Title

Baby-MONITOR: A Composite Indicator of NICU Quality

Journal

Pediatrics, 134(1)

ISSN

0031-4005

Authors

Profit, Jochen

Kowalkowski, Marc A

Zupancic, John AF

et al.

Publication Date

2014-07-01

DOI

10.1542/peds.2013-3552

Peer reviewed

Baby-MONITOR: A Composite Indicator of NICU Quality

WHAT'S KNOWN ON THIS SUBJECT: The traditional process-focused approach to quality improvement has not remedied NICUs' inconsistency in quality of care delivery across clinically important measures. Global measurement of quality may induce broad, systems-based improvement, but must be formally studied.

WHAT THIS STUDY ADDS: We present a systematically developed and robust composite indicator, the Baby-MONITOR, to assess the quality of care delivered to very low birth weight infants in the NICU setting.

abstract

BACKGROUND AND OBJECTIVES: NICUs vary in the quality of care delivered to very low birth weight (VLBW) infants. NICU performance on 1 measure of quality only modestly predicts performance on others. Composite measurement of quality of care delivery may provide a more comprehensive assessment of quality. The objective of our study was to develop a robust composite indicator of quality of NICU care provided to VLBW infants that accurately discriminates performance among NICUs.

METHODS: We developed a composite indicator, Baby-MONITOR, based on 9 measures of quality chosen by a panel of experts. Measures were standardized, equally weighted, and averaged. We used the California Perinatal Quality Care Collaborative database to perform cross-sectional analysis of care given to VLBW infants between 2004 and 2010. Performance on the Baby-MONITOR is not an absolute marker of quality but indicates overall performance relative to that of the other NICUs. We used sensitivity analyses to assess the robustness of the composite indicator, by varying assumptions and methods.

RESULTS: Our sample included 9023 VLBW infants in 22 California regional NICUs. We found significant variations within and between NICUs on measured components of the Baby-MONITOR. Risk-adjusted composite scores discriminated performance among this sample of NICUs. Sensitivity analysis that included different approaches to normalization, weighting, and aggregation of individual measures showed the Baby-MONITOR to be robust (r = 0.89–0.99).

CONCLUSIONS: The Baby-MONITOR may be a useful tool to comprehensively assess the quality of care delivered by NICUs. *Pediatrics* 2014;134:74–82

AUTHORS: Jochen Profit, MD, MPH,[a,b] Marc A. Kowalkowski, PhD,[c] John A. F. Zupancic, MD, ScD,[d,e] Kenneth Pietz, PhD,[f,g] Peter Richardson, PhD,[f,g] David Draper, PhD,[h,i] Sylvia J. Hysong, PhD,[d,e] Eric J. Thomas, MD, MPH,[j] Laura A. Petersen, MD, MPH,[d,e] and Jeffrey B. Gould, MD, MPH[a,b]

[a]Perinatal Epidemiology and Health Outcomes Research Unit, Division of Neonatal and Developmental Medicine, Department of Pediatrics, Stanford University School of Medicine and Lucile Packard Children's Hospital; Palo Alto, California; [b]California Perinatal Quality Care Collaborative; Palo Alto, California; [c]Levine Cancer Institute, Carolinas HealthCare System, Charlotte, North Carolina; [d]Department of Neonatology, Beth Israel Deaconess Medical Center, Boston, Massachusetts; [e]Division of Newborn Medicine, Harvard Medical School, Boston, Massachusetts; [f]Section of Health Services Research, Department of Medicine, Baylor College of Medicine, Houston, Texas; [g]Houston Veterans Affairs (VA) Health Services Research and Development Center of Excellence, Health Policy and Quality Program, Michael E. DeBakey VA Medical Center; Houston, Texas; [h]Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, Santa Cruz, California; [i]eBay Research Labs, San Jose, California; [j]University of Texas at Houston – Memorial Hermann Center for Healthcare Quality and Safety, University of Texas Medical School, Houston, Texas

KEY WORDS

infant, newborn, quality of care, performance measurement

ABBREVIATIONS

CPQCC—California Perinatal Quality Care Collaborative
VLBW—very low birth weight

Dr Profit acquired funding for this study, conceptualized and designed the study, selected data for inclusion in analyses, analyzed the data, assisted with interpretation of the results, and drafted the initial manuscript; Dr Kowalkowski analyzed the data, assisted with interpretation of the results, and revised the manuscript; Dr Zupancic conceptualized and designed the study, selected data for inclusion in analyses, assisted with interpretation of the results, and revised the manuscript; Dr Pietz helped acquire funding for the study, conceptualized the study, analyzed the data, assisted with interpretation of the results, and revised the manuscript; Dr Richardson analyzed the data, assisted with interpretation of the results, and revised the manuscript; Dr Draper assisted with designing the analysis and interpretation of the results and revised the manuscript; Dr Hysong conceptualized and helped design the study, assisted with interpretation of the results, and revised the manuscript; Dr Thomas helped acquire funding for the study, conceptualized and helped design the study, assisted with interpretation of the results, and revised the manuscript; Dr Petersen helped acquire funding for the study, conceptualized and helped design the study, assisted with interpretation of the results, and revised the manuscript; Dr Gould helped acquire funding for the study, conceptualized and helped design the study, selected data for inclusion in analyses, assisted with interpretation of the results, and revised the manuscript; and all authors approved the final manuscript as submitted.

