Published in final edited form as:
J Perinatol. 2011 November ; 31(11): 702–710. doi:10.1038/jp.2011.12.

Formal selection of measures for a composite index of NICU quality of care: Baby-MONITOR

J Profit[1,2,3], **JB Gould**[4,5], **JAF Zupancic**[6,7], **AR Stark**[1], **KM Wall**[1,3,8], **MA Kowalkowski**[2,3], **M Mei**[2,3], **K Pietz**[2,3], **EJ Thomas**[9], and **LA Petersen**[2,3]

[1]Section of Neonatology, Department of Pediatrics, Baylor College of Medicine, Texas Children's Hospital, Houston, TX, USA

[2]Section of Health Services Research, Department of Medicine, Baylor College of Medicine, Houston, TX, USA

[3]Houston Veterans Affairs (VA) Health Services Research and Development Center of Excellence, Health Policy and Quality Program, Michael E DeBakey VA Medical Center, Houston, TX, USA

[4]California Perinatal Quality Care Collaborative, Palo Alto, CA, USA

[5]Division of Neonatology, Perinatal Epidemiology and Health Outcomes Research Unit, Department of Pediatrics, Stanford University, Palo Alto, CA, USA

[6]Department of Neonatology, Beth Israel Deaconess Medical Center, Boston, MA, USA

[7]Division of Newborn Medicine, Department of Pediatrics, Harvard Medical School, Boston, MA, USA

[8]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, USA

[9]University of Texas—Memorial Hermann Center for Healthcare Quality and Safety, University of Texas Medical School, Houston, TX, USA

Abstract

Objective—To systematically rate measures of care quality for very low birth weight infants for inclusion into Baby-MONITOR, a composite indicator of quality.

Study Design—Modified Delphi expert panelist process including electronic surveys and telephone conferences. Panelists considered 28 standard neonatal intensive care unit (NICU) quality measures and rated each on a 9-point scale taking into account pre-defined measure characteristics. In addition, panelists grouped measures into six domains of quality. We selected measures by testing for rater agreement using an accepted method.

Result—Of 28 measures considered, 13 had median ratings in the high range (7 to 9). Of these, 9 met the criteria for inclusion in the composite: antenatal steroids (median (interquartile range)) 9(0), timely retinopathy of prematurity exam 9(0), late onset sepsis 9(1), hypothermia on

Correspondence: Dr J Profit, Houston VA HSR&D Center of Excellence (152), 2002 Holcombe Boulevard, Houston, TX 77030, USA. profit@bcm.edu .

Conflict of interest Drs Profit, Zupancic and Gould will serve as Expert Consultants with the Vermont Oxford Network's NICQ 7 Quality Improvement Collaborative. Dr Gould is the principal investigator for the California Perinatal Quality Care Collaborative.

The results of this study were presented as a poster at the Pediatric Academic Societies' Annual Meeting in Baltimore on 5 May 2009.

Supplementary Information accompanies the paper on the Journal of Perinatology website (http://www.nature.com/jp)

admission 8(1), pneumothorax 8(2), growth velocity 8(2), oxygen at 36 weeks postmenstrual age 7(2), any human milk feeding at discharge 7(2) and in-hospital mortality 7(2). Among the measures selected for the composite, the domains of quality most frequently represented included effectiveness (40%) and safety (30%).

Conclusion—A panel of experts selected 9 of 28 routinely reported quality measures for inclusion in a composite indicator. Panelists also set an agenda for future research to close knowledge gaps for quality measures not selected for the Baby-MONITOR.

Keywords

infant; newborn; quality of health care; measurement; composite indicator; Delphi method

Introduction

The release of the Institute of Medicine's influential reports on patient safety and quality[1,2] has invigorated the focus on improving quality of care. Quality deficits and variations have been documented in the neonatal intensive care unit (NICU) setting.[3–7] These variations are important, as they are often associated with preventable morbidity and mortality.[3–5,8]

Reducing variation in quality and outcomes through measuring, reporting and rewarding quality of health-care delivery has become a national health policy priority.[9] In other areas of medicine, multi-stakeholder initiatives and health-care payers are experimenting with comparative quality measurement (benchmarking),[10] public release of performance information[11,12] and financial incentives to improve quality.[12–14]

Composite indicators of care quality have been used in adult medicine to measure and track global provider performance.[15–18] A composite performance measure combines two or more indicators into a single number to summarize multiple dimensions of provider performance and to facilitate comparisons.[19] Composites have several potential desirable attributes, including improving communication among stakeholders through reduced complexity of data generated, providing global insights and trends on quality for internal and external benchmarking, and improving the reliability and efficiency of provider measurement.[20–23] In addition, global measurement solutions via composite indicators have the potential to foster comprehensive approaches to quality improvement. Such approaches might change the current practice of addressing individual domains of quality (for example, use of a specific medication) to a more systems-based approach, which may improve care across multiple domains (for example, improving teamwork and safety culture[24,25]).

