Published in final edited form as:
Circ Cardiovasc Qual Outcomes. 2012 January 1; 5(1): 44–51. doi:10.1161/CIRCOUTCOMES.111.961474.

Estimating and Reporting on the Quality of Inpatient Stroke Care by Veterans Health Administration Medical Centers

Greg Arling, PhD[1,3,4], **Mathew Reeves, PhD**[1,5], **Joseph Ross, MD**[6], **Linda S. Williams, MD**[1,4,7,8], **Salomeh Keyhani, MD**[9,10], **Neale Chumbler, PhD**[1,4,7,11], **Michael S. Phipps, MD**[12], **Christianne Roumie, MD, MPH**[13,14], **Laura J. Myers, PhD**[1,3,7], **Amanda H. Salanitro, MD**[14,15], **Diana L. Ordin, MD, MPH**[1,2], **Jennifer Myers, MSW**[1,7], and **Dawn M. Bravata, MD**[1,3,4,7]

[1]VHA Health Services Rsrch & Development (HSR&D) Stroke Quality Enhancement Rsrch Initiative (QUERI), Indianapolis, IN

[2]Veterans Health Administration (VHA) Office of Informatics & Analytics, Washington, DC

[3]Dept of Med, Indiana Univ School of Med

[4]Regenstrief Inst, Indianapolis, IN

[5]Dept of Epidemiology, Michigan State Univ, East Lansing, MI

[6]Section of General Internal Med, Dept of Med, Yale Univ School of Med, and Ctr for Outcomes Rsrch & Evaluation, Yale-New Haven Hospital, New Haven, CT

[7]VHA HSR&D Ctr of Excellence on Implementing Evidence-Based Practice (CIEBP); Richard L. Roudebush VHA Med Ctr

[8]Dept of Neurology, Indiana Univ School of Med, Indianapolis, IN

[9]Division of Gen Internal Med, Univ of California at San Francisco, San Francisco, CA

[10]HSR&D Rsrch Enhancement Award Program San Francisco VA Med Ctr, San Francisco, CA

[11]Dept of Sociology, Indiana Univ School of Liberal Arts, Indiana Univ Purdue Univ Indianapolis, Indianapolis, IN

[12]Robert Wood Johnson Foundation Clinical Scholars Program & Dept of Neurology, Yale Univ School of Med, New Haven, CT

[13]HSR&D Targeted Rsrch Enhancement Program Ctr, GRECC, and Clinical Rsrch Training Ctr of Excellence, Veterans Affairs–Tennessee Valley Healthcare System, Nashville, TN

[14]Inst of Med & Public Health, Vanderbilt Univ Med Ctr, Nashville, TN

[15]Section of Hospital Med, Vanderbilt Univ Med Ctr, Nashville, TN

Abstract

Background—Reporting of quality indicators (QIs) in Veterans Health Administration Medical Centers is complicated by estimation error due to small numbers of eligible patients per facility. We applied multilevel modeling and empirical Bayes (EB) estimation in addressing this issue in performance reporting of stroke care quality in the Medical Centers.

Correspondence: Greg Arling, PhD, Regenstrief Institute, 410 West 10th Street, Suite 2000, Indianapolis, IN 46202-3012, Phone 317-423-5634, Fax 317-423-5695, GArling@IUPUI.edu.

Conflicts of Interest Disclosures: None

Methods and Results—We studied a retrospective cohort of 3812 veterans admitted to 106 Medical Centers with ischemic stroke during fiscal year 2007. The median number of study patients per facility was 34 (range: 12-105). Inpatient stroke care quality was measured with thirteen evidence-based QIs. Eligible patients could either pass or fail each indicator. Multilevel modeling of a patient's pass/fail on individual QIs was used to produce facility-level EB estimated QI pass rates and confidence intervals. The EB estimation reduced inter-facility variation in QI rates. Small facilities and those with exceptionally high or low rates were most affected. We recommended 8 of the 13 QIs for performance reporting: dysphagia screening, NIH Stroke Scale documentation, early ambulation, fall risk assessment, pressure ulcer risk assessment, Functional Independence Measure documentation, lipid management, and deep vein thrombosis prophylaxis. These QIs displayed sufficient variation across facilities, had room for improvement, and identified sites with performance that was significantly above or below the population average. The remaining 5 QIs were not recommended because of too few eligible patients or high pass rates with little variation.

Conclusions—Considerations of statistical uncertainty should inform the choice of QIs and their application to performance reporting.

Keywords

acute stroke; performance measurement; quality indicators; stroke management; statistics

Quality assessment systems are being adopted widely in the health care industry for performance incentives and public reporting. Controversy surrounds the choice of the quality measures, risk adjustment methods, and quality thresholds or standards[1-4]. Another issue receiving increased attention is the problem of estimation error or the statistical uncertainty associated with quality measures based on small numbers of observations per facility [5-10]. Estimation error, if unrecognized or inadequately dealt with, can undermine the performance assessment and make it difficult to draw fair comparisons among organizations.

The Veterans Health Administration (VHA) has a long-standing tradition of performance measurement across a variety of disease conditions and often out-performs the private sector [11-13]. We recently conducted a study of the quality of inpatient ischemic stroke care in Veterans Administration Medical Centers [14]. A goal of the study was to recommend stroke quality indicators (QIs) for potential adoption by the VHA quality reporting system. Dealing with uncertainty in facility QI rates was a major concern in making recommendations. The QI rates represented the number of eligible patients “passing” a recommended care process. Although the QI rates were informative at the population level, we found it difficult to draw conclusions about facility-level performance because of the low volume of stroke cases at many hospitals. The median number of annual stroke cases per facility was 34 among the 129 hospitals in the study. Exclusions of ineligible patients further reduced the denominators for several QIs. The smaller the denominator for a facility's QI rate, i.e., number patients eligible for a QI, the greater the imprecision or uncertainty of the QI estimate and the more likely a hospital being evaluated would appear as an outlier, with either exceptionally high or low quality, through chance alone.