(Continued on last page)

Neonatal intensive care is a complex and multidimensional activity, which the measurement of its quality should reflect. There is value in summarizing performance by combining the information from multiple measures, as such a summary can convey quality from many different perspectives.[1] The Institute of Medicine noted that composite measures can enhance measurement to extend beyond tracking performance on individual measures, and can provide a potentially deeper view of the reliability of the care system.[2] A multidimensional measure can convey quality from many different perspectives and may provide new insights into effective improvement strategies.

The National Quality Forum defines composite measures as "a combination of two or more individual measures into a single measure that results in a single score."[1] They are created by compiling individual measures into a single indicator, on the basis of an underlying model of the multidimensional concept that is being measured.[3] Their primary appeal is the ability to simplify and summarize otherwise complex issues, and to provide global insights and trends about quality of care.

On the other hand, composite indicators may be susceptible to unsound conceptual or statistical approaches, and they may be less transparent than individual measures of quality. Therefore, the construction of composites requires that explicit and transparent methods are used to ensure conceptual and statistical soundness,[4] so that they do not (1) fall short of their aim to improve quality of care, (2) fail to elicit buy-in from providers, (3) misclassify providers as outliers, or (4) encourage overly simplistic conclusions.

In this article, we describe the construction of a composite indicator of quality of care delivered to very low birth weight (VLBW; <1500 g) infants, building on previous work. We have coined the term Baby-MONITOR (Measure Of Neonatal InTensive care Outcomes Research) for the instrument,[14] which we present as a prototype for the next generation of quality assessment. Our primary objective was to test whether the Baby-MONITOR would discriminate global NICU performance on quality of care delivery. Our secondary objective was to test the robustness of the Baby-MONITOR.

METHODS

Overview

We followed a systematic and explicit approach based on recommendations by the European Commission Joint Research Center and the Organization for Economic Cooperation and Development, thought leaders in this area.[3,4,15] Preliminary steps in the development of the Baby-MONITOR included development of a theoretical framework[4,15]; expert-informed selection of its measure components[14,16]; initial data analysis (1) to investigate the completeness of the data, (2) to develop and test adequate measure definitions and restrictions, and (3) to minimize systematic selection and transfer biases; and construction of risk adjustment models.[17]

With these building blocks in place, we standardized and risk-adjusted outcomes, weighted the individual components, and aggregated measures to form a composite indicator. After defining a base-case composite, we evaluated its sensitivity to the underlying assumptions and explored the effects of alternative computational approaches.

Sample

Patients

Patient data for this analysis were obtained from the California Perinatal Quality Care Collaborative (CPQCC). Local NICU personnel are trained to abstract data. Annual training sessions help to promote accuracy and uniformity in data abstraction. Each record has range and logic checks both at the time of data collection and data closeout, with auditing of records with excessive missing data. A detailed description of the patient-selection criteria has been published elsewhere.[14] In brief, the goal was to create a sample of VLBW infants that would represent the "common" preterm infant. For this study, 9023 unique VLBW infants cared for at 22 California regional NICUs between 2004 and 2010 met the inclusion criteria. Of these centers, 15 are designated as level 4 (access to pediatric surgical subspecialists) and the remainder as level 3.[18] We used multiyear analyses because of the small number of VLBW infants cared for in some institutions.

To ensure that patient outcomes reflected the quality of care of the NICU under observation, we excluded infants who died before 12 hours of life, those transferred in after 3 days of age, those transferred out for reasons other than convalescent and chronic care, and those who had severe congenital anomalies. Finally, to avoid systematic bias based on decisions to withhold resuscitation at the threshold of viability, we restricted the analysis to infants born after 24 completed weeks of gestation.

Measures

Dependent Variables

Quality-of-care measures: Measures were selected by an expert panel via a formal modified Delphi process,[4] and subsequently affirmed by a sample of practicing neonatologists.[16] Measure definitions were derived from standard CPQCC/Vermont Oxford Network (VON) algorithms. The measures were expressed as binary variables at the

patient level and as proportions at the unit level. They included: (1) any antenatal steroid administration; (2) moderate hypothermia (<36°C) on admission; (3) non–surgically-induced pneumothorax; (4) health care-associated bacterial or fungal infection; (5) chronic lung disease (oxygen requirement at 36 weeks' gestational age); (6) timely eye exam (retinopathy of prematurity screening at the age recommended by the American Academy of Pediatrics); (7) discharge on any human breast milk; (8) mortality during the birth hospitalization, and (9) growth velocity (less or more than the median of 12.4 g/kg/day). Growth velocity was determined according to a logarithmic function described by Patel.[19] We aligned variables so that a higher value represented a better outcome. Other restrictions with regard to transfers and hospital of birth are described elsewhere.[14]

Independent Variables

We applied CPQCC standard operational definitions for all independent variables, including gender, weight for gestational age below the 10th percentile, birth outside a regional center, and Cesarean birth. Gestational age at birth was categorized into 25 weeks to 27 weeks, 6 days; 28 weeks to 29 weeks, 6 days; and 30 weeks or more gestation groups, based on similar patient numbers among groups. Apgar score was categorized as 3 or below, between 4 and 6, and above 6.