However, the process of developing composite indicators is complex, and developers have to make choices in their construction that may significantly influence performance ratings.[26] For example, a previous report compared differences between composite scores derived from the hospital quality scorecard generated by the Centers for Medicare and Medicaid Services and from the quality scorecard generated by *US News and World Report*.[27] Based on these composite scores, the authors found frequent discordance in hospital performance between the scorecards. This discrepancy highlights the importance of a standardized and explicit approach to scorecard and composite indicator development. A particularly explicit approach to composite indicator development has been described by JRC (European Commission Joint Research Center).[23]

Specifically, the JRC advocates a purposeful 10-step approach to composite indicator development, which guides developers through various stages of conceptualization, computation, testing and dissemination.[22] Although the JRC targets global measurement of social and economic domains, the methods are broadly applicable to the health-care setting.

Our overarching goal is to develop a composite indicator that could be used to assess the global quality of care provided by each NICU member of a quality improvement consortium. Initially, we are developing a composite indicator of quality delivered to very low birth weight infants (VLBW), which we will name Measure Of Neonatal InTensive care Outcomes Research or Baby-MONITOR. Here, we report on an essential step for indicator development, selection of quality measures for inclusion in the Baby-MONITOR.

Methods

Framework for quality measurement

The framework for development of the Baby-MONITOR has been described in detail elsewhere.[22] In brief, ideally the Baby-MONITOR would consist of six subpillars, each representing one of the domains of quality described by the Institute of Medicine (safety, effectiveness, efficiency, patient-centeredness, timeliness and equity).[2] Each subpillar would combine several measures of quality. The Baby-MONITOR was specifically created using outcome and process measures routinely collected at the NICU level by the California Perinatal Quality Care Collaborative (CPQCC) and the VON (Vermont Oxford Network).

Patient sample and measurement techniques

Our sampling framework attempts to compare similar populations of VLBW infants among hospitals while minimizing the number of patients that would be excluded. Exclusion criteria and their rationale are shown in Table 1. These criteria were vetted by a panel of experts described below and attempt to reconcile the competing demands of accuracy, fairness and generalizability.

Measure selection and definition

From the CPQCC and the VON operations manuals, we pre-selected 28 candidate measures for expert vetting. The pre-selection process was designed for inclusiveness, limited only to exclude measures of surgical quality and based on consensus among participating researchers. Measures included VLBW annual volume, antenatal steroids, temperature measured within 1 h of NICU admission, hypothermia at NICU admission, surfactant administration within 2 h of birth, timely retinopathy of prematurity (ROP) examination, severe ROP (>stage 2), ROP surgery, any intracranial hemorrhage, intracranial hemorrhage severity >grade 2, cystic periventricular leukomalacia, use of assisted ventilation, duration of assisted ventilation, pneumothorax, postnatal steroids for chronic lung disease, oxygen on day 28, oxygen at 36 weeks postmenstrual age, oxygen at initial discharge, discharge home or to long-term care facility on assisted ventilation, necrotizing enterocolitis, necrotizing enterocolitis surgery, feeding with human milk only at discharge, feeding any human milk at discharge, growth velocity, healthcare-associated infection, length of stay, 28-day mortality and mortality during NICU admission. Measure details are shown in Supplementary information 1.

Expert panel and Delphi process

A panel of 15 experts rated commonly reported measures of quality via a modified Delphi process. Panelists were selected based on peer recommendations using criteria to include recognized expertise in neonatal outcomes and quality research, and geography. Participating panelists are acknowledged below.

The rating exercise was conducted between March and August 2008. We provided each panelist with introductory materials including explicit definitions of the selected quality measures, rating sheets for each measure and instructions regarding the Delphi process.[28] Panelists were instructed to consider the applicability of all measures at the population level rather than their relevance to the individual patient. Panelists were asked to put themselves in the position of a NICU administrator evaluating performance reports in the context of local improvement efforts with current and historical clinical performance. We clarified that the Baby-MONITOR could potentially be used for either internal quality improvement or external comparative measurement of quality of care.

In addition, we asked panelists whether any measure not selected for the composite should be used as a sentinel indicator. Sentinel indicators describe individual events that are undesirable and represent the extreme of poor performance, triggering further analysis. Measures so designated would be added to the CPQCC report card in case of significant performance deviation (>2 s.d.) from the group mean. A two-thirds majority was required for measure designation as a sentinel indicator.

Measure ratings

Measure rating sheets included a summary of each measure including its classification, description, rationale, risk adjustment method, numerator and denominator, and variable type. Measures were rated on a scale of 1 to 9 (9 being best) for importance, reliability, validity, scientific soundness, usability and overall score. The specific prompts used to clarify these criteria are shown in Table 2.

In the first round of ratings, panelists individually rated the 28 measures. Panelists then received both individual and group ratings and discussed each measure and its ratings during two conference calls in May and June 2008. In all, 12 of 15 panelists participated in at least one of the conference calls, and conference call minutes were distributed among all panelists. The panelists then re-rated the 28 measures for overall score. Potential differences in ratings between conference call participants and non-participants were evaluated using *t*-tests.

