263 (±97), with a range of 163–740 patients per facility. There was also variation among facilities in the mean levels of A1c, SBP, and LDL-C.

Facility level performance for individual and summary measures based on dichotomous and continuous weighted measures is summarized in Table 2. Overall facility performance, as assessed by the summary dichotomous measure was 65% (range 0–76%), and was lower than that assessed by % maximal QALY gained [mean (range)] summary measure which was [75 (59–84%)]. Performance [mean (range)] as assessed by dichotomous measures was best for LDL-C [78% (56–89%)], followed by A1c [62% (48–75%)] and SBP [56% (37–70%)]. By contrast, percent maximal QALY gained was greatest for LDL-C [86 (68–97%)], followed by SBP [73 (55–84%)], with A1c trailing further behind [45 (31–60%)]. At the aggregate individual level, both QALYsS and % adherence for A1c, in general, increased with age to 65–74 years and then declined. Both QALYsS and % adherence for A1c decreased for SBP. QALYsS and % adherence for LDL-C showed no pattern with respect to age.

At the facility level, the correlation between the QALYsS and the dichotomous measures was excellent for A1c, BP, and LDL-C; r = 0.86, 0.94, and 0.95 (Pearson’s correlation coefficient), respectively (Table 3). There was also high agreement between the summary of percentage adherence to the dichotomous measures and the percentage of maximal QALYsS achieved (0.92). However, there was modest-to-poor correlation among the dichotomous measures and QALYsS for individual measures or the summary QALYsS (0.19–0.36).

The assessment of rankings for summary measures showed that only 46% of facilities remained in the same or within 1 decile for both the dichotomous and the continuous measures (Figure 1). Facilities changed position as much as 7 deciles. Similar changes were observed with the single measures.
Discussion

Our study suggests that a common metric—QALYs—can be used to develop a summary performance measure for the major intermediate outcomes for diabetes-related morbidity and mortality for several reasons. Firstly, dichotomous measures and continuous weighted measures for the individual measures of A1c, SBP, and LDL-C were highly correlated. Similarly, there was a high correlation between summary scores based on dichotomous and continuous measures. In addition, the marked changes in facility level rankings (including best/worst performers) between the two summary methods despite the high levels of correlation among individual dichotomous and weighted continuous measures demonstrate
The importance of choosing measures that can best capture the domain of health benefit. Indeed, the differences in rankings can most likely be attributed to the fact that by providing ‘partial credit’ towards achieving dichotomous measure goals, the use of the QALYsS measure is assessing progress towards achieving thresholds rather than whether the targets were completely met. Furthermore, because QALYsS are age-adjusted and dichotomous thresholds are not, the use of the continuous weighted measure does to a degree adjust for differences among facilities in patients’ ages, which in this study had a 2 SD magnitude of >11 years. This is particularly relevant for A1c values that decline on a population basis as a function of age [23]. Finally, because performance in one domain, e.g. glycemic control, did not correlate well with performance in other domains, i.e. blood pressure and lipid control, overall facility performance in improving population health is best reflected conceptually in a summary measure.

Our results have important implications for the developers of performance measures, especially because ‘optimal’ levels for A1c (<7%), SBP (<130 mm/Hg), and LDL-C (<100 mg/dl) have recently been adopted for public reporting by the National Committee for Quality Assurance, and the use of all three optimal thresholds in a composite measure has been proposed [24]. However, we are concerned that ‘optimal threshold’-based composite measures will lead to a misperception of poor care when, in reality, most benefit has been achieved because absolute risk reduction is log-linear. Therefore, a weighted intermediate outcome measure that assesses population health benefit is conceptually consistent with the IOM definition of quality of care ‘as the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge’ [11]. The use of QALYsS would make opportunities for improving population health more transparent to clinicians, managers, and the public. Individual risk factors could be used for internal quality improvement [25] rather than public reporting, as is currently the case in the American Diabetes Association/NCQA Bridges to Excellence Program [26].

Another important consideration for physicians and payer groups may be the use of QALYSs as being fairer than an all or none threshold, especially if there is no other allowance for risk adjustment. In addition, it has recently been suggested that ‘fair’ thresholds in health care may result in patients being less likely to consider patient preferences. Although the use of QALYSs has been proposed to improve performance in other domains, i.e. blood pressure and lipid control, the use of all three optimal thresholds in a composite measure has been proposed [24].

Our results have important implications for the developers of performance measures, especially because ‘optimal levels’ for A1c (<7%), SBP (<130 mm/Hg), and LDL-C (<100 mg/dl) have recently been adopted for public reporting by the National Committee for Quality Assurance, and the use of all three optimal thresholds in a composite measure has been proposed [24]. However, we are concerned that ‘optimal threshold’-based composite measures will lead to a misperception of poor care when, in reality, most benefit has been achieved because absolute risk reduction is log-linear. Therefore, a weighted intermediate outcome measure that assesses population health benefit is conceptually consistent with the IOM definition of quality of care ‘as the degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge’ [11]. The use of QALYSs would make opportunities for improving population health more transparent to clinicians, managers, and the public. Individual risk factors could be used for internal quality improvement [25] rather than public reporting, as is currently the case in the American Diabetes Association/NCQA Bridges to Excellence Program [26].