In the present study we selected an approach, multilevel modeling, that could address estimation error and identify candidate QIs for facility performance assessment in the VHA. Our study objectives were to: (1) construct facility-level estimated inpatient stroke QI rates based on multilevel models and empirical Bayes (EB) estimation methods, (2) compare the EB-estimated with observed QI rates to determine effects of EB estimation on facility QI rates and outlier status, (3) evaluate how well each QI discriminated between facilities, and

(4) draw conclusions about which stroke QIs would be most useful for performance reporting across hospitals.

Methods

Patients and Setting

The VHA Office of Quality Performance and the VA Stroke Quality Enhancement Research Initiative (QUERI) collaborated to conduct the first VHA study of national in-patient ischemic stroke care quality using 14 process indicators based largely on the Joint Commission Standardized Stroke Measure Set [15]. The study design and primary findings have been reported elsewhere [14] (see: The Quality of VA Inpatient Ischemic Stroke Care, Supplemental Materials).

The total study sample consisted of 3,931 veterans admitted to 129 hospitals in fiscal year 2007 with a primary discharge diagnosis of ischemic stroke based on ICD-9 codes from administrative data [16]. All patients admitted with stroke to lower volume centers (≤55 patients in fiscal year 2007) and a random 80% sample of patients admitted with stroke at higher volume centers (>55 patients in fiscal year 2007) were sampled. The cut-off of 55 patients was chosen in order to optimize the number of sampled patients from smaller volume facilities while staying within budget constraints. Patients were excluded from the sample if they were admitted for elective carotid endarterectomy, admitted only for post-stroke rehabilitation, were already admitted for a non-stroke condition when the ischemic stroke event occurred, or were admitted to a hospital that did not use the VHA electronic medical record system. For purposes of our current analysis we excluded from the total sample 23 hospitals with fewer than 12 patients (fewer than 1 stroke admission per month on average) because results for these very low volume facilities would be highly unreliable. The cut-off of 12 patients was selected arbitrarily. We set a low cut-off with the intention of including as many facilities as possible. The resulting analysis sample was 3,812 patients from 106 hospitals. The median number of patients per facility was 34 with a range of 12 to 105.

Data were collected through retrospective chart review of medical records using remote electronic medical record data only (not paper medical records), performed by abstractors from the West Virginia Medical Institute (WVMI) who were specially trained for this study. Among the 307 data elements, 90% had an inter-observer agreement ≥70%. The institutional review board approved this study.

Quality Indicators

Thirteen of the 14 study QIs were included in this analysis: dysphagia screening before oral intake, documentation of stroke severity using the NIH Stroke Scale (NIHSS), thrombolysis (tPA) given, antithrombotic therapy by hospital day two and at discharge, deep vein thrombosis (DVT) prophylaxis, early ambulation, fall risk assessment, pressure ulcer risk assessment, and rehabilitation needs assessment based on documentation of the Functional Independence Measure (FIM), atrial fibrillation management at discharge, lipid management at discharge, and smoking cessation counseling [14] (see: The Quality of VA Inpatient Ischemic Stroke Care, Supplemental Materials, for detailed definitions). The 14th QI that was included in the original study pertained to stroke education but this QI was excluded from the current analysis because of unreliable documentation.

Patient-level QIs were measured as a binary variable for eligible patients: pass or fail. A facility's QI rate was the number of passes divided by the number of patients who were eligible for that QI. Patients were excluded from the calculation for an indicator if they had characteristics or conditions making them ineligible [14] (see: The Quality of VA Inpatient

Ischemic Stroke Care, Supplemental Materials). The analysis did not include additional patient risk-adjusters because appropriateness of the care process had already been dealt with in the QI definition through patient exclusion criteria.

Multilevel Modeling and Empirical Bayes Estimates

We based our analysis on a multilevel modeling approach, which takes into account the nested and correlated structure of quality assessment data (e.g., patients nested within hospitals). It is better suited to deal with estimation error and other estimation problems compared with conventional regression methods that do not deal adequately with the hierarchal nature of these data [4, 17-20]. Multilevel models generally provide more precise and reliable estimates and more valid conclusions about performance, particularly when providers differ widely in the volume of patient populations [21-26].

The multilevel models produce empirical Bayes (EB) estimates of facility QI rates. The EB-estimated QI rates were developed in stages with the HLM 6.0 statistical package (Scientific Software International, Chicago IL). First, we constructed hierarchical general linear (HGLM) models for each of the 13 quality indicators. The binary patient-level QI outcome (pass or fail) was modeled with facility treated as a random effect [27]. The models assumed a logit link function and Bernoulli distribution and we used EM Laplace estimation [28]. These were null or unconditional models containing a facility error term but no patient or facility variables. Only patients who were eligible for a QI were included in the model for that QI.

From the patient-level models we output facility-level empirical Bayes (EB) residuals to represent the variance in a facility's QI rate [28]. The EB residuals differ from ordinary least squares residuals in that they take into account differences between facilities in the reliability of their estimates. As noted earlier, facilities with small numbers of stroke patients are likely to have less reliable estimates compared to facilities with a higher volume of stroke admissions. The EB estimator, sometimes referred to as a "shrinkage estimator," produces more precise and reliable and usually more conservative estimates because rates for small outlier facilities are pulled to the population mean [28-30]. The HLM software produces EB residuals for each QI in each facility. For ease of interpretation we calculated standardized QI rates by adding the EB residuals to the grand mean QI rate for the facility population. Facilities with a positive residual for a QI had a standardized rate above the mean, while negative residuals resulted in standardized rates below the population mean. For example, the mean pass rate for the fall risk assessment QI was .78. If a facility had an EB residual of .05 then its standardized rate would be .83; if its EB residual were -.05 then its standardized rate would be .73. By standardizing the EB QI rates in this manner, we are able to make direct comparisons to the observed QI rates. The HLM software also outputs a standard error for each residual that was used to construct 90% confidence intervals around the EB estimated rate.