Measure domains

In addition, we asked experts to attribute measures to the quality of care domains defined by the Institute of Medicine:[2] safety (the ability to provide care with minimal detrimental errors), effectiveness (the proper use of evidence-based care and the ability of that care to attain a specific therapeutic objective), efficiency (the non-wasteful use of health-care resources), patient-centeredness (the ability to prioritize patient desires and values for guidance of clinical decisions), timeliness (the ability to provide timely care) and equity (care delivery independent of patients' gender, race or socioeconomic status). Panelists also summarized measures according to their applicability for benchmarking. Each measure was graded as level 1, 2 or 3 as defined in Table 3.

Measure selection process

Measures were selected for the Baby-MONITOR if they passed three criteria adopted from accepted gold standard methods developed by researchers at RAND studying the appropriateness of medical care delivery.[28,29] The first was a high median rating (7 to 9). The second criterion tested for agreement via the hypothesis that 80% of the ratings were within the high range (7 to 9). If the hypothesis could not be rejected at $P<0.33$, the measure was rated 'with agreement.' The third criterion tested for disagreement via the hypothesis that 90% of ratings are within one of two ranges (1 to 6 or 4 to 9). If that hypothesis could be rejected on a binomial test at the $P<0.1$, the measure was rated 'with disagreement.' These significance levels were selected by RAND to accommodate variable panel sizes and provide a statistical solution to measure selection.[30] We validated the results of the RAND approach with a method used by multinational European panels carried out as part of the BIOMED Concerted Action on Appropriateness designed for 15 panel members.[31]

Validation of the selection process

We assessed clinician agreement with the Delphi panel by surveying a national sample of 46 neonatal intensive care practitioners. Study subjects were selected based on recommendation by the District Chairs of the American Academy of Pediatrics Perinatal Section on account of the following qualities: board certification, experience with quality improvement, peer respect, public/private mix and geographic distribution. Contacts for enrollment were limited to three attempts. Clinician participants received electronic surveys, which provided general instructions, and detailed information on each quality measure, including attribute ratings by the Delphi panel. We then asked the clinician group to indicate their level of agreement with the Delphi panel for each measure using a 5-point scale (much too high, slightly too high, reasonably, slightly too low, much too low). In addition, we assessed whether clinicians would select the same measures as the Delphi panel for inclusion in the composite based on an up or down vote on each measure and a pre-specified two-thirds majority for inclusion. In this study, we only inquired about 27 measures of quality as one (duration of ventilation) had not been consistently recorded in the database.

Results

Measure ratings

Table 4 exhibits Delphi panelist ratings of measures according to importance, reliability, validity, scientific soundness, usability, and overall score in Round 1 and overall score in Round 2. Of 28 measures considered, 13 had high median ratings (7 to 9). Of these, 9 met the criteria for inclusion in the composite (bolded) using either the RAND or the European BIOMED study selection criteria.

Overall scores changed little between the two rounds of ratings. In general, measures not favored by panelists in the first round became less favored in the second round and the ones that were favored became more so. Discussions among panelists led to the inclusion of three measures (oxygen at 36 weeks postmenstrual age, any human milk at discharge and growth velocity) in the final round of rating that would not have been selected following the initial round of ratings. These new inclusions were the result of either an increased median score or decreased variability in scores from initial to final round ratings.

Among the four measures not selected for the final composite, one was too similar to another more highly rated measure (panelists preferred in-hospital mortality over 28-day mortality). The other three measures failed the agreement criterion. For example, although the median rating was high, the early surfactant measure generated substantial disagreement among the panel with regard to whether or not infants given a trial of continuous positive airway pressure should be included in the denominator. Therefore, following RAND methods, the measure was excluded. (A table summarizing the expert discussions for each measure is available in electronic version as Supplementary information 2).

Final ratings from panelists not participating in either of the two conference calls did not differ significantly from panelists that did participate in at least one of the conference calls (P>0.1). Among measures not selected for the composite, severe intracranial hemorrhage (>grade 2), temperature measurement within the first hour of admission, postnatal steroids for chronic lung disease and length of stay were selected as ‘sentinel indicators’ based on ≥10 of 15 votes by panelists.

Measure domains

Panelist attributions of measures to domains of quality are presented in Figures 1 and 2. Most measures, whether selected for the Baby-MONITOR or not, map onto the domains of safety and effectiveness. Among the nine measures selected for the composite panelists assigned four (timely ROP exam, hypothermia on admission, late onset infection and pneumothorax) to the domain of safety and five (antenatal steroids, oxygen at 36 weeks postmenstrual age, growth velocity, any human milk at discharge and in-hospital mortality) to the domain of effectiveness.

Validation

Of clinician nominees, 23 (47.8%) responded to the survey. We found high levels of agreement between clinical neonatologists and the Delphi panel. Survey participants selected the same nine measures for inclusion in the composite as the panel from the original study. For these nine measures, 74% of clinicians indicated that Delphi panel rating was reasonable; 18% thought Delphi panel ratings were slightly too high.

