EDITORIAL

Transforming risk factors into an unbiased mortality model

What explains 'higher mortality' at neonatal intensive care units (NICUs) in children's hospitals compared to perinatal center NICUs? Patients at children's hospital NICUs may differ in many ways from those at other NICUs, and some of those differences might explain the unadjusted association between NICU type and mortality. Severity of illness does not explain the unadjusted association between NICU type and mortality, so in this issue Berry et al.[1] explore additional risk factors.

Neonatal intensive care is an exceedingly complex web of processes interacting with myriad individual-level and group-level factors that modulate outcomes.[2] The main tool for disentangling these determinants is risk adjustment. The equation describing the rationale is deceptively simple:

$$\text{Outcome} = \text{treatment effectiveness} + \text{patient characteristics affecting response to treatment} + \text{quality of care} + \text{random chance.}[3] \quad \text{(Equation (1))}$$

Because it is so difficult to assign accurate values to the requisite equation terms, adjusted outcomes often have a low signal to noise ratio for quality of care.[4] Predictive models may achieve statistical significance, but correlation between quality of care and mortality remains low.[4,5]

To improve the outcome/quality signal to noise ratio, Berry et al. propose three adjusters: referral center, congenital anomalies and surgeries. Their work begs this question: After accounting for specified predictors, is it valid to consider outcome differences among institutions a surrogate for differences in quality of care? Well, it depends. 'All models are wrong; the practical question is how wrong do they have to be to not be useful.'[6] Here are just a few considerations germane to their model.

Rate denominators

NICU outcome rates are usually computed with institutional-based denominators: mortality rate = x deaths per 1000 NICU admissions per year. In general, a rate denominator counts the number of individuals who potentially can experience the outcome counted in the numerator. Institutional- and population-based rates answer different questions.[7] Institutional rates can actually misrepresent institutional performance. Berry et al. analyzed similar patient numbers for the two study hospitals using 6 months of data from Toronto and 12 months from Montreal. This suggests that the

facilities serve different size populations; despite comparable institutional-based denominators, population-based rate denominators might differ.

Imagine NICU A and NICU B, each with 1000 admissions per year and 80 deaths per year. Each institutional mortality rate = 8%. NICU A is its regional referral center, with population base = 10 000 births per year. NICU B is its regional referral center, with population base = 50 000 births per year. Assume that lower-level NICUs in each region transfer their sickest patients (highest mortality risk) to their regional referral center. NICU A and NICU B therefore have *different population-based mortality rates* despite identical institutional-based rates. NICU A's population-based rate = 80 deaths per 10 000 births per year = 0.8% and NICU B's population-based rate = 80 deaths per 50 000 births per year = 0.16%.

Assuming constant excess risk across strata

Statisticians adjust for case mix to compare, metaphorically, apples with apples when supplied with a bowl of apples, oranges and so on. The magic works, provided methodological assumptions are not violated. Berry et al.'s model assigned a particular coefficient value to the categorical variable 'admission from another NICU'; it assumed that the risk associated with initial care at another NICU is constant, the same for all infants and across all referring NICUs.

This is an important issue that is persuasively explained by Shwartz et al.[8] and summarized below. The essential idea is that if case mix strongly influences outcome and also varies widely among providers, then despite explicit accounting it may continue to distort performance comparisons.

Suppose that patients were categorized by baseline factors as either (a) low neonatal mortality risk = 1% or (b) high neonatal mortality risk = 5% and that in the overall population, half the patients are low risk and half are high risk, with an overall mortality rate of 3%. The task is to compare mortality at NICUs A–C. Each cares for 1000 patients, but the case mix varies greatly (Table 1). To render the argument specific to Berry et al., simply interchange 'high risk' with 'transferred from another NICU' and 'low risk' with 'transferred from non-NICU.'

NICU A cares for 800 low-risk patients; its 1% rate equals that for the overall low-risk population: 8 deaths per 800 patients. In contrast, the 10% rate for its 200 high-risk patients is twice that in the overall high-risk population: 20 deaths per 200 patients.

