Reliability of the American College of Surgeons Commission on Cancer's Quality of Care Measures for Hospital and Surgeon Profiling

Jason B Liu, MD, Kristopher M Huffman, MS, Bryan E Palis, MA, Lawrence N Shulman, MD, David P Winchester, MD, FACS, Clifford Y Ko, MD, MS, MSHS, FACS, Bruce L Hall, MD, PhD, MBA, FACS

BACKGROUND: Efforts to improve healthcare quality involve profiling hospitals and providers. Whether cancer-specific measures can be used reliably for profiling purposes has not been reported.

STUDY DESIGN: Hospitals and surgeons were profiled with 3 measures assessing the adequacy of lymphadenectomy for colon (ie at least 12 regional lymph nodes [12RLN] are removed and pathologically examined for resected colon cancer), gastric (ie at least 15 regional lymph nodes [G15RLN] are removed and pathologically examined for resected gastric cancer), and non-small cell lung (ie at least 10 regional lymph nodes [10RLN] are removed and pathologically examined for American Joint Committee on Cancer stage IA, IB, IIA, and IIB resected non-small cell lung cancer) cancers using hierarchical models. National Cancer Data Base cases spanning 2010 to 2013 were included if they met measure eligibility. Reliability estimates for hospital and surgeon performance across cumulative years of data (2013, 2012 to 2013, 2011 to 2013, and 2010 to 2013) were calculated with and without risk adjustment. Surgeon caseload minimums were projected to achieve reliabilities of 0.40 and 0.70.

RESULTS: Reliability estimates tended to increase with longer periods of data collection but at different rates, depending on measure, level of aggregation, and performance outlier status. Profiling hospitals using 12RLN with 2 years of data yielded a median reliability of 0.72 (interquartile range [IQR] 0.55 to 0.83); however, 4 years of data yielded a median reliability of only 0.31 (IQR 0.14 to 0.54) for surgeons. The G15RLN performance was poor overall; 10RLN had high reliability at both hospital (0.74; IQR 0.50 to 0.86) and surgeon (0.61; IQR 0.34 to 0.80) levels using 1 year of data, but the literature questions this measure's validity. Few surgeons could achieve appropriate levels of reliability regardless of increased data collection duration.

CONCLUSIONS: Profiling hospitals based on measures such as these can achieve acceptable reliability in reasonable timeframes, but does not always. Either lower levels of reliability should be accepted to profile surgeons with these measures or longer timeframes should be used. (J Am Coll Surg 2017;224:180−190. © 2016 by the American College of Surgeons. Published by Elsevier Inc. All rights reserved.)

The Centers for Medicare and Medicaid Services will soon apply payment adjustments based partly on provider performance.[1] The appropriateness of certain performance measures for benchmarking, profiling, and public reporting has been heavily scrutinized, given the potential for misclassification.[2-5] Profiling efforts prominently feature the hospital and now the provider. In some cases, these efforts have also lacked specialty or disease specificity. Stakes are high for these quality measures to be valid and reliable.[6,7]

Disclosure Information: Nothing to disclose.

Support: Dr Liu is supported by a research fellowship from the Department of Surgery, University of Chicago Hospitals and the American College of Surgeons Clinical Scholar in Residence Program.

Received August 13, 2016; Revised October 21, 2016; Accepted October 23, 2016.

From the Division of Research and Optimal Patient Care, American College of Surgeons (Liu, Huffman, Palis, Winchester, Ko, Hall), Department of Surgery, University of Chicago Hospitals (Liu), Chicago, IL, Abramson Cancer Center, Department of Medicine, University of Pennsylvania, Philadelphia, PA (Shulman), Department of Surgery, University of California Los Angeles David Geffen School of Medicine (Ko), VA Greater Los Angeles Healthcare System (Ko), Los Angeles, CA, Department of Surgery (Hall), Center for Health Policy and the Olin Business School (Hall), Washington University in St Louis, Saint Louis Veterans Affairs Medical Center (Hall), and BJC Healthcare (Hall), St Louis, MO.