Discussion

We report a systematic rating process to select measures of neonatal intensive care quality as candidates for inclusion into the Baby-MONITOR, a composite indicator. As composite indicators are being used increasingly for provider profiling, rigorous development methods are necessary to provide accurate feedback regarding performance. Multiple strategies for indicator selection have been developed. These strategies can be classified into two basic methodologies: participatory and statistical. Participatory methods, such as the modified Delphi method used in this study, optimize face validity. While this results in a composite indicator that may be acceptable to users, it may contain measures that contribute little to the measurement of overall quality of care. On the other hand, statistical methods (for example, factor analysis and principal component analysis) provide a more mathematically parsimonious indicator set, but may lack face validity.[23] We favored using the participatory method for indicator selection to enhance acceptability of the Baby-MONITOR among neonatologists. While other participatory group methods exist (for example, consensus development conference), we preferred the widely used and generally accepted Delphi method because it does not require consensus among a conferencing group, and therefore, is less dictated by the opinions of dominant individuals.

The NICU setting may be ideal for the development of composite indicators, as many patients remain in a single physical location or a defined network and are under the control of one care group for the duration of the initial hospital stay. This allows for better attribution of responsibility for care quality to individual NICUs than is the case for other inpatient and ambulatory care settings, in which patients are treated by a multitude of providers in different locations. In addition, standardized clinical quality measures have been developed and are being collected by numerous NICUs. The NICU setting therefore provides a good framework to test whether global measurement of quality via composite indicators will support comprehensive improvements in care delivery.

Measure selection

This paper presents an explicit quantification of the quality of commonly recorded measures of neonatal intensive care. Panelists rated 9 of 28 measures sufficiently high and with agreement to satisfy criteria for inclusion in a composite indicator of quality. Our results have important implications for the neonatal quality improvement enterprise in that selected measures suggest areas of priority, having been rated as highly important, valid and amenable to improvement. In addition, our results guide the need for future research and measure refinement to ensure that data collection efforts yield measures of high value and little dispute among users. For example, ‘severe ROP’ did not meet the panel’s approval for inclusion in the composite, despite its clinical importance and prominence as a potential target for quality improvement. Panelists were concerned about transfer bias and lack of ascertainment after early discharge, which may provide undue credit to NICUs that transfer out their patients for higher level care or discharge them before the peak incidence of ROP. Concerns regarding transfer bias already led to efforts by CPQCC and the VON to ensure better linkage of patient outcomes with treating hospitals in order to avoid giving credit to hospitals that transfer their poor outcomes and punishing hospitals who receive them. However, our study indicates an urgent need to research the postdischarge conversion rate to severe ROP so that a true NICU-specific severe ROP rate can be calculated. The need for such longitudinal research also highlights the need for further integration of care services and the measurement thereof, so that quality of care can be evaluated and improved comprehensively.

It is notable that the most highly rated measure of quality, antenatal steroid administration, is really a measure of perinatal care quality. Panelists acknowledged this but affirmed neonatologists' responsibility to influence their obstetrical colleagues' care provision with respect to this therapy within their institutions. Some even suggested that an NICU's sphere of influence should extend beyond its own walls to include its referral network (currently, outborn infants are excluded from analysis for this measure).

The panelists' view is concordant with current health policy priorities, which aim at improving care coordination among specialties through ACOs (Accountable Care Organizations).[32] An ACO is a health systems model, which aims to integrate services along the continuum of care across different institutions and care settings. One way to promote the development of ACOs is to align quality measurement with underlying health policy intent. Specifically, longitudinal measurement of quality across different care settings may give an impetus to providers to coordinate high quality of care delivery for patients in which they share joint responsibility.

Measure domains

Safety and effectiveness were the primary domains of quality assigned to the selected nine measures. These results imply that in its first iteration, the Baby-Monitor will contain only two rather than all six of the Institute of Medicine's domains of quality. While safety and effectiveness reflect areas of health policy priority, our results highlight the need for additional research to develop new measures in other domains, or refine existing measures or data collection methods for existing measures. For example, one major concern regarding length of stay as an efficiency measure was the inability to assess the safety of earlier discharge due to the lack of data on postdischarge medical resource utilization. Once such issues are resolved, future revisions of the composite can accommodate additional measures of quality.

Limitations

Our findings should be viewed within the context of the study design. The Baby-Monitor is based on measures available through the CPQCC and the VON. Therefore, measures may not be entirely generalizable to other data sources. However, these consortia receive data from over 900 NICUs worldwide, representing a robust sample for indicator development. In addition, quality measures are very comparable to those collected nationally and internationally by other large consortiums, such as Pediatrix Medical Group or the Australian and New Zealand Neonatal Network.

Measure ratings may vary between and even within groups of experts. However, the high level of agreement between academic researchers and clinical neonatologists with regard to selecting measures of quality for a composite index of neonatal intensive care quality provides important face validity for the Baby-MONITOR. Although a response rate of close to 50% is common among physician surveys, we cannot exclude bias in our survey response. However, the direction of any potential bias is not determined easily.