Quality Thresholds

The study had no predetermined standards or thresholds for stroke care quality performance. We sought to test three approaches for threshold setting that have been applied generally in health care quality assessment. One approach relied on tests of statistical significance where EB QI estimates and confidence intervals were compared to a quality threshold. We chose the population mean as one threshold. Facilities with QI confidence intervals entirely above the population mean would have significantly better than average quality; facilities with QI confidence intervals entirely below the mean would have significantly poorer quality; and facilities overlapping the mean would be uncertain (i.e., neither above nor below average). The population average threshold becomes problematic when performance is generally poor in a QI area (e.g., dysphagia screening or NIHSS documentation) or quite good overall (e.g., antithrombotics by day 2). Therefore, we selected an alternative threshold based on an absolute standard of an 85% pass rate that could be applied across all of the QIs. The 85% pass rate is the benchmark used by the American Heart Association Get with the Guidelines Stroke Recognition Program in recognizing achievements in stroke care quality [31]. Facilities would be designated as good, poor, or uncertain if their QI confidence intervals fell above, below or crossed over an 85% threshold. Finally, a third approach relied on outlier status: a facility ranked below the 10th percentile would be a low performer and above the 90th percentile a high performer. We chose the 90% CI rather than a 95% CI because we wanted greater sensitivity in identifying facilities that might be providing good or poor care, i.e., CIs above or below our thresholds.

Results

Facilities varied considerably in the number of stroke patients admitted and patients eligible for each QI (Table 1). No facility had 12 or more patients for the tPA administration or atrial fibrillation management QIs. Therefore, we did not report results for these QIs. Only 31 facilities had ≥12 patients eligible for the DVT prophylaxis QI and 41 facilities had ≥12 patients eligible for the smoking cessation counseling QI. At least 90 facilities had ≥12 eligible patients for the remaining QIs; the median number of eligible patients per facility ranged from 19-35. Even among the highest volume facilities (90th percentile) the number of eligible patients per QI ranged from only 47-58.

Facility Quality Indicator (QI) Rates

Table 2 shows the observed and estimated QI rates for the 11 QIs that had analyzable results. Facilities performed very well on several QIs. The mean observed pass rates were above 85% on the QIs for antithrombotics by day 2 and at discharge, early ambulation, pressure ulcer risk assessment, and smoking cessation counseling. Performance on other QIs was generally poor: the dysphagia screening prior to oral intake process had an observed mean pass rate of only 0.173 and the NIHSS stroke scale documentation rate was only 0.250. The observed mean pass rates for the other QIs ranged from 0.749 (DVT prophylaxis) to 0.815 (lipid management at discharge). The facilities displayed modest variation in observed QI rates with the NIH Stroke Scale and fall risk assessment having the greatest variation in rates (SD = 0.380 and SD = 0.335, respectively; see Table 2).

EB-Estimated Facility Performance

The EB-estimated QI rates displayed a pattern similar to the observed rates. Means for the facility QI rates changed very little between observed and EB estimated (see Table 2). The main effect of EB estimation was to reduce (shrink) the variation in QI rates; standard deviations declined for all of the EB QIs compared to the observed QIs. Similarly, the range in rates between facilities at the 10th and 90th percentiles narrowed. Facilities farther away from the mean were likely to experience greater shrinkage.

Table 3 shows the impact of EB estimation on the facility QI rates and outlier status. The table displays the average absolute difference in observed and EB QI rates (absolute value of the facility's EB rate minus its observed rate). The mean absolute differences ranged from a low of 0.004 (SD=0.004, interquartile range=.002-.005) for NIHSS documentation to a high of 0.053 (SD=0.048, interquartile range=.021-.088) for DVT prophylaxis.

We would expect EB estimation to have its greatest impact at the tails of the distribution where the greatest EB shrinkage is likely to occur. Table 3 summarizes facility movement into or out of outlier status (i.e., above the 90th percentile or below the 10th percentile). Some QIs having the highest observed rates (i.e., antithrombotics by day 2, antithrombotics at discharge, fall risk assessment, and early ambulation) had several facilities with observed QIs rates of 1.000. This created tied rankings for top outlier status. The EB estimation removed the ties through QI shrinkage away from 1.000. As a consequence, the greatest movement out of top outlier status occurred for QIs having the most tied facilities. Nonetheless facilities displayed considerable movement in or out of outlier status for all of the QIs. For example, 5 facilities became bottom outliers while 5 facilities moved out of the bottom outlier status on the FIM administration QI; and, 2 facilities became top outliers and 13 facilities moved out of top outlier status on the early ambulation QI.

Confidence Intervals

Given the small denominators for many of the facility QIs, we would expect considerable uncertainty in the estimates even among EB rates. The confidence intervals around these estimates can be a valuable tool in evaluating performance while taking into account the uncertainty of the QI rate. Figures 1 and 2 illustrate the application of confidence intervals to EB QI rates for lipid management that has generally high pass rates and dysphagia screening that has generally low pass rates. The width of the confidence intervals (i.e., precision of the QI estimates) is mainly a function of the facility QI denominators. Facilities having more patients eligible for a QI will have more precise and reliable estimates. Although we chose a 90% CI because of its greater sensitivity in detecting good or poor performance, a 95% CI could be used in a QI reporting system where specificity, i.e., minimizing false positives, was of greater concern.

Quality Thresholds

We relied on EB rates and confidence intervals in deciding which facilities were good or poor performers on the 11 analyzable QIs (Table 4). Our first approach, comparing the confidence intervals for the facilities' EB QI rates to the population mean, discriminated relatively well for 8 of the 11 QIs. However, this method discriminated poorly for three QIs with the highest pass rates (i.e., antithrombotics by day 2, antithrombotics at discharge, and Smoking cessation counseling) where no more than 5% of facilities were significantly above or below the population average.