Table 1 Case mix and mortality for a hypothetical population and three hypothetical NICUs

	Overall population		NICU A		NICU B		NICU C	
Risk category	Case mix %	Mortality %	N	Mortality %	N	Mortality %	N	Mortality %
Low	50	1	800	1	200	1	800	1.25
High	50	5	200	10	800	10	200	12.50
Performance								
Mortality rate	3%		28/1000 = 2.8%		82/1000 = 8.2%		35/1000 = 3.5%	
SMR (O/E)			28/18 = 1.56		82/42 = 1.95		35/18 = 1.94	
Difference (O−E)			2.8−1.8 = 1%		8.2−4.2 = 4%		3.5−1.8 = 1.7%	

Abbreviations: NICU, neonatal intensive care unit; O/E, observed and expected rate; SMR, standardized mortality ratio; O−E, observed minus expected rate.
Adapted with permission from Ash AS, Shwartz M, Pekoz EA. Comparing outcomes across providers. In: Iezzoni LI (ed). *Risk Adjustment for Measuring Healthcare Outcomes*, 3rd edn. Chicago: Health Administration Press, 2003, p 307.

Thanks to NICU A's low proportion of high-risk patients, its crude rate is slightly better than average = 2.8%.

Obviously, crude rates can lead comparisons astray. One possible preventative is a method called indirect standardization. Risk stratum-specific rates for the population are applied to the numbers in each risk stratum for each NICU to yield expected outcomes (Table 1). Although Berry *et al*. use regression modeling, not standardization, regression may be considered to generalize standardization. Standardization is more transparent in revealing the fallacy of interest.

Expected outcomes may then be compared with the observed ones. When evaluating mortality, the ratio of no. of observed outcomes to no. of expected outcomes, O/E, is called the standardized mortality ratio, SMR.[9] Observed minus expected rates, O−E, is called excess mortality. Thus, applying the all low-risk patients' 1% mortality rate to the 800 low-risk NICU A patients yields eight expected deaths. Similarly, applying the all high-risk patients' 5% mortality rate to the 200 high-risk NICU A patients yields 10 expected deaths. The sum, 18 deaths, yields an expected mortality rate of 1.8%. NICU A's SMR of 1.56 means it experienced 56% more deaths than expected by case mix. Excess mortality rate, O−E, was 1%.

Next, compare NICU B with NICU A. Each experienced identical stratum-specific rates. However, NICU B cares for a high proportion of high-risk patients, so its crude rate is much higher than the average = 8.2%, highlighting the stimulus for some means of risk adjustment. NICU B's SMR = 1.95 and O−E = 4%, calculated as for NICU A. Similar patients experienced the same outcome at either facility, but NICU B appears to perform worse than NICU A.

Now consider NICU C. Its case mix is identical to NICU A's, but stratum-specific rates are 25% higher. Because of NICU C's case mix, its crude rate is lower than NICU B's. Indirect standardization is unable to reveal the performance gap. What is worse, excess mortality at NICU C suggests that it performs *better* than NICU B.

Table 1 assumes *constant* excess risk across strata in comparing observed outcomes to expected outcomes. Providers are inaccurately characterized despite case mix accounting because case mix influences outcome *differently at each hospital*.[10] One remedy is direct standardization: apply NICU-specific risk stratum rates to an externally defined 'standard population.' This produces a weighted average of the differences between NICUs for low- and high-risk patients based on their prevalences in the standard population (Table 2). Alternatively, one may limit NICU comparisons to those with similar patient mixes,[8] as discussed below in the section on product lines.

Collinearity

Logistic regression assumes that risk factors are independent of each other. When they are substantially correlated, the model can yield misleading results owing to a problem called collinearity.[11,12] In Berry *et al*.'s model, referral center, surgeries and congenital anomalies are plausibly correlated with illness severity.