Panelist discussions and ratings may be dominated by the most vocal participants. We attempted to minimize this effect in several ways. The first panel discussion was co-moderated by an independent researcher experienced in quality of care. In addition, we allotted time for additional comments and assigned each participant a group of measures for introduction to the group.

Our initial exclusion criteria should not be regarded as normative. It is our intent to develop a data set with the smallest degree of systematic bias against any individual hospital type. Any decision to include or exclude certain patients may lead to biases. In addition, data collection may change over time and richer data sets may allow for exclusion criteria to be altered. Moreover, some variation in inclusion criteria will not significantly alter NICU performance. We have shown that NICU performance ratings are largely insensitive to variations in definitions of mortality with regard to in/exclusion of delivery room deaths, deaths before 12 h of life, 28-day mortality and in-hospital mortality. While the positions of top and bottom performing hospitals are very stable most of the rank switching occurs in the middle tier.[33] These results are consistent with the findings of others.[26] Therefore, while one can reasonably disagree regarding the exact definitions of quality measures, their effect on comparative performance is often marginal.[33] Nevertheless, in the development of the Baby-MONITOR we will address the uncertainty in any given measure through sensitivity analysis at the stage of measure aggregation so that the extent of bias can be explored.[34]

Conclusion

In a modified Delphi experiment, a panel of 15 experts selected 9 of 28 measures of quality for inclusion into a composite indicator of neonatal intensive care quality delivered to VLBW infants. In future work, we will aggregate the individual measures and test whether the resulting composite is robust and valid. Our systematic andtransparent approach to indicator construction may serve as a template for developers in other health-care settings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

JP, JBG, JAFZ and LAP led the design, data analysis, interpretation of results and writing of the manuscript. JAFZ, JBG and ARS participated as panelists and helped with panelist recruitment. ARS also contributed to writing of the manuscript. KMW and MAK coordinated the study, prepared study materials, assisted with data acquisition and contributed to writing of the manuscript. MM and KP undertook data analysis and interpretation. EJT contributed to the study design, interpretation of results, writing of the manuscript and moderated panel discussions. JP is guarantor of the paper. We gratefully acknowledge the contributions and effort by all expert panelists listed here alphabetically: Judy L Aschner, MD (Vanderbilt University), Reese H Clark, MD (Pediatrix Medical Group), Edward F Donovan, MD (University of Cincinnati), William H Edwards, MD (Dartmouth University), Gabriel E Escobar (Kaiser Permanente Medical Care Program), Donald A Goldmann, MD (Harvard University and Institute for Healthcare Improvement), Jeffrey H Horbar, MD (Vermont Oxford Network and University of Vermont), Martin J McCaffrey, MD (University of North Carolina, Chapel Hill), Lu-Ann Papile, MD (Baylor College of Medicine), Roger F Soll, MD (Vermont Oxford Network and University of Vermont), Jon E Tyson, MD, MPH (University of Texas, Houston), Michele C Walsh, MD (Case Western University). The following experts co-authored this manuscript: Jeffrey B Gould, MD, MPH, Ann R Stark, MD, John A Zupancic, MD, ScD. We also acknowledge the contribution and support from Dr Carl Bose in facilitating the validation survey component of this research. Jochen Profit's contribution is supported in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development K23 HD056298-01 (PI Jochen Profit, MD, MPH). Dr Petersen was a recipient of the American Heart Association Established Investigator Award (Grant number 0540043N) at the time this work was conducted. Drs Petersen, Pietz, and Mr Mei also receive support from a Veterans Administration Center Grant (VA HSR&D CoE HFP90-20). Dr Thomas is supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development Grant K24 HD053771.