Facilities also varied significantly in their performance compared to the 85% pass rate standard on most of the QIs. For example, 41% of facilities were significantly above and 15% significantly below the standard on the early ambulation QI, 43% were above and 22% below on fall risk assessment, and 60% above and 8% below on pressure ulcer assessment. Only 12% of facilities were significantly above while 80% were significantly below the standard for NIH Stroke Scale administration. The standard discriminated very little between facilities on the other QIs. All facilities fell significantly below the 85% pass rate for the dysphagia screening, none was significantly above the standard for DVT prophylaxis, and none was significantly below the standard for smoking cessation counseling. Almost all facilities were above the 85% standard for the antithrombotic QIs at day 2 and at discharge.

Discussion

Our study demonstrates an approach for dealing with the problem of estimation error, a serious issue in assessing the quality of inpatient ischemic stroke care in the VHA and other health care settings. The low volume of stroke admissions in many VHA facilities leads to uncertainty about their true QI rates. We approached this problem by first setting a threshold of at least 12 patients (one patient per month) in the QI denominator before a facility's QI rate would be reported. Secondly, we developed multilevel models for estimating patient pass rates in each facility for the 13 QIs. The models generated empirical Bayes (EB) estimates and 90% confidence intervals around these estimates. The EB estimation narrowed the between-facility variance in QI rates. The QI rates in smaller facilities with extreme values were pulled toward the population average QI rate, reducing the likelihood that these facilities would be labeled QI outliers. By constructing 90% confidence intervals around the EB QI rates we were able to discriminate statistically between good and poor performing facilities using both the population mean and an 85% pass rate as quality thresholds.

Advantages over Conventional Methods of QI Reporting

The EB estimation offers advantages over conventional methods of calculating QI rates. One method of QI reporting is to simply ignore estimation error and report observed QI rates leaving the user with little guidance as to the uncertainty in the estimates. Another method is to report observed rates but excludes small volume facilities from QI calculations if they do not attain a minimum number of patients in the QI denominator. This method is problematic because it leaves unreported the QI rates for small volume facilities and because it fails to take into account differences in estimation error among facilities above the threshold [30]. A third method is to report observed rates with conventional confidence intervals. The confidence intervals convey the amount of uncertainty associated with the rates in small facilities, but having such broad confidence intervals makes it difficult to draw inference about care quality.

The EB estimation approach is an improvement over these methods. It is grounded in Bayesian statistical theory, uses information about QI rates in the patient population and variation in rates between facilities to inform the estimation process, and produces estimates that are more precise and reliable and with narrower confidence intervals than conventional methods[27, 29]. One concern frequently raised about EB estimates is the conservative bias inherent in shrinkage. Low volume, outlier facilities having poor QI scores may be “let off the hook” as their QI rates shrink toward the population mean. We should point out that shrinkage is bidirectional: small volume facilities with exceptionally good and poor scores will both experience shrinkage. We feel that a conservative approach is warranted from a policy perspective. We want to avoid penalizing low volume facilities by falsely labeling then as having exceptionally poor care. This approach runs the risk of some facilities with poor quality being undetected or good quality unrewarded. We recognize that outlier status based on rankings can be highly variable and uncertain[19, 20] and, thus, we would caution against this approach, recommending instead the application of confidence intervals in relation to benchmarks (85% pass rate).

Recommended QIs for Performance Reporting of VHA Stroke Quality

Seven of the 13 stroke quality indicators in our analysis would be good candidates for public reporting: dysphagia screening, NIH Stroke Scale documentation, early ambulation, fall risk assessment, pressure ulcer assessment, FIM documentation, and lipid management. These QIs are recommended for use for the following reasons. First, they can be calculated and reported for the majority of hospitals (i.e., at least 95 facilities had 12 or more patients in the denominators for these QIs). Second, they had reasonably large inter-facility variation in QI pass rates even after EB estimation. As an indication of variation, their standard deviations ranged from 0.071 (lipid management) to 0.375 (NIHSS administered). Third, their mean EB pass rates were neither exceptionally low nor high; they ranged from 0.172 to 0.920. Finally, these QIs discriminated reasonably well between facilities according to our quality threshold approaches. Among the seven QIs, no fewer than 8% of facilities had EB pass rates significantly below their population average, and at least 10% of facilities were significantly above the their population average. The proportion of facilities significantly below an 85% pass rate standard was no less than 8%, and the proportion above 85% was no greater than 60%.

Some of the QIs in our study differ from measures in other national reporting systems. Dysphagia screening was recently dropped from the set of stroke measures proposed by National Quality Forum (NQF) despite evidence linking dysphagia screening with improvements in important clinical outcomes [14]. Also, fall risk assessment, pressure ulcer assessment, FIM documentation are not NQF recommended measures for stroke; however, these processes are embedded strongly in the VHA culture and have important implications for clinical outcomes.

The DVT prophylaxis QI is not among the seven recommended measures because it had only 31 facilities with 12 or more eligible patients. In other respects this QI meets our criteria. The DVT prophylaxis QI has a strong evidence base, an important consideration in QI reporting [32]. The 31 facilities included in the QI denominator had a reasonable population average QI rate (0.751); 13% of facilities were significantly below and 12% significantly above average on this QI; and 36% were significantly below the 85% pass rate standard, leaving room for improvement. The DVT prophylaxis QI could be informative for facilities with sufficient numbers of eligible patients.

The remaining 5 measures, while not recommended for facility performance reporting, are still clinically important and worthy of measurement and tracking. The tPA administration for the treatment of acute stroke and management of atrial fibrillation processes have a strong evidence base [32]. Although these two QIs had too few patients per facility for reliable performance reporting, these clinically meaningful care processes could be treated as sentinel events with a focus on individual occurrences of failure. The remaining QIs – antithrombotic by day 2, antithrombotic by discharge, and smoking cessation counseling have very high average pass rates (>0.950). These measures should be tracked because of their clinical importance, yet they should not be part of a performance reporting system because they do not offer meaningful comparisons between facilities.