Collinearity leads to unstable regression coefficient estimates, increases their standard errors and affects their computed *P*-values.[11,13] This instability reflects the difficulty in distinguishing the influence of one or the other of the highly correlated variables on the outcome.[12,13] The modeling method partials all the other independent variables from the relationship between each independent variable and the outcome.[11] If independent variables are substantially correlated with each other, the model assigns to them relatively small coefficient values and large standard errors. Risk factors could appear unimportant in predicting an outcome because they are correlated with other risk factors, not necessarily because they are indeed trivial. For example, in a study of the relationship between low income and low birth weight, low income had no effect.[14] Additional covariates included maternal and paternal education. Because the latter three variables are highly correlated, including them all in the model obscured the true relationships.[11]

Table 2 Direct standardization of crude data in Table 1

Risk category	N (1)	Mortality rate			Expected no. of deaths in standard population, using rates from		
		NICU A (2)	NICU B (3)	NICU C (4)	NICU A (1) × (2)	NICU B (1) × (3)	NICU C (1) × (4)
Low	1800	1%	1%	1.25%	18	18	22.5
High	1200	10%	10%	12.5%	120	120	150
Totals	3000				138	138	172.5
Direct standardized mortality rate					138/3000 = 4.6%	138/3000 = 4.6%	172.5/3000 = 5.75%

Abbreviation: NICU, neonatal intensive care unit.
Direct standardized mortality rate = Expected no. of deaths in standard population under specific NICU's rates/total no. of in standard population
The standard population is composed of all patients at NICUs A–C. When available, broader standard population data, such as state or national, may be used.

Bundled variables

A referral center bundles in one variable several outcome determinants, each of which might vary by patient, by NICU and over time. In addition to reflecting lead-time bias, discussed by Berry *et al.*, it also contains information about the quality of care (itself, a bundled variable) at the referring NICU. The model designed to characterize the quality of care at the receiving institution could be confounded by the quality of care at the referring institution. Adjusting for referral center could thus camouflage a quality problem at a referring hospital. Further, unless diagnoses for patients in each source group are similar, constant excess risk is assumed across the strata.

Assuming quality trumps chance

Investigators sometimes characterize quality by rearranging (Equation (1)) to leave only the quality of care and random chance on the same side of the identity. To be a valid evaluative strategy, random chance must be unimportant; quality of care must dominate.

Berry *et al.*'s variables account for a measurable, but not provided, proportion of the observed variation in outcomes. This proportion is denoted by a model performance measure called R^2, explanatory power. The explanatory power of logistic regression models is often surprisingly small (more specifically called pseudo-R^2) and necessarily constrained by low outcome incidence.[15] For example, R^2 for the Vermont Oxford Network (VON) mortality model $= 0.16$;[16] 84% of the observed variation in mortality is *not* explained by the model. Even if one was confident that the model indeed accounted for all the other substantive factors, one must still allocate the unexplained variation among quality and chance.

The respective allocation may be lopsided.[17] A study of mortality among 2671 infants in UK NICUs adjusted for illness severity, congenital malformations, gestation and birth weight; examined outcome variation over time; and accounted for systematic differences among NICUs. Variation among NICUs explained 0% (!) of the total variation in risk-adjusted mortality.[17] Within-NICU variation, not between-NICU variation, explained virtually all the differences in observed vs expected outcomes and for each NICU these differences varied over time.

Counterfactual reasoning

Berry *et al.*'s model may *define* the quality of care at the two study hospitals more than it *evaluates* it. This is because it is based only on the infants admitted to those two hospitals. In order to evaluate quality at those hospitals, data are required on infants similar to those studied but who were not transferred there and yet received the same interventions and procedures they would have received at the study hospital(s).

Product lines

Much of the NICU performance evaluation literature rests on the implicit yet crucial assumption that the appropriate unit of analysis is the NICU; the unit of observation tends to be the patient. Perhaps a more appropriate unit of analysis would be the *NICU product line*.

Reports comparing providers for adults commonly avoid describing a hospital's overall results; they stratify by product line. For example, providers are compared according to their 30-day mortality rates for coronary artery bypass graft surgery or survival rates after acute myocardial infarction. This practice reflects the agreement that the hospital as a unit of analysis aggregates too heterogeneous a group of patients, disease conditions and interventions. Risk adjusting cannot disentangle all the complexity.

Analytically, NICUs may be hospitals in their own right—for neonates with diverse medical and surgical problems, they are fundamentally heterogeneous service entities.

Correlated outcomes, multi-level analysis

Whatever the unit of analysis, logistic regression as used by Berry *et al.* assumes that observations are independent of each other. But what happens to one patient may not be independent of what happens to another patient in the same NICU. Moreover, it may be fallacious to use patient-level data to predict institution-specific outcome risk for subsequent patients.[18] Associations observed at the patient level may not hold at the NICU level.