Analyses

Standardization/Risk-Adjustment

Because some NICUs will have higher morbidity and mortality rates simply because they care for sicker infants, we developed risk-adjustment models[20–22] for all measures, except the eye examination (which is a process that should be performed on all infants independent of illness severity at birth; Section 1 in the Supplemental Web Appendix gives additional details on the risk-adjustment model). Variables included in these models include a combination of prenatal care, gestational age at birth, small for gestational age status, multiple birth, cesarean delivery, inborn or outborn, and 5-minute Apgar score.[17]

We used the Draper-Gittoes[21] method of risk adjustment, which has long been used successfully in the UK higher education system. With this method, a standardized *z* score is constructed that is suitable for combining via unweighted or weighted averaging. These *z* scores should be approximately normally distributed with mean 0 and SD 1. Additional details are available in the Supplemental Web Appendix, Section 2.

Weighting

We adopted an equal weighting scheme for the base case. In sensitivity analyses, we explored a variety of weighting schemes based on expert opinion. Our panel of experts[14] was asked to distribute 100 imaginary dollars across the measures, according to the relative contribution of each measure to overall NICU quality (see Table 1). Mean and median weights derived from this exercise were applied in sensitivity analyses.

Aggregation and Discrimination

Measures were aggregated by averaging the 9 *z* scores for each NICU. The 95% confidence intervals were computed via bootstrapping[23] (a simulation in which each NICU's patients are resampled with replacement 1000 times) and are plotted in Fig 1; failure of 2 such intervals to overlap corresponds to a highly statistically significant difference. This criterion was used to discriminate between NICUs. In addition, we assigned star ratings to groups of NICUs depending on whether their entire confidence interval fell below 0 (3 stars), overlapped 0 (4 stars), or above 0 (5 stars).

Sensitivity Analysis/Robustness

We investigated the effects of our methodological choices on the composite score by varying measure weights and by alternative methods of measure aggregation. In the base case all measures were equally weighted and averaged; this is an easily understood format. However, different approaches are possible, and if they result in substantially different performance assessments, might be favored on theoretical grounds.

We tested several alternatives to weighting, including using the mean and median weights derived by our group of experts (median weights are less prone to outlier opinions). In addition, rather than adding the *z* scores, we assigned a rank to each NICU and then added the ranks across different measures. This method may provide more separation between NICUs,

TABLE 1 Comparison of Median and Mean Expert Weights Assigned to Each Measure

Measure	Median (IQR)	Mean (SD, Range)
Antenatal steroids	20 (10)	16.5 (5.8, 5–25)
Health care-associated infection	15 (5)	16.5 (5.5, 5–25)
Survival	8 (15)	12.8 (11.7, 4–40)
Growth velocity	10 (10)	10.1 (5.5, 3–20)
Hypothermia on admission	9 (10)	10.1 (6.5, 3–25)
Discharge on any human milk	10 (6)	9.7 (5.4, 3–20)
Timely eye examination	9 (5)	9.1 (5.4, 3–20)
Chronic lung disease	5 (10)	8.9 (6.3, 0–25)
Pneumothorax	5 (3)	6.3 (2.4, 3–11)

IQR, interquartile range.
Weights derived using the Budget Allocation Technique described in the main paper.[2]

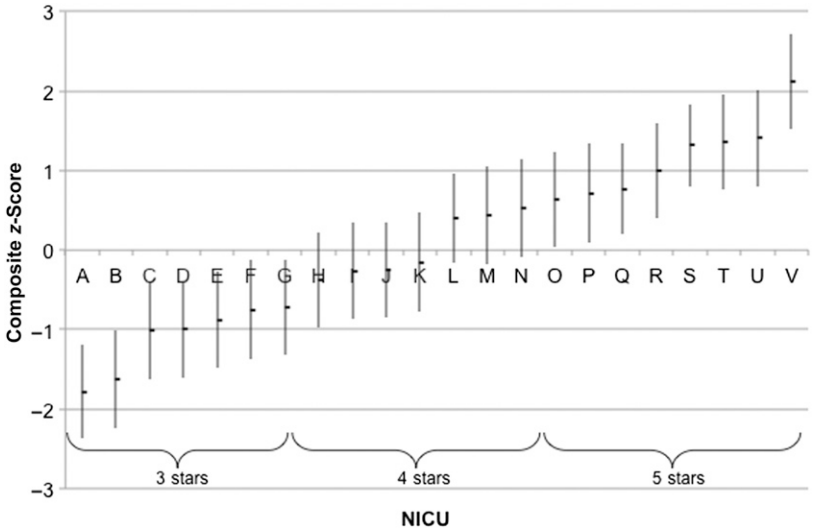

FIGURE 1
The base case is obtained by averaging the *z* scores of quality measures for each NICU (see Table 2) 2004 to 2007 data. Note: failure of 2 95% intervals to overlap corresponds to a statistically significant difference at approximately the 99% level.

although one must bear in mind that a 1-place difference in ranks could reflect large or small difference in *z* scores.