Because of their greater precision the EB rates should be a more reliable basis for performance assessment not only in a single time period but also in drawing comparisons over multiple time periods. Stroke admission rates have been stable over the last 5 years. If the recommended QIs are to be applied to performance assessment and publicly reported, then data should be collected in a consistent manner from year to year. Key data elements should be incorporated into the VHA administrative and clinical systems. We are exploring the feasibility of measuring the QIs from administrative and clinical systems on an ongoing basis without abstracting. Our study suggests that 12 months of data would be required for QI reporting. Data could be reported every calendar quarter using a 12-month rolling average with data updated each quarter.

We tested the sensitivity of the findings to our choice of a 12-patient minimum for facility QI reporting. Increasing the minimum number of patients required for calculating a facility's QI rate yielded more reliable QI estimates and reduced the difference between observed and EB estimated QI rates. Yet, increased reliability came at a price. For example, a 25- patient minimum reduced the number of VAMCs per QI from an average of 88 to 55, with a maximum of 70 facilities (pressure sore risk assessment) and minimum of 6 facilities (DVT prophylaxis) having reportable QI rates.

Limitations

Our study has limitations. First, we concentrated on the issue of estimation error. Other issues such as the clinical significance, measurement validity and reliability, eligibility criteria, and potential for quality improvement may have an equal or even greater impact on the choice of QIs for the inpatient stroke quality reporting system [10, 33]. Second, the study was conducted within the VHA, which has unique delivery system characteristics and patient population characteristics that may make findings difficult to generalize to other settings [11]. Third, statistical estimation is complex and may be difficult for providers or consumers to understand or difficult for hospital systems to calculate. The EB QI rates with confidence intervals may be most appropriate for system-level applications where facilities must be compared against their peers in quality report cards or pay-for-performance programs. The EB rates should be accompanied by observed rates, particularly for quality improvement purposes where providers may want to investigate individual passes and failures. Finally, we did not attempt to develop a composite QI where scores on individual QIs could be combined into a single measure. Although composite measures tend to be more reliable than individual measures, they may obscure differences in performance between care domains and some individual measures may dominate others.[34]

Conclusion

Despite these limitations the study demonstrates an approach for dealing with statistical uncertainty using multilevel modeling and empirical Bayes estimation; it resulted in QI estimates and confidence intervals that discriminate well between good and poor performing facilities; and it identified reportable stroke care quality indicators that met our criteria for performance reporting.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The Department of Veterans Affairs, Veterans Health Administration (VHA), Office of Quality and Performance and Health Services Research and Development Service Quality Enhancement Research Initiative (RRP 09-184) supported this project. Dr. Bravata is the principal investigator at the Richard L. Roudebush VA Medical Center (Indianapolis, IN). The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs. Dr. Ross is supported by the National Institute on Aging and by the American Federation of Aging Research through the Paul B. Beason Career Development Award Program (K08 AG032886). Dr. Keyhani is supported by a VA HSR&D Career Development Award.

References

1. Hibbard JH, Stockard J, Tusler M. Does publicizing hospital performance stimulate quality improvement efforts? Health Aff (Millwood). 2003; 22:84–94. [PubMed: 12674410]
2. Robinowitz DL, Dudley RA. Public reporting of provider performance: Can its impact be made greater? Annu Rev Public Health. 2006; 27:517–536. [PubMed: 16533128]
3. Dudley RA, Robinowitz DL, Talavera JL, Broadhead P, Luft HS. Strategies to support quality-based purchasing: A review of the evidence. 2004
4. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, Howard G, Peterson ED, Rathore SS, Smith SC Jr, Spertus JA, Wang Y, Normand SL. Standards for statistical models used for public reporting of health outcomes. Circulation. 2006; 113:456–462. [PubMed: 16365198]
5. Fung V, Schmittdiel JA, Fireman B, Meer A, Thomas S, Smider N, Hsu J, Selby JV. Meaningful variation in performance: A systematic literature review. Med Care. 2010; 48:140–148. [PubMed: 20057334]
6. Davidson G, Moscovice I, Remus D. Hospital size, uncertainty, and pay-for-performance. Health Care Financ Rev. 2007; 29:45–57. [PubMed: 18624079]
7. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: The problem with small sample size. Jama. 2004; 292:847–851. [PubMed: 15315999]
8. Hayward RA, Heisler M, Adams J, Dudley RA, Hofer TP. Overestimating outcome rates: Statistical estimation when reliability is suboptimal. Health Serv Res. 2007; 42:1718–1738. [PubMed: 17610445]

9. O'Brien SM, Delong ER, Peterson ED. Impact of case volume on hospital performance assessment. Arch Intern Med. 2008; 168:1277–1284. [PubMed: 18574084]

10. Normand S-LT, Shahian DT. Statistical and clinical aspects of hospital outcomes profiling. Statistical Science. 2007; 22:206–226.