References

1. Kohn, LT.; Corrigan, JM.; Donaldson, MS. To Err Is Human. National Academy Press; Washington, DC: 1999. Institute of Medicine.
2. Institute of Medicine. Crossing the Quality Chasm: A New Health System for the 21st Century. National Academy Press; Washington, DC: 2001.
3. Horbar JD, Badger GJ, Lewit EM, Rogowski J, Shiono PH. Hospital and patient characteristics associated with variation in 28-day mortality rates for very low birth weight infants. Vermont Oxford Network. Pediatrics. 1997; 99:149–156. [PubMed: 9024438]
4. Sankaran K, Chien L-Y, Walker R, Seshia M, Ohlsson A, Lee SK. Variations in mortality rates among Canadian neonatal intensive care units. CMAJ. 2002; 166:173–178. [PubMed: 11826939]
5. Rogowski JA, Staiger DO, Horbar JD. Variations in the quality of care for very-low-birthweight infants: implications for policy. Health Aff. 2004; 23:88–97.
6. Morales LS, Staiger DO, Horbar JD, Carpenter J, Kenny M, Geppert J, et al. Mortality among very low-birthweight infants in hospitals serving minority populations. Am J Public Health. 2005; 95:2206–2212. [PubMed: 16304133]
7. Profit J, Zupancic JA, McCormick MC, Richardson DK, Escobar GJ, Tucker J, et al. Moderately premature infants at Kaiser Permanente Medical Care Program in California are discharged home earlier than their peers in Massachusetts and the United Kingdom. Arch Dis Child Fetal Neonatal Ed. 2006; 91:245–250.
8. Phibbs CS, Baker LC, Caughey AB, Danielsen B, Schmitt SK, Phibbs RH. Level and volume of neonatal intensive care and mortality in very-low-birth-weight infants. N Engl J Med. 2007; 356:2165–2175. [PubMed: 17522400]
9. Institute of Medicine. Rewarding Provider Performance: Aligning Incentives in Medicare. National Academy Press; Washington, DC: 2006.
10. Mattke S, Epstein AM, Leatherman S. The OECD Health Care Quality Indicators Project: history and background. Int J Qual Health Care. 2006; 18:1S–14.
11. Hibbard JH, Stockard J, Tusler M. Hospital performance reports: impact on quality, market share, and reputation. Health Aff. 2005; 24:1150–1160.
12. Epstein AM. Paying for performance in the United States and abroad. N Engl J Med. 2006; 355:406–408. [PubMed: 16870921]
13. Lindenauer PK, Remus D, Roman S, Rothberg MB, Benjamin EM, Ma A, et al. Public reporting and pay for performance in hospital quality improvement. N Engl J Med. 2007; 365:486–496. [PubMed: 17259444]
14. Petersen LA, Woodard LD, Urech T, Daw C, Sookanan S. Does pay-for-performance improve the quality of health care? Ann Intern Med. 2006; 145:265–272. [PubMed: 16908917]
15. Schoen C, Davis K, How SKH, Schoenbaum SC. U.S. health system performance: a national scorecard. Health Aff. 2006; 25:w457–w475.
16. Epstein AJ. Do cardiac surgery report cards reduce mortality? Assessing the evidence. Med Care Res Rev. 2006; 63:403–426. [PubMed: 16847071]
17. Grossbart SR. What's the return? Assessing the effect of 'pay-for-performance' initiatives on the quality of care delivery. Med Care Res Rev. 2006; 63:29S–248. [PubMed: 16688923]
18. Petra E, Varughese P, Epifania L, Buneo L, Scarfone K. Use of quality index tracking to drive improvement in clinical outcomes. Nephrol News Issues. 2006; 20:67–83. [PubMed: 16859235]
19. Peterson ED, Delong ER, Masoudi FA, O'Brien SM, Peterson PN, Rumsfeld JS, et al. 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]

20. Zaslavsky AM, Shaul JA, Zaborski LB, Cioffi MJ, Cleary PD. Combining health plan performance indicators into simpler composite measures. Health Care Financ Rev. 2002; 23:101–115. [PubMed: 12500473]

21. Kaplan, SH.; Normand, SL. Conceptual and Analytical Issues in Creating Composite Measures of Ambulatory Care Performance. Robert Wood Johnson Foundation; Washington, DC: 2006.

22. Profit J, Typpo KV, Hysong SJ, Woodard LD, Kallen MA, Petersen LA. Improving benchmarking by using an explicit framework for the development of composite indicators: an example using pediatric quality of care. Implement Sci. 2010; 5:13. [PubMed: 20181129]

23. Nardo, M.; Saisana, M.; Saltelli, A.; Tarantolo, S.; Hoffman, A.; Giovanini, E. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing; Paris, France: 2005. p. 1-08.

24. Sexton, BJ.; Grillo, S.; Fullwood, C.; Pronovost, PJ. Assessing and improving safety culture. In: Frankel, A.; Leonard, M.; Simmonds, T.; Haraden, C.; Vega, KB., editors. The Essential Guide for Patient Safety Officers. Joint Commission Resources with the Institute for Healthcare Improvement; Chicago, IL: 2009. p. 11-20.

25. Thomas EJ, Sherwood GD, Mulhollem JL, Sexton JB, Helmreich RL. Working together in the neonatal intensive care unit: provider perspectives. J Perinatol. 2004; 24:552–559. [PubMed: 15141266]

26. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? Med Care. 2005; 43:1177–1184. [PubMed: 16299428]

27. Halasyamani LK, Davis MM. Conflicting measures of hospital quality: ratings from ‘Hospital Compare’ versus ‘Best Hospitals’. J Hosp Med. 2007; 2:128–134. [PubMed: 17549759]

28. Brook, RH. The RAND/UCLA Appropriateness Method. Agency for Health Care Policy and Research; Rockville, MD: 1994.

29. Fitch, K.; Bernstein, SJ.; Aguilar, MS.; Burnand, B.; LaCalle, JR.; Lazaro, P., et al. The RAND/UCLA Appropriateness Method User’s Manual. RAND Corporation; Santa Monica, CA: 2001.

30. Leape, LL.; Hilborne, LH.; Kahan, JP.; Stason, WB.; Park, RE.; Kamberg, CJ., et al. Coronary Artery Bypass Graft: A Literature Review and Ratings of Appropriateness and Necessity. RAND Corporation; Santa Monica, CA: 1991.