Correspondence address: Jason B Liu, MD, American College of Surgeons, 633 N St Clair, 22nd Floor, Chicago, IL 60611. email: jliu@facs.org

http://dx.doi.org/10.1016/j.jamcollsurg.2016.10.053
ISSN 1072-7515/16

Validity is necessary but not sufficient to characterize the overall appropriateness of a quality measure for performance evaluation.[8,9] Reliability is distinct from validity, yet depends on it for meaning: one cannot achieve reliable depictions of true performance differences if one is not measuring what the metric claims to measure to start. Specifically, reliability is a ratio of "signal" to "signal plus noise" and quantifies the degree to which a performance measure is based on true differences in performance.[2,10,11] It typically relates measurement error with apparent performance differences. Practically, reliability allows one to gauge whether measured differences in performance are, in fact, due to true provider differences or are more likely to be "noise." Therefore, assessing a measure's reliability is vital for performance measurement.

The American College of Surgeons' Commission on Cancer (CoC), in collaboration with multidisciplinary stakeholders, defines and publishes quality of care measures used for maintaining accreditation standards, benchmarking efforts, and driving internal quality improvement.[12] These measures span a variety of cancers, are categorized by their evidence base, include process and outcomes measures, and, in some cases, have been endorsed by the National Quality Forum.[13] Currently, these measures are used to evaluate hospital performance. However, 3 measures judging the adequacy of lymph node harvest are potentially more attributable to the surgeon compared with the applicability of other CoC quality measures as provider-specific measures. Therefore, the objective of this study was to determine whether these 3 measures had acceptable levels of reliability to profile hospitals, surgeons, or both. The numbers of surgeons with acceptable ($\geq$0.4) and good ($\geq$0.7) levels of reliability were projected. Increasing data collection duration and risk-adjustment are potential strategies to improve reliability, and therefore these were considered. Whether these measures should be considered process or intermediate-outcomes measures was also explored.

METHODS

Data source

The CoC Quality of Care measures use data from the National Cancer Data Base, a joint project of the American Cancer Society and the American College of Cancer CoC, which captures approximately 70% of all newly diagnosed malignancies nationwide across approximately 1,500 CoC-accredited hospitals.[14,15] The National Cancer Data Base collects patient demographics, tumor characteristics, and treatment details. Patient information is de-identified and clustered by hospital and by provider. Trained registrars abstract data based on North American Association of Central Cancer Registries standards,[16] the *Facility Oncology Registry Data Standards* (FORDS),[17] and Collaborative Stage Data Collection System schemas.[18] Hospitals are periodically audited to ensure data integrity.

Study cohorts

Cohorts were created using operations performed from 2010 to 2013 using publically available measure specifications as follows: at least 12 regional lymph nodes were removed and examined for resected colon cancer (12RLN),[19] at least 15 regional lymph nodes were removed and examined for resected gastric cancer (G15RLN),[20] and at least 10 regional lymph nodes were removed and examined for American Joint Committee on Cancer stage IA, IB, IIA, and IIB resected non-small cell lung cancer (10RLN).[21] Measure compliance is binary at the patient level, meaning that a patient either had sufficient lymph node yield or not. Again, these measures were chosen for study because they were potentially most attributable to an individual provider (the surgeon) compared with the other measures.[12] In addition, they represented clinical diversity and varying degrees of evidence. The 12RLN measure for colon cancer, for instance, is endorsed by the National Quality Forum[13] and is supported by other national associations, given the strong evidence supporting its validity,[22-27] and is used to benchmark CoC hospitals. In comparison, the 10RLN measure for non-small cell lung cancer is currently neither National Quality Forum-endorsed nor used for hospital benchmarking because of its controversial link to improved outcomes.[15,28-30]

All resections were identified using FORDS site-specific surgery codes and staged according to the American Joint Committee on Cancer manual, 7th edition.[26] Patients with missing surgeon National Provider Identifier numbers (used to cluster by surgeon) or that had invalid National Provider Identifier numbers[31] were excluded (12RLN: n = 20,460 [12%], G15RLN: n = 1,855 [17%], 10RLN: n = 14,050 [14%]). National Provider Identifier numbers were recoded before any subsequent analyses to preserve surgeon anonymity, thereby exempting this study from formal IRB review.

Data collection periods

Because the NCDB captures all resections from CoC-accredited hospitals, lengthening data collection (vs increasing sampling) would be the primary mechanism for increasing case volumes for profiling purposes. To simulate the effect of increasing case volumes on reliability, operations performed in 2013, 2012 to 2013, 2011 to 2013, and 2010 to 2013 represented data collection periods of 1, 2, 3, and 4 years, respectively.