11. Asch SM, McGlynn EA, Hogan MM, Hayward RA, Shekelle P, Rubenstein L, Keesey J, Adams J, Kerr EA. Comparison of quality of care for patients in the veterans health administration and patients in a national sample. Ann Intern Med. 2004; 141:938–945. [PubMed: 15611491]

12. Ross JS, Keyhani S, Keenan PS, Bernheim SM, Penrod JD, Boockvar KS, Federman AD, Krumholz HM, Siu AL. Use of recommended ambulatory care services: Is the veterans affairs quality gap narrowing? Arch Intern Med. 2008; 168:950–958. [PubMed: 18474759]

13. Kerr EA, Gerzoff RB, Krein SL, Selby JV, Piette JD, Curb JD, Herman WH, Marrero DG, Narayan KM, Safford MM, Thompson T, Mangione CM. Diabetes care quality in the veterans affairs health care system and commercial managed care: The triad study. Ann Intern Med. 2004; 141:272–281. [PubMed: 15313743]

14. Bravata DM, Wells CK, Lo AC, Nadeau SE, Melillo J, Chodkowski D, Struve F, Williams LS, Peixoto AJ, Gorman M, Goel P, Acompora G, McClain V, Ranjbar N, Tabereaux PB, Boice JL, Jacewicz M, Concato J. Processes of care associated with acute stroke outcomes. Arch Intern Med. 2010; 170:804–810. [PubMed: 20458088]

15. Joint Commission. Standardized stroke measure set (harmonized measures). 2011

16. Reker DM, Hamilton BB, Duncan PW, Yeh SC, Rosen A. Stroke: Who's counting what? J Rehabil Res Dev. 2001; 38:281–289. [PubMed: 11392661]

17. Burgess JF Jr, Christiansen CL, Michalak SE, Morris CN. Medical profiling: Improving standards and risk adjustments using hierarchical models. J Health Econ. 2000; 19:291–309. [PubMed: 10977193]

18. DeLong E. Hierarchical modeling: Its time has come. Am Heart J. 2003; 145:16–18. [PubMed: 12514649]

19. Goldstein H, Spiegelhalter D. League tables and their limitations: Statistical issues in comparisons of institutional performance. Journal of the Royal Statistical Society Series A (Statistics in Society). 1996:385–443.

20. Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: Issues and applications. Journal of the American Statistical Association. 1997; 92:803–814.

21. Glance LG, Dick A, Osler TM, Li Y, Mukamel DB. Impact of changing the statistical methodology on hospital and surgeon ranking: The case of the new york state cardiac surgery report card. Med Care. 2006; 44:311–319. [PubMed: 16565631]

22. Austin PC, Tu JV, Alter DA. Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: Should we be analyzing cardiovascular outcomes data differently? Am Heart J. 2003; 145:27–35. [PubMed: 12514651]

23. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Stat Med. 1997; 16:2645–2664. [PubMed: 9421867]

24. Hannan EL, Wu C, DeLong ER, Raudenbush SW. Predicting risk-adjusted mortality for cabg surgery: Logistic versus hierarchical logistic models. Med Care. 2005; 43:726–735. [PubMed: 15970789]

25. Hinchey JA, Shephard T, Tonn ST, Ruthazer R, Selker HP, Kent DM. Benchmarks and determinants of adherence to stroke performance measures. Stroke. 2008; 39:1619–1620. [PubMed: 18323510]

26. Moerbeek M, van Breukelen GJ, Berger MP. A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. J Clin Epidemiol. 2003; 56:341–350. [PubMed: 12767411]

27. Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage Publications; 2002.

28. Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage Publications; 2002.

29. Greenland S. Principles of multilevel modelling. Int J Epidemiol. 2000; 29:158–167. [PubMed: 10750618]

30. Arling G, Lewis T, Kane RL, Mueller C, Flood S. Improving quality assessment through multilevel modeling: The case of nursing home compare. Health Serv Res. 2007; 42:1177–1199. [PubMed: 17489909]

31. American Heart Association. Get with the guidelines-stroke recognition criteria. 2011

32. Adams HP Jr, del Zoppo G, Alberts MJ, Bhatt DL, Brass L, Furlan A, Grubb RL, Higashida RT, Jauch EC, Kidwell C, Lyden PD, Morgenstern LB, Qureshi AI, Rosenwasser RH, Scott PA, Wijdicks EF. Guidelines for the early management of adults with ischemic stroke. Circulation. 2007; 115:e478–534. [PubMed: 17515473]

33. Reeves MJ, Parker C, Fonarow GC, Smith EE, Schwamm LH. Development of stroke performance measures: Definitions, methods, and current measures. Stroke. 2010; 41:1573–1578. [PubMed: 20489174]

34. Peterson ED, Delong ER, Masoudi FA, O'Brien SM, Peterson PN, Rumsfeld JS, Shahian DM, Shaw RE, Goff DC Jr, Grady K, Green LA, Jenkins KJ, Loth A, Radford MJ. Accf/aha 2010 position statement on composite measures for healthcare performance assessment: A report of the american college of cardiology foundation/american heart association task force on performance measures (writing committee to develop a position statement on composite measures). Circulation. 2010; 121:1780–1791. [PubMed: 20351232]

What is known?

- Quality of inpatient stroke care is a serious concern that can affect patient outcomes such as functional status and mortality
- Quality indicators (QIs) can provide useful information for public reporting, performance incentives, and quality improvement, although prior to this study the Veterans Health Administration did not have QIs for the quality of inpatient stroke care.
- Drawing inference about the “true” quality of stroke care for a specific hospital and making valid comparisons between hospitals is complicated by estimation error, particularly when hospitals have a low volume of stroke patients or they vary widely the volume of their patients.

What this article adds?

- We present a method for dealing with QI estimation error that relies on a multilevel framework, i.e., patients nested within hospitals, and an empirical Bayes estimator.
- We demonstrate the practical application of this method in estimating and reporting on the quality of stroke care using 14 care process QIs in an annual cohort of 3812 inpatient ischemic stroke patients in 106 Veterans Health Administration medical centers.