31. Fitch K, Lazaro P, Aguilar MD, Kahan JP, van het LM, Bernstein SJ. European criteria for the appropriateness and necessity of coronary revascularization procedures. Eur J Cardiothorac Surg. 2000; 18:380–387. [PubMed: 11024372]

32. Kocher R, Sahni NR. Physicians versus hospitals as leaders of accountable care organizations. N Engl J Med. 2010; 363:2579–2582. [PubMed: 21067374]

33. Profit J, Zupancic JA, Gould JB, Pietz K, Petersen LA. NICU performance ratings are not sensitive to variations in definitions of mortality. E-PAS. 2009; 5510:5175.

34. Saisana M, Saltelli A, Tarantola S. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. J R Stat Soc Ser A. 2005; 168:307–323.

35. Peerzada JM, Richardson DK, Burns JP. Delivery room decision-making at the threshold of viability. J Pediatr. 2004; 145:492–498. [PubMed: 15480373]

36. Escobar GJ, McCormick MC, Zupancic JAF, Coleman-Phox K, Armstrong MA, Greene JD, et al. Unstudied infants: outcomes of moderately premature infants in the neonatal intensive care unit. Arch Dis Child Fetal Neonatal Ed. 2006; 91:F238–F244. [PubMed: 16611647]

37. Profit J, Petersen LA, McCormick MC, Escobar GJ, Coleman-Phox K, Zheng Z, et al. Patient-to-nurse ratios and outcomes of moderately preterm infants. Pediatrics. 2010; 125:320–326. [PubMed: 20064868]

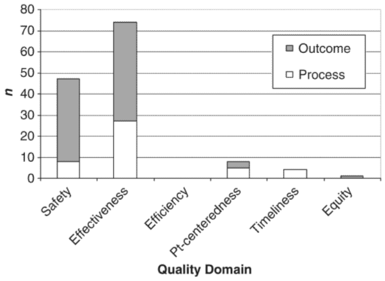

Figure 1.
Primary domain attribution of measures selected for the composite. Votes cast for quality domains for each of the nine measures selected for the composite; attribution of each measure to only one primary quality domain. No structural measures of quality were selected for the composite (*n* = 15 raters).

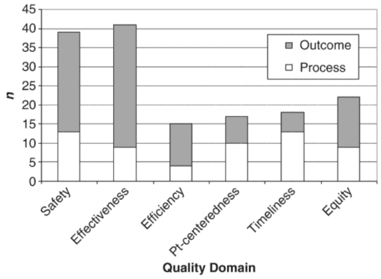

Figure 2.

Secondary attribution of measures selected for the composite. Votes cast for quality domains for each of the nine measures selected for the composite; attribution of each measure to multiple secondary domains permitted. No structural measures of quality were selected for the composite (*n* = 15 raters).

Table 1

Exclusion criteria and rationale

Exclusion criterion	Pros	Cons
<25 weeks gestational age at birth	Avoids bias from care variations at the threshold of viability. Most neonatologists resuscitate at ≥25 weeks[35]	Loss of information
>1500 g birth weight	Common criterion used for data collection. Fewer adverse outcomes in infants close to this threshold[36,37]	Favors institutions with more SGA infants
Major congenital anomalies	Avoids selection bias against NICUs that treat complex, high-risk infants	Loss of information
Death before 12 h of life	Avoids counting perinatally moribund patients. Variable admission procedures for moribund infants to NICU are avoided	Loss of information. Potentially, poor neonatal resuscitation rather than perinatal event responsible for early death
Transferred in after day of life 3[a]	Reduces systematic bias introduced by late transfer of infants. Adjustment for inborn status is made via risk adjustment	Loss of information. Potential of gaming by transferring out bad outcomes
Transfer out except for convalescent and chronic care, excludes readmissions[a]	Increases sample size compared with excluding all transfers. Most infants transferred from level III NICUs will be convalescing	Loss of information. May create bias toward transferring institutions. Requires editorial choices for certain outcomes that are ascertained late in the hospital course (CLD, ROP, LOS)

Abbreviations: CLD, chronic lung disease; LOS, length of stay; NICU, neonatal intensive care unit; ROP, retinopathy of pre-maturity; SGA, small for gestational age.

[a] ‘Present on admission’ codes introduced in 2008 may allow for inclusion of all transfers.

Table 2

Rating criteria for each measure

Criteria	Prompt/definition
Importance	Does this measure represent a priority or high impact aspect of healthcare? Are improvements in this measure important to the health outcomes of VLBW infants?
Reliability	Given that this measure is adequately abstracted, is it precisely defined, reproducible between and within raters, and reflective of action when implemented over time? Reliability is necessary but not sufficient for validity
Validity	Given that this measure is adequately abstracted, does this measure identify the true condition of the patient? Are exclusions and risk adjustment adequate?
Scientific soundness	Is scientific soundness evidenced in the literature, suggesting that a measure is a surrogate for quality of care and can be improved using QI methodology?
Usability	Does this measure provide information that is actionable (can be used to improve quality)? Is the information meaningful and understandable? Can real differences between NICUs be identified?
Overall score	What is your overall impression of this measure’s ability to help discriminate quality differences among NICUs? Is it a useful component of a composite indicator of quality? Can conclusions about quality be inferred from performance on this measure? The overall score is not derived from the individual domain ratings (e.g., it is not an average of previous ratings), but is a reflection of overall opinion and depends on your weighting of each domain

Abbreviations: NICU, neonatal intensive care unit; QI, quality improvement; VLBW, very low birth weight.