We also explored a different aggregation method by multiplying rather than adding the *z* scores. Geometric (multiplicative) aggregation is theoretically appealing. Whereas linear addition of measures allows NICUs to fully trade off low performance in 1 measure with high performance in another, multiplicative aggregation allows only for partial compensation. Consider a scenario including 2 NICUs and 2 quality measures. Suppose NICU A achieves a score of 1 on 1 measure and a score of 9 on the other, whereas NICU B achieves a score of 5 on both. Under additive aggregation both NICUs perform equally (9 + 1 = 5 + 5). However, the extreme performance of NICU A results in a much lower rating under multiplicative aggregation (9 < 25). Multiplicative aggregation is thus intriguing for settings where policymakers aim to promote broad standards of care.

We assessed robustness (our use of the term *robustness* in this article is synonymous with the term *stability*) of NICU performance under these different scenarios with the base case using both Pearson and Spearman rank correlation coefficients. Correlation coefficients >0.7 imply strong correlation.[24] See Supplemental Web Appendix, Section 3, for additional details.

Longitudinal Analysis

We evaluated stability over time of the base case as follows: the measures were generated separately using data from 2004 to 2007 and 2008 to 2010, and the results were compared. Parry[24] showed that mortality alone was not a good indicator of quality in that NICUs tended to bounce between top and bottom performance. Because we do expect some drift in quality of care over time, our analysis aimed to find moderate correlations of performance between the 2 time periods.

In 2004 to 2007, 22 NICUs, and in 2008 to 2010, 21 NICUs, met the criteria for inclusion, so the analyses were done on these units. First, we conducted Pearson and Spearman correlation analyses across the 2 time periods. Second, we applied the nonparametric Wilcoxon signed-rank test to examine the extent, if any, of the temporal differences. We also looked at the number of NICUs that changed quartiles between the 2 time periods. Finally, we computed the κ statistic for the quartiles.

Human Subjects Compliance

This study was approved by the CPQCC and the Baylor College of Medicine Internal Review Board.

RESULTS

Patient Characteristics

Table 2 shows the population and NICU characteristics for the combined sample, as well as for the 2 study periods 2004 to 2007 and 2008 to 2010. Of note is the improvement in absolute and risk-adjusted components of the Baby-MONITOR between the time periods in all measures, except for the rate of occurrence of pneumothoraces (which roughly held constant).

Performance on Individual Measures of Quality

Table 3 shows the standardized *z* scores for the clinical measures, with units ordered with regard to ascending composite score. The variation in performance within and between these regional NICUs is notable (see Supplemental Web Appendix, Section 4): it would be difficult to draw inferences on overall performance based on any individual measure of quality.

Base Case Baby-MONITOR

The base-case composite indicator was derived by averaging the *z* scores from each NICU; measures were assigned equal weights. Performance on the Baby-MONITOR is not an absolute marker of quality but indicates

TABLE 2 Characteristics of Infants and NICUs

Characteristic	2004 to 2007		2008 to 2010	
	Infant Level *n* = 5444	NICU Level *n* = 22	Infant Level *n* = 3579	NICU Level *n* = 21
	n (%) or Mean (SD)	Mean (SD)	*n* (%) or Mean (SD)	Mean (SD)
Gestational age at birth, wks	28.5 (2.4)	28.6 (0.3)	28.6 (2.4)	28.5 (0.4)
Birth weight, g	1092 (259)	1095 (25.7)	1096 (263)	1096 (35.8)
Small for gestational age	1526 (28.0)	28.3 (6.3)	1052 (29.4)	28.5 (5.7)
Female	2670 (49.0)	49.0 (4.8)	1753 (49.0)	48.0 (4.0)
Apgar score at 5 min	6.6 (1.1)	6.6 (0.2)	6.5 (1.2)	6.5 (0.3)
Cesarean delivery	3994 (73.4)	72.4 (7.4)	2714 (75.8)	75.6 (7.9)
Multiple gestation	1641 (30.2)	28.5 (8.0)	1066 (29.8)	27.8 (10.8)
Any prenatal care	5183 (96.0)	95.5 (4.2)	3447 (96.9)	96.3 (3.4)
Infant race/ethnicity				
Non-Hispanic black	646 (11.9)	12.2 (8.3)	411 (11.5)	11.8 (8.5)
Non-Hispanic white	1721 (31.6)	28.4 (14.0)	1067 (29.8)	26.7 (14.4)
Hispanic	2343 (43.0)	45.7 (20.4)	1527 (42.7)	46.6 (20.2)
Asian	587(10.8)	10.0 (7.3)	442 (12.4)	10.7 (9.0)
Other	147 (2.7)	3.6 (6.1)	132 (3.7)	4.3 (7.1)
VLBW admissions		247 (161)[a]		170 (96)[a]
Inborn babies	3838 (70.5)	72.2 (35.5)[a]	2579 (72.1)	71.1 (35.9)[a]
Baby-MONITOR quality measures, (*n* varies owing to definitions)	*n* (%) (based on measure definition)	Risk-adjusted mean (SD)[b]	*n* (%) (based on measure definition)	Risk-adjusted mean (SD)[b]
Survival of birth hospitalization	4954 (94.8)	94.2 (3.4)	3311 (95.2)	94.7 (3.5)
Any antenatal steroids	4269 (79.6)	78.0 (12.4)	2989 (84.3)	83.8 (9.9)
No hypothermia on admission	2227 (81.3)	78.6 (11.5)	3052 (86.0)	84.2 (13.1)
No pneumothorax	5243 (96.5)	96.4 (1.9)	3421 (95.9)	95.8 (1.9)
No health care-associated infection	4273 (81.3)	82.1 (6.4)	3033 (87.4)	86.7 (4.3)
Timely eye examination[c]	3844 (70.8)	68.2 (12.5)	2696 (75.6)	73.8 (10.4)
High growth velocity	2355 (50.6)	51.4 (14.1)	1785 (55.8)	54.6 (13.2)
No chronic lung disease	3606 (74.6)	74.6 (9.7)	2579 (79.1)	77.5 (9.6)
Any human breast milk at discharge	2345 (59.5)	61.1 (18.9)	1847 (64.1)	63.0 (17.7)

All NICU-level data are statistics of NICU means or proportions except for admissions.