| Table 3 | Correlation in facility rankings between multiple performance measures using Pearson’s rank correlation coefficient (n = 144 facilities) |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Glycemic control <8 | Glycemic control weighted | Hypertension control <140 | Hypertension control weighted | Low-density lipoprotein (LDL) good control <130 | LDL good control weighted | Average control (8 140 130) | Average control weighted |
| Glycemic control <8 | 0.86 (<0.001) | 0.22 (0.009) | 0.19 (0.022) | 0.34 (<0.001) | 0.32 (<0.001) | 0.71 (<0.001) | 0.47 (<0.001) |
| Glycemic control weighted | 0.23 (0.005) | 0.20 (0.016) | 0.36 (<0.001) | 0.34 (<0.001) | 0.66 (<0.001) | 0.50 (<0.001) |
| Hypertension control <140 | 0.94 (<0.001) | 0.27 (<0.001) | 0.27 (<0.001) | 0.22 (<0.01) | 0.20 (0.017) | 0.66 (<0.001) | 0.79 (<0.001) |
| Hypertension control weighted | 0.86 (0.001) | 0.22 (0.009) | 0.19 (0.022) | 0.34 (<0.001) | 0.32 (<0.001) | 0.71 (<0.001) | 0.47 (<0.001) |
| LDL good control <130 | 0.95 (<0.001) | 0.74 (<0.001) | 0.74 (<0.001) | 0.74 (<0.001) | 0.74 (<0.001) | 0.74 (<0.001) | 0.74 (<0.001) |
| LDL good control weighted | 0.68 (<0.001) | 0.68 (<0.001) | 0.68 (<0.001) | 0.68 (<0.001) | 0.68 (<0.001) | 0.68 (<0.001) | 0.68 (<0.001) |
| Average control (8 140 130) | 0.71 (<0.001) | 0.71 (<0.001) | 0.71 (<0.001) | 0.71 (<0.001) | 0.71 (<0.001) | 0.71 (<0.001) | 0.71 (<0.001) |
| Average control weighted | 0.47 (<0.001) | 0.50 (<0.001) | 0.50 (<0.001) | 0.50 (<0.001) | 0.50 (<0.001) | 0.50 (<0.001) | 0.50 (<0.001) |
legitimate patient factors that are not easily obtained from administrative records [9, 27]. Whether risk adjustment can be applied and how clinicians can receive credit based on intensity of treatment would be the subject of future research.

Although we did not directly address cost-effectiveness, presenting population health care outcomes as QALYs would permit policy makers to more fairly compare the health impact of different interventions [14] and permit payer systems to evaluate the cost of quality in a more explicit manner by using the principle of risk stratification [28]. For example, blood pressure reduction would be cost saving among all age groups in our analysis except individuals >85 years of age, whereas ‘tight’ control of A1c, at least on a population basis, is more cost effective in younger populations (<55–60 years) than in the elderly. By minimizing the cognitive dissonance between the business case for quality (public reporting capitation) and true population health (QALYS and cost-effectiveness), the use of percentage of QALYs saved can target quality improvement efforts by demonstrating whether a health care organization can prioritize resources to improve population health, as opposed to ‘meeting thresholds’. There are analogous implications at the facility and individual patient level. For example, improvement efforts at a facility level or therapeutic decisions for the different risk factors could (and one could argue should) be prioritized based on their impact on QALY increase.

Our study has a number of limitations. Firstly, the results depend on the validity of the estimates of utilities for the different health states including the assumptions underlying the calculations of QALYs. Although we used a published model from the CDC that has been applied to the US population as a whole [7], altering the utilities for complications could alter the overall results [29]. Consequently, it would be necessary to obtain the consensus of developers and methodologists. Secondly, our results are based on cross-sectional results but lend themselves to longitudinal analyses; with the availability of electronic medical records, individual patient change scores could be calculated from year to year [30]. Thirdly, although the continuous measure is sufficiently sensitive to demonstrate facility differences, we assessed this measure only at this level where the number of patients exceeded 100 individuals, which has previously been demonstrated to be reliable in demonstrating difference among practices [31]. Fourthly, our population was a largely male veteran population, which could limit its generalizability. On the contrary, recent articles have established the comparability between VHA and private sector commercial plan performance, and gender differences in QALYs have not been reported for A1c, blood pressure, and cholesterol [4].

In conclusion, our study demonstrates that the use of QALYS in a weighted continuous performance measure is a feasible alternative to dichotomous thresholds in the assessment.
of quality of care for diabetes and has implications for payment for physician performance. We suggest that this alternative be investigated in other disorders and other settings.

Acknowledgements

This project was partly funded from Veterans Administration Health Service Grants IIR 00-072-1 and QUERI-Diabetes 98-001 (Dr Pogach) and REA 01-100 (Dr Aron). This work was presented in part at the American Academy of Health Services Research Meeting in Boston, June 2005. We thank Ms Christina Croft and Ms Michelle Davidson for their editorial assistance in preparing the manuscript. The authors thank VHA’s Office of Quality and Performance for providing access to the EPRP data. The opinions expressed are those of the authors and do not represent the opinions of the VA.

Funding

Veterans Administration Health Service Grants IIR 00-072-1 and QUERI–Diabetes 98-001 (Dr. Pogach) and REA 01-100 (Dr. Aron).

References


Accepted for publication 8 November 2006