Table 3

Levels of measure quality for use in benchmarking

Level	Definition
1	Current evidence indicates that collaborative improvement efforts have used a measure successfully to facilitate improvements in quality of care and that this measure is ready for application in VLBW infant quality of care improvement
2	Current evidence is limited but indicates improvements in VLBW infant quality of care in small studies with the need of further research
3	Evidence exists for wide performance variation among NICUs but current evidence is limited for successful use in quality improvement efforts. Strategies for improvement are unclear or undefined, and more research is needed

Abbreviations: NICU, neonatal intensive care unit; VLBW, very low birth weight.

Table 4

Panel median scores from Rounds 1 and 2

Measures selected for the Baby-MONITOR sorted by median overall score in Round 2									
	Round 1							**Round 2**	
	Measure	**Importance**	**Reliability**	**Validity**	**Scientific soundness**	**Usability**	**Overall score (IQR)**	**Overall score (IQR)**	**Level rating**
2.	**Antenatal steroids**	9	9	8	9	9	**9 (1)**	**9 (0)**	1
6.	**Timely ROP exam**	9	9	9	9	9	**9 (1)**	**9 (0)**	3
25.	**Infection**	9	8	8	9	8	**8 (1)**	**9 (1)**	1
4.	**First hour hypothermia**	9	7	8	8	8	**8 (1)**	**8 (1)**	2
14.	**Pneumothorax**	8	7	7	8	6	**7 (2)**	**8 (2)**	2
24.	**Growth velocity**	8	7	7	7	7	**7.5 (2)**	**8 (2)**	2
17.	**Oxygen at 36 weeks**	8	8	7	7	7	**7 (3)**	**7 (2)**	2
23.	**Any human milk at discharge**	7.5	7	7	7	7	**7 (3)**	**7 (2)**	2
28.	**NICU mortality**	9	8	8	8	7	**7 (2)**	**7 (2)**	3
Measures NOT selected for the Baby-MONITOR sorted by median overall score in Round 2									
27.	**28-Day mortality**	9	8	8	7	7	**8 (2)**	**7 (2)**	3
15.	**Steroids for CLD**	9	8	6	7	7	**7 (3)**	**7 (2)**	3
5.	**Early surfactant**	8	9	7	8	7	**7 (3)**	**7 (3)**	1
3.	**First hour temperature measured**	9	8	8	7	7	**7 (4)**	**7 (4)**	2
10.	**Severe IH**	9	6	6	6	5	**6 (2)**	**6 (3)**	3
7.	**Severe ROP**	8	6	6	7	6	**6 (3)**	**6 (4)**	3
8.	**ROP surgery**	8	7	7	6	6	**6 (3)**	**5 (1)**	3
22.	**Only human milk at discharge**	6	7	6	5	5	**5 (1)**	**5 (2)**	2
21.	**NEC surgery**	8.5	8	6	5.5	5	**6 (1)**	**5 (2)**	3
20.	**NEC**	9	5.5	6	6	5	**6 (1)**	**5 (3)**	3
26.	**Length of stay**	7	8	5	5	6	**7 (2)**	**5 (3)**	2
1.	**VLBW volume**	4	9	4	4	4	**5 (3)**	**4 (2)**	3
12.	**Use of AV**	5	7	4	3	4	**4 (2)**	**4 (2)**	3
13.	**Duration of AV**	8	6	5	5	5	**5 (2)**	**4 (2)**	3
9.	**Any IH**	7	6	6	6	5	**5 (2)**	**4 (3)**	3
11.	**Cystic PVL**	9	4	5	6	4	**5 (3)**	**4 (3)**	3
19.	**Discharge on AV**	8	8	6	4	4	**5 (3)**	**4 (3)**	3

Measures selected for the Baby-MONITOR sorted by median overall score in Round 2

	Round 1							Round 2	
	Measure	**Importance**	**Reliability**	**Validity**	**Scientific soundness**	**Usability**	**Overall score (IQR)**	**Overall score (IQR)**	**Level rating**
18.	**Oxygen at discharge**	8	6.5	5	5	4	**5 (2)**	**4 (3)**	3
16.	**Oxygen on day 28**	5	5	3.5	4	4.5	**5 (2)**	**4 (4)**	3

Abbreviations: AV, assisted ventilation; CLD, chronic lung disease; IH, intracranial hemorrhage; IQR, interquartile range; NEC, necrotizing enterocolitis; NICU, neonatal intensive care unit; PVL, periventricular leukomalacia; ROP, retinopathy of pre-maturity; VLBW, very low birth weight.

Measure rating from 1 to 9 (9 is highest). Level attribution from 1 to 3 (1 is best). All ratings are expressed as medians, IQR. Bolded measures are selected for the composite indicator.