[a] Annual mean (SD) of number of infants per hospital.

[b] Risk-adjusted percentages (SD).

[c] Not risk-adjusted. See Supplemental Web Appendix, Section 1 for details on risk adjustment.

overall performance relative to that of the other NICUs (Fig 1). NICUs can evaluate their absolute performance by investigating the composite's individual components.

Several observations can be made regarding California regional NICUs:

1. Considerable variation, evinced by non-overlapping confidence intervals, exists.
2. The composite scores for NICU V are significantly better than for NICUs A to Q.
3. The scores for NICUs A to G are below 0 (lower than expected), include 0 for NICUs H to N, and are above 0 (better than expected) for NICUs O to V.

A classification system was derived based on item 3 above, in which NICUs A to G are assigned 3 stars, NICUs H to N are assigned 4 stars, and NICUs O to V are assigned 5 stars. Special recognition was awarded to NICU V as the top performer, with its composite score exceeding the upper limit of the next best NICU's 95% confidence interval.

Sensitivity/Robustness Analyses

Figure 2 presents composite scores for the 22 NICUs based on 5 methods of weighting and aggregation across outcomes, including the base case (equal weights), mean and median expert weights, the base case using ranks, and the geometric mean (Supplemental Web Appendix Table B provides additional detail). Results are presented as ranks and show that the base case exhibits a high degree of stability. The Pearson and Spearman correlations between the base case and the other 4 approaches in Supplemental Table B varied from 0.89 to 0.99.

However, the figure also juxtaposes interesting characteristics of multiplicative aggregation in the geometric composite vis-à-vis additive aggregation in the base case. Note that NICU K, ranked 12th in the base case, was second lowest using the geometric approach. Table 3 shows that this NICU had the lowest score in the "no hypothermia" measure and the highest in the "timely eye exam"

TABLE 3 Standardized *Z* Scores for the Individual Measures

NICU	Survival	ANS	Not Cold	No CLD	No PTX	No HAI	High GV	Dc on any HM	Timely EE
A	−2.79	−6.11	−0.14	−0.79	0.63	−1.78	−0.76	2.59	−6.86
B	−2.62	0.27	−0.61	−0.40	−0.67	−1.38	−4.39	0.49	−5.31
C	1.14	−1.59	0.24	−2.89	−0.46	−0.34	0.90	−7.08	1.01
D	−0.96	−3.13	1.04	−0.87	−1.49	−2.29	−4.50	5.89	−2.65
E	−0.96	−5.75	0.04	−1.22	−2.20	−0.40	4.60	−1.49	−0.50
F	−0.22	0.07	−1.39	−4.24	−0.95	0.45	−0.61	3.60	−3.45
G	−0.23	−4.79	2.00	−4.15	−0.41	0.84	−0.75	1.96	−0.90
H	−0.09	−2.01	−2.15	6.56	−0.45	−5.44	−5.95	0.80	5.36
I	0.72	2.23	2.97	−0.04	−1.73	0.20	1.09	−1.65	−6.11
J	−1.34	0.95	0.61	4.06	0.02	0.59	−0.77	−5.03	−1.34
K	−0.97	−1.90	−9.21	4.26	0.58	−2.52	3.91	−4.34	8.80
L	0.57	2.66	−3.77	−0.51	2.60	−0.21	5.51	−6.97	3.69
M	4.36	−1.31	2.50	0.70	1.46	1.05	0.13	−4.45	−0.53
N	1.54	−1.13	−3.90	0.56	−0.26	3.40	0.54	−0.50	4.51
O	−0.41	5.04	6.93	−2.14	1.05	−2.39	−2.05	2.45	−2.72
P	−1.39	−0.10	0.61	−1.61	−0.38	2.29	2.84	3.16	1.94
Q	2.92	5.07	1.65	−5.88	1.52	−2.00	−1.72	−0.23	5.57
R	0.40	2.34	3.05	−2.21	1.75	1.81	−0.75	2.70	−0.12
S	−0.17	0.42	−0.98	3.62	1.56	2.53	0.10	5.28	−0.49
T	1.55	3.90	2.61	−2.52	0.03	2.85	1.94	4.46	−2.53
U	−0.16	6.12	5.04	3.95	−0.34	−0.83	−2.70	1.49	0.10
V	−0.38	1.02	−2.30	4.92	−0.37	7.74	6.00	0.08	2.42

Values above 1.96 and below −1.96 are significant at the $P < .05$ level.

ANS, antenatal steroids; CLD, chronic lung disease; cold, hypothermia (<36.0°C) on admission; Dc, discharge; EE, eye examination; GV, growth velocity; HAI, health care-associated infection; HM, human breast milk; PTX, pneumothorax

measure; because the geometric composite rewards stable performance across all measures, NICU K is unable to compensate for low performance on 1 measure with good results on others.