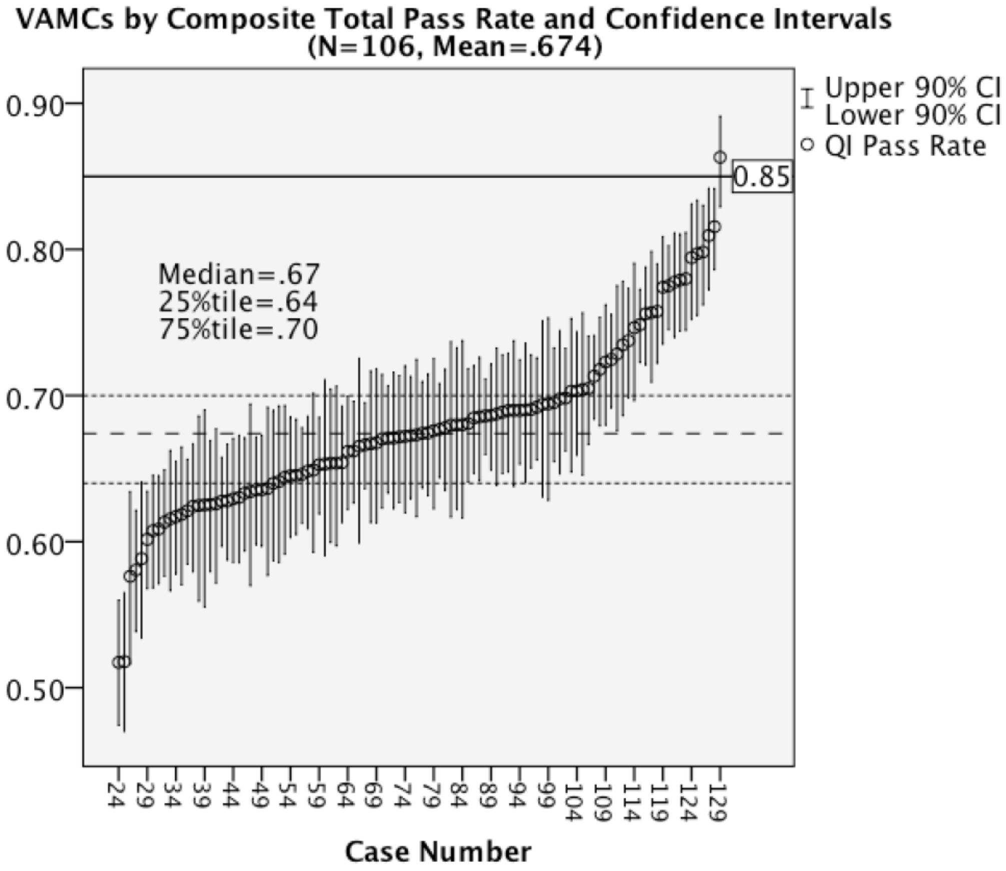

Figure 1.

The vertical axis is the EB estimated QI rate. Case numbers on the horizontal axis are randomly assigned facility ID numbers. Not all cases are labeled due to space constraints.

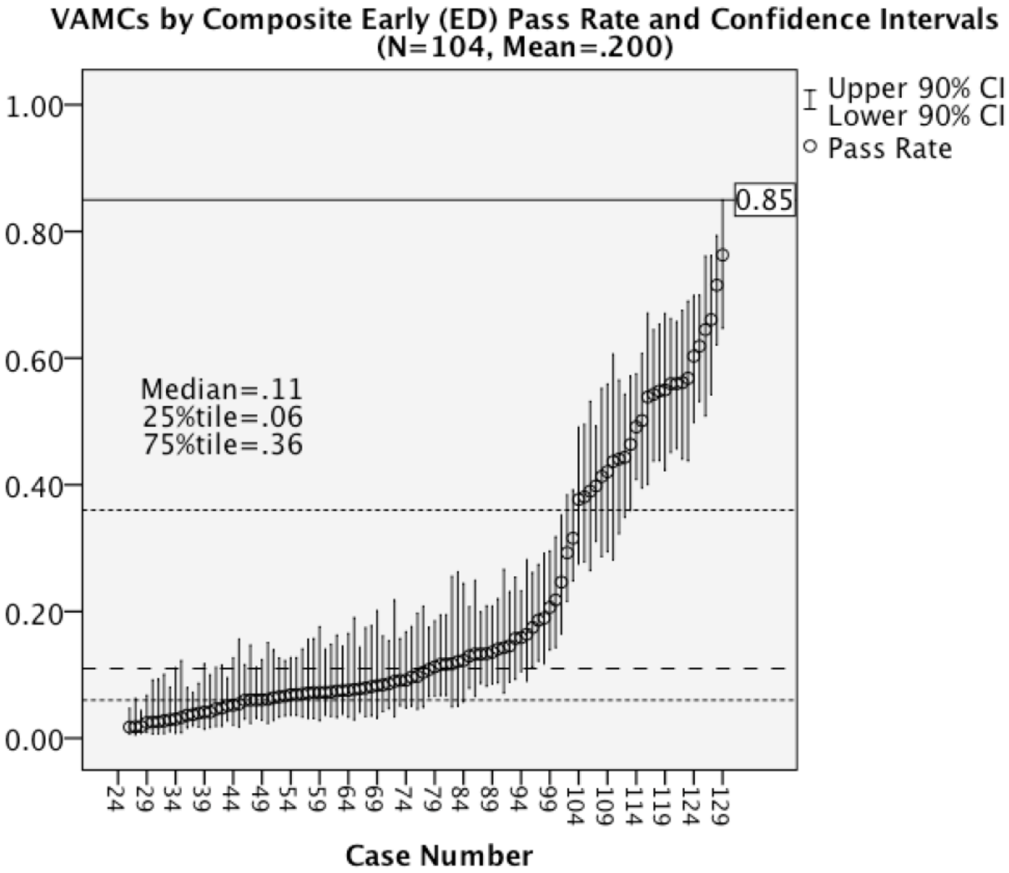

Figure 2.

The vertical axis is the EB estimated QI rate. Case numbers on the horizontal axis are randomly assigned facility ID numbers. Not all cases are labeled due to space constraints.