Accounting for the varying roles of individual-level and group-level factors in jointly determining outcomes requires multi-level, or random-effects, modeling. This was how the 'systematic differences among NICUs' in an earlier cited study were accounted for,[17] and this is how the VON computes 'shrunken' performance estimates provided to participating centers (http://www.vtoxford.org/).

Conclusion

The challenges in translating candidate risk adjusters into unbiased and equitable models are many and daunting. The practical test that a model provides operational insight, ' ... how wrong do they have to be to not be useful,'[6] is straightforward: Do the findings lead to improved population-based rates?

Acknowledgments

J Schulman
*Department of Pediatrics/Division of Newborn Medicine
and Department of Public Health/Division of Outcomes
and Effectiveness,
New York-Presbyterian Hospital,
Weill Cornell Medical Center, Weill Medical
College of Cornell University, New York, NY, USA*
E-mail: jos2039@med.cornell.edu

References

1 Berry MA, Shah PS, Brouillette RT, Hellmann J. Predictors of mortality and length of stay for neonates admitted to children's hospital neonatal intensive care units. *J Perinatol* 2008; **28**: 297–302.

2 Schulman J. *Evaluating the Processes of Neonatal Intensive Care*. BMJ Books: London, 2004.

3 Iezzoni LI. The risks of risk adjustment. *JAMA* 1997; **278**(19): 1600–1607.

4 Lilford RJ, Brown CA, Nicholl J. Use of process measures to monitor the quality of clinical practice. *BMJ* 2007; **335**: 648–650.

5 Pitches DW, Mohammed MA, Lilford RJ. What is the empirical evidence that hospitals with higher-risk adjusted mortality rates provide poorer quality care? A systematic review of the literature. *BMC Health Serv Res* 2007; **7**: http://www.biomedcentral.com/1472-6963/7/91.

6 Box GEP, Draper NR. *Empirical Model-Building and Response Surfaces* 1987. New York: Wiley p 74.

7 Wennberg JE, Gittelsohn AM. *A Small Area Approach to the Analysis of Health System Performance*. U.S. Department of Health and Human Services: Washington, DC, 1980.

8 Shwartz M, Ash AS, Iezzoni LI. Comparing outcomes across providers. In: Iezzoni LI (ed). *Risk Adjustment for measuring Healthcare Outcomes*, 2nd edn. Health Administration Press: Chicago, 1997, pp 471–516.

9 Kirkwood BR, Sterne JAC. *Essential Medical Statistics*. 2 edn. Blackwell Science Ltd.: Oxford, UK, 2003.

10 Zaslavsky AM. Statistical issues in reporting quality data: small samples and casemix variation. *Int J Qual Health Care* 2001; **13**(6): 481–488.

11 Kiely JL. Some conceptual problems in multivariable analyses of perinatal mortality. *Paediatr Perinat Epidemiol* 1991; **5**: 243–257.

12 Schulman J. *Managing Your Patients' Data in the Neonatal and Pediatric ICU: an Introduction to Databases and Statistical Analysis*. Blackwell: Oxford, UK, 2006.

13 Dallal GE. Collinearity. 2001, http://www.tufts.edu/∼gdallal/collin.htm;accessed 15 October 2007.

14 Gortmaker SL. Poverty and infant mortality in the United States. *Am Sociolo Rev* 1979; **44**: 280–297.

15 Ash AS, Schwartz M. Evaluating the performance of risk-adjustment methods: dichotomous outcomes. In: Iezzoni LI (ed). *Risk Adjustment for Measuring Healthcare Outcomes*. Health Administration Press: Chicago, 1997, pp 427–470.

16 Rogowski JA, Horbar JD, Staiger DO, Kenny M, Carpenter J, Geppert J. Indirect vs direct hospital quality indicators for very low-birth-weight infants. *JAMA* 2004; **291**(2): 202–209.

17 Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO, International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study. Annual league tables of mortality in neonatal intensive care units: longitudinal study. *BMJ* 1998; **316**: 1931–1935.

18 Diez-Roux AV. Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *Am J Public Health* 1998; **88**(2): 216–222.