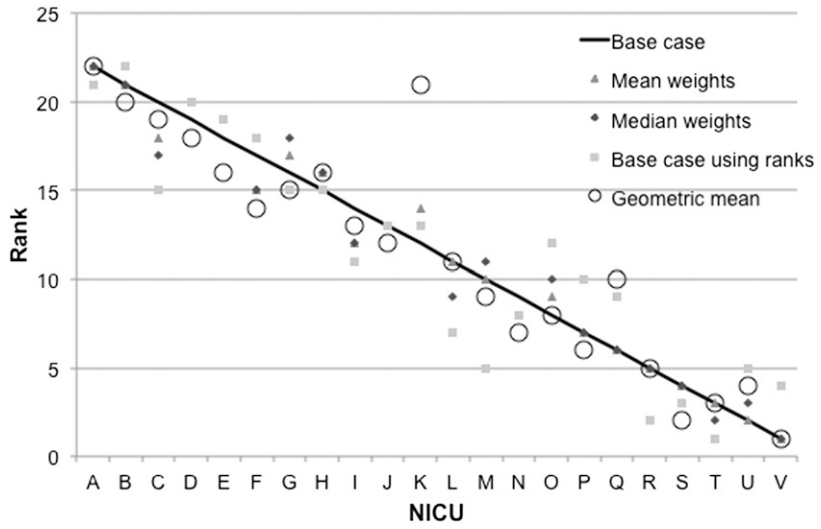

FIGURE 2

The difference of each symbol from the trend line of the base case is the key purpose of this graph. NICUs are ordered according to the base case rank. For each NICU, ranks were computed using 5 different weighting and aggregation schemes. The base case uses equal weighting and additive aggregation of *z* scores. In addition, we used mean and median expert weights, ranks rather than *z* scores for aggregation, and multiplicative aggregation (see Supplemental Web Appendix, Section 4, Table B for numerical detail). If <5 symbols are displayed per NICU this is attributable to overlap. NICU ranks for all schemes are highly correlated ($r = 0.89–0.99$). However, of particular interest is the comparison of the base case using additive aggregation (trend line) and the geometric mean using multiplicative aggregation (large white circles). Geometric aggregation penalizes NICUs with extreme performance (NICU K drops from rank 12 to rank 21).

Robustness of the Baby-MONITOR Over Time

The Pearson correlation between base-case composites derived from 2004 to 2007 and 2008 to 2010 data was 0.67; the Spearman (rank) correlation was 0.74. The *P* value for the Wilcoxon signed-rank test was .68, providing support for the null hypothesis that there was no difference between the 2 time periods. Of the 21 NICUs, 3 changed quartiles in the positive direction and 3 in the negative direction. The κ statistic was 0.43, indicating moderate agreement.[25]

DISCUSSION

We present the first iteration of the Baby-MONITOR, a composite indicator of neonatal intensive care quality provided to VLBW infants. The development of the Baby-MONITOR followed a formal, stepwise, and explicit process that has been peer reviewed and is widely applied in health and non-health settings.[3,26–28]

In previous work, we developed a theoretical framework for the Baby-MONITOR,[4,15] selected measures of quality

using rigorous methods,[14] validated the selection,[16] conducted initial data analyses, and developed risk models for individual measures.[17,29] In this study, we aggregated the measures, assessed the composite's ability to discriminate among NICUs, and conducted extensive sensitivity analyses with regard to weighting and aggregation. We found the Baby-MONITOR to be a robust measure and able to discriminate overall quality of care delivery.

Composite measurement of quality is becoming more prominent in health care and is being used to support consumer choices. Already, composites such as the *US News and World Report* hospital rankings exert great influence on hospital strategic planning and marketing. Whether they accurately discriminate higher from lower performing hospitals is less evident. In fact, several studies have highlighted the variable nature of performance ratings according to different methods.[30,31] Such divergence can be addressed only by adopting explicit and transparent standards for indicator development.[3,32]

In this study, we demonstrated high correlation with alternative methods of composite construction. We therefore retained the base case with equal weighting and additive aggregation. Clinicians may think that equal weighting poorly reflects quality priorities in the NICU, as few would consider mortality on par with hypothermia on admission. However, an equal weighting scheme is supported by the literature, which shows that unit weights would have to differ dramatically to substantially affect performance assessments.[33] Figure 2 confirms this literature and supports our decision for equal weighting, as it demonstrates little effect of various weighting schemes on NICU performance.

With regard to aggregation, we decided against grouping measures into subdimensions despite our previous research showing that the 9 measures assessed only 4 latent factors.[17] Grouping would add to the composite's complexity and require decisions about weighting within and between subindicators. If replicated in larger datasets we may revise the Baby-MONITOR, but for this initial iteration, we selected a simpler format.

We were intrigued by the results generated by the geometric composite. In the base case, the additive computation allows a high score in 1 measure to cancel a low score in another; this compensation does not occur in the geometric composite. In our sample, NICU K would have been classified as a 3-star rather than a 4-star NICU under multiplicative aggregation, revealing its quality deficits in several domains. Arguably, from the standpoint of policymakers, achievement of a certain performance benchmark across all aspects of care, as promoted by the multiplicative composite, may be more desirable. A multiplicative approach to measure aggregation may also be better aligned with the premise of composite measurement with regard to its role in incentivising systems-based multidimensional improvement. However, before a recommendation to use multiplicative rather than additive aggregation can be made, these findings require affirmation in a larger, more diverse sample of NICUs, as well as validation against other indicators.