Table 1

Number of Patients per Facility by Quality Indicator for Facilities with a Minimum of 12 Patients

Quality Indicators	Facilities	Total Patients	Mean # Patients/Facility	Median # Patients/Facility	Std. Dev. Patients/Facility	10th %tile Patients/Facility	90th %tile Patients/Facility
Antithrombotic by day 2	101	3349	33.2	31	16.2	15	54
Antithrombotic at discharge	100	3334	33.3	31	16.2	16	54
Atrial fibrillation management	0	0					
Deep vein thrombosis (DVT) prophylaxis	31	595	19.2	16	9.3	12	30
Dysphagia screening	102	3474	34.1	32	16.8	16	58
Early ambulation	96	2845	29.6	28	14.4	14	49
Fall risk assessment	102	3498	34.3	33	17.0	16	54
FIM™ documentation	101	3347	33.1	31	16.1	16	52
Lipid management	95	2827	29.8	28	13.6	15	47
NIH Stroke Scale Administration	101	3460	34.3	33	16.8	16	54
Pressure sore risk assessment	102	3606	35.4	33	17.5	16	56
Smoking cessation counseling	41	789	19.2	17	6.5	12	29
Thrombolysis (tPA) administration	0	0					

Table 2

Quality Indicator (QI) Rates for Facilities with a Minimum of 12 Patients per QI

		Observed QI Rates					**EB Estimated QI Rates**				
Quality Indicator	**Number of Facilities**	**Mean**	**Median**	**Std. Dev.**	**90th %tile**	**10th %tile**	**Mean**	**Median**	**Std. Dev.**	**90th %tile**	**10th %tile**
Antithrombotic by day 2	101	0.951	0.950	0.051	0.882	1.000	0.955	0.957	0.019	0.930	0.974
Antithrombotic at discharge	100	0.957	0.971	0.057	0.893	1.000	0.960	0.970	0.028	0.931	0.982
Deep vein thrombosis (DVT) prophylaxis	31	0.749	0.800	0.163	0.515	0.985	0.756	0.787	0.093	0.597	0.868
Dysphagia screen	102	0.173	0.125	0.160	0.000	0.417	0.172	0.126	0.137	0.047	0.366
Early ambulation	96	0.857	0.923	0.150	0.600	1.000	0.859	0.919	0.125	0.665	0.970
Fall Risk Assessment	102	0.784	0.925	0.335	0.000	1.000	0.785	0.924	0.327	0.033	0.991
FIM™ administration	101	0.803	0.816	0.144	0.625	0.960	0.803	0.818	0.105	0.668	0.912
Lipid management	95	0.815	0.824	0.117	0.637	0.961	0.819	0.828	0.071	0.727	0.906
NIH Stroke Scale administration	101	0.250	0.024	0.380	0.000	0.975	0.248	0.024	0.375	0.002	0.967
Pressure sore risk assessment	102	0.916	0.958	0.124	0.779	1.000	0.920	0.957	0.102	0.810	0.984
Smoking cessation counseling	41	0.947	1.000	0.085	0.839	1.000	0.955	0.982	0.052	0.903	0.985

Table 3

Difference between Observed and EB Estimated Facility QI Rates and Outlier Status

		Absolute Difference EB Estimated – Observed Rate#				Change in Outlier Status (Top or Bottom 10%) after EB Estimation*			
	Number of Facilities	Mean	Std. Dev.	Range	Inter-Quartile Range	Become Bottom Outlier	Out of Bottom Outlier	Become Top Outlier	Out of Top Outlier
Antithrombotic by day 2	101	0.027	0.022	.000-.159	.014-.031	2	1	0	28
Antithrombotic at discharge	100	0.022	0.021	.000-.143	.014-.023	10	0	2	2
Deep vein thrombosis (DVT) prophylaxis	31	0.053	0.048	.000-.202	.021-.088	0	0	1	1
Dysphagia screen	102	0.020	0.018	.001-.071	.006-.029	5	7	1	1
Early ambulation	96	0.022	0.017	.001-.087	.008-.032	1	2	2	13
Fall risk assessment	102	0.006	0.007	.000-.035	.001-.010	1	2	0	20
FIM™ administration	101	0.030	0.028	.000-.206	.011-.045	5	5	1	0
Lipid management	95	0.038	0.030	.000-.139	.011-.058	6	6	1	1
NIH Stroke Scale administration	101	0.004	0.004	.000-.021	.002-.005	10	0	0	0
Pressure sore risk assessment	102	0.017	0.018	.000-.159	.006-.021	0	0	0	11
Smoking cessation counseling	41	0.026	0.023	.001-.117	.015-.024	4	0	1	2

#Differences are calculated by taking the absolute value of the facility’s EB estimated QI rate minus its observed QI rate.

*Change in outlier status between facility rankings according to observed QI rates and facility rankings according to EB estimated QI rates.

Table 4

EB Estimated QI Rates Significantly Above or Below the Mean Facility Quality Indicator (QI) Rate or an 85% Standard Pass Rate

Quality Indicators	Number of Facilities	Mean Facility QI Rate	Facilities Significantly* Below the Mean	Facilities Significantly* Above the Mean	Facilities Significantly* Below 85%0	Facilities Significantly* Above 85%
Antithrombotic at discharge	100	0.961	5%	0%	0%	89%
Antithrombotic by day 2	101	0.957	3%	1%	0%	94%
Deep vein thrombosis (DVT) prophylaxis	31	0.751	13%	12%	36%	0%
Dysphagia screen	102	0.166	21%	18%	100%	0%
Early ambulation	96	0.877	21%	31%	15%	41%
Fall risk assessment	102	0.785	18%	55%	22%	43%
FIM™ administration	101	0.809	15%	15%	25%	5%
Lipid management	95	0.821	8%	10%	14%	1%
NIH Stroke Scale administration	101	0.246	68%	26%	80%	12%
Press sore risk assessment	102	0.920	12%	22%	8%	60%
Smoking cessation counseling	41	0.952	5%	0%	0%	66%

[*] Statistical significance is based on the QI’s 90% confidence interval falling entirely above or below the population mean or 85% pass rate standard.