Users of the Baby-MONITOR must respect its limitations and recognize that this initial iteration is merely a first step toward comprehensively measuring NICU quality of care delivery. Interpretation of composite ratings must take into account that absolute differences may be small or not statistically significant. This is a particular concern for rank-based performance assessments that include only point estimates of ranks. We therefore chose to present the Baby-MONITOR results as a caterpillar plot, and for further user-friendliness included a star rating based on normative criteria (ie, overall performance falling below, meeting, or exceeding expectations), even though simple star ratings may overstate quality differences and must be interpreted with caution. We suggest that the composite be used to generate system-based improvement efforts that "lift the boat" on multiple measures simultaneously. Such efforts should accompany, not replace, traditional quality improvement efforts.

One important concern for the Baby-MONITOR is validity, that is, does it measure overall NICU quality? Support for its validity has several sources. Content validity (the measure represents all facets of the underlying construct) was conferred by a panel of independent experts in the original measure selection process[14] and strengthened by a sample of recognized clinicians.[16] Construct validity (the composite actually measures what it is supposed to measure) is supported by formally including each measure's reliability, validity, importance, scientific soundness, and usability in the selection process. In addition, each component of the Baby-MONITOR achieves statistical separation of NICUs and so does the composite. Nevertheless, given the absence of a gold standard comparison, additional research is needed to further solidify construct validity, including comparison with other measures of quality. We are currently investigating convergent validity of the Baby-MONITOR and NICU safety culture. In addition, future research will need to address whether performance on the Baby-MONITOR correlates with long-term infant outcomes (predictive validity).

For reasons of sample size, we combined 4 years of data to generate the initial estimates for the Baby-MONITOR.

For a composite to meet the needs of clinicians at the frontline, additional research will need to focus on generating real-time estimates of the composite results based on moving averages or Bayesian updating methods.

Finally, although our current analysis is restricted to regional, or mostly level 4, NICUs, recent improvements in data collection, with linkage of outcomes across hospitals and post-discharge outcomes, will allow us to generalize assessment beyond the regional NICUs to the larger universe of lower level NICUs in California and beyond.

CONCLUSIONS

We present the first iteration of the Baby-MONITOR and display information regarding its ability to discriminate quality of care and robustness to different assumptions in its construction and over time.

ACKNOWLEDGEMENTS

We thank Aloka Patel and Rush University Medical Center for granting Dr Profit a nonexclusive license to use Rush's exponential infant growth model for noncommercial research purposes.

REFERENCES

1. NQF. Composite Measure Evaluation Framework and National Voluntary Consensus Standards for Mortality and Safety–Composite Measures: A Consensus Report. National Quality Forum. Washington, DC; 2009
2. Institute of Medicine. *Rewarding Provider Performance: Aligning Incentives in Medicare*. Washington, DC: National Academy Press; 2006
3. Nardo M, Saisana M, Saltelli A, Tarantolo S, Hoffman A, Giovanini E. *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris, France: OECD Publishing; 2005
4. 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
5. Medicare. Hospital Compare. 2013 [cited March 31, 2013]. Available at: www.medicare.gov/hospitalcompare/
6. Bradley EH, Herrin J, Elbel B, et al. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. *JAMA*. 2006;296(1):72–78
7. Fonarow GC, Abraham WT, Albert NM, et al; OPTIMIZE-HF Investigators and Hospitals. Association between performance measures and clinical outcomes for patients hospitalized with heart failure. *JAMA*. 2007; 297(1):61–70
8. Jha AK, Orav EJ, Li Z, Epstein AM. The inverse relationship between mortality rates and performance in the Hospital Quality Alliance measures. *Health Aff (Millwood)*. 2007;26(4):1104–1110
9. Werner RM, Bradlow ET. Relationship between Medicare's hospital compare performance measures and mortality rates. *JAMA*. 2006;296(22):2694–2702
10. Meehan TP, Fine MJ, Krumholz HM, et al. Quality of care, process, and outcomes in elderly patients with pneumonia. *JAMA*. 1997;278(23):2080–2084
11. Kanwar M, Brar N, Khatib R, Fakih MG. Misdiagnosis of community-acquired pneumonia and inappropriate utilization of antibiotics: side effects of the 4-h antibiotic administration rule. *Chest*. 2007;131(6): 1865–1869
12. Houck PM, Bratzler DW, Nsa W, Ma A, Bartlett JG. Timing of antibiotic administration and outcomes for Medicare patients hospitalized with community-acquired pneumonia. *Arch Intern Med*. 2004;164(6):637–644
13. Relman AS. Assessment and accountability: the third revolution in medical care. *N Engl J Med*. 1988;319(18):1220–1222
14. Profit J, Gould JB, Zupancic JA, et al. Formal selection of measures for a composite index of NICU quality of care: Baby-MONITOR. *J Perinatol*. 2011;31(11):702–710
15. Profit J, Zupancic JA, Gould JB, Petersen LA. Implementing pay-for-performance in the neonatal intensive care unit. *Pediatrics*. 2007;119(5):975–982
16. Kowalkowski M, Gould JB, Bose C, Petersen LA, Profit J. Do practicing clinicians agree with expert ratings of neonatal intensive care unit quality measures? *J Perinatol*. 2012;32(4):247–252
17. Profit J, Zupancic JA, Gould JB, et al. Correlation of neonatal intensive care unit performance across multiple measures of quality of care. *JAMA Pediatr*. 2013;167(1): 47–54
18. Stark AR; American Academy of Pediatrics Committee on Fetus and Newborn. Levels of neonatal care. *Pediatrics*. 2004;114(5): 1341–1347
19. Patel AL, Engstrom JL, Meier PP, Kimura RE. Accuracy of methods for calculating postnatal growth velocity for extremely low birth weight infants. *Pediatrics*. 2005;116 (6):1466–1473
20. Escobar GJ, Greene JD, Scheirer P, Gardner MN, Draper D, Kipnis P. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Med Care*. 2008;46(3):232–239
21. Draper D, Gittoes M. Statistical analysis of performance indicators in UK higher education. *J R Stat Soc [Ser A]*. 2004;167:449–474
22. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. *J R Stat Soc [Ser A]*. 1996; 159:385–443
23. Efron BT, An RJ. *Introduction to the Bootstrap*. New York, NY: Chapman & HallHall/ CRC Monographs on Statistics & Applied Probability; 1994
24. Aberson CL. *Applied Power Analysis for the Behavioral Sciences*. New York, NY: Taylor and Francis Group, LLC; 2010
25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174
26. Kelley ET, Hurst J. Health Care Quality Indicator Project—conceptual framework. 2006; 23. Available at: www.oecd.org/data-oecd/1/36/36262363.pdf
27. Mattke S, Epstein AM, Leatherman S. The OECD Health Care Quality Indicators Project: history and background. *Int J Qual Health Care*. 2006;18(suppl 1):1–4
28. Brand DA, Saisana M, Rynn LA, Pennoni F, Lowenfels AB. Comparative analysis of alcohol control policies in 30 countries. *PLoS Med*. 2007;4(4):e151
29. Profit J, Gould JB, Draper D, Zupancic JA, Kowalkowski MA, Woodard L, et al. Variations in definitions of mortality have little influence on neonatal intensive care unit performance ratings. *J Pediatr*. 2013;162:50–55.e2

30. Williams SC, Koss RG, Morton DJ, Loeb JM. Performance of top-ranked heart care hospitals on evidence-based process measures. *Circulation*. 2006;114(6):558–564
31. Shahian DM, Normand SL. Comparison of "risk-adjusted" hospital outcomes. *Circulation*. 2008;117(15):1955–1963
32. Peterson ED, Delong ER, Masoudi FA, et al; ACCF/AHA Task Force on Performance Measures. 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(15):1780–1791
33. Bobko P, Roth PL, Buster MA. The usefulness of unit weights in creating composite scores: a literature review, application to content validity, and meta-analysis. *Organ Res Methods*. 2007;10:689–709

(Continued from first page)

Dr Profit and Dr Pietz had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

At the time of the research, Dr Profit was on faculty at Baylor College of Medicine, Texas Children's Hospital, Department of Pediatrics, Section of Neonatology. He held a secondary appointment in the Department of Medicine, Section of Health Services Research and conducted his research at the VA Health Services Research and Development Center of Excellence where he collaborated with Dr Kowalkowski.

www.pediatrics.org/cgi/doi/10.1542/peds.2013-3552

doi:10.1542/peds.2013-3552

Accepted for publication Mar 24, 2014

Address correspondence to Jochen Profit, MD, MPH, Perinatal Epidemiology and Health Outcomes Research Unit, Division of Neonatology, Department of Pediatrics, Stanford University School of Medicine, MSOB Room x115, 1165 Welch Rd, Stanford, CA 94305. E-mail: profit@stanford.edu

PEDIATRICS (ISSN Numbers: Print, 0031-4005; Online, 1098-4275).

FINANCIAL DISCLOSURE: Drs Profit, Zupancic, and Gould have served in consultant roles to the Vermont Oxford Network NICQ 7 and 8 Quality Improvement Collaboratives. The other authors have indicated they have no financial relationships relevant to this article to disclose.

FUNDING: Dr Profit's contribution is supported in part by the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development grant 1 K23 HD056298 (Principal Investigator: Dr Profit). Dr Petersen was a recipient of the American Heart Association Established Investigator Award (0540043N) at the time this work was conducted. Drs Petersen and Hysong also receive support from a Veterans Administration Center Grant (VA HSR&D CoE HFP90-20). Dr Hysong's contribution is supported in part by the Department of Veterans Affairs Health Services Research and Development Program (CD2-07-0181). Dr Thomas's effort was supported in part by grants from the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development grant 1 K24 HD053771-01 (Principal Investigator: Dr Thomas). Funded by the National Institutes of Health (NIH).

POTENTIAL CONFLICT OF INTEREST: Dr Gould is the Principal Investigator for the California Perinatal Quality Care Collaborative. Dr Profit is a researcher at the California Perinatal Quality Care Collaborative. The other authors have indicated they have no potential conflicts of interest to disclose.