Published in final edited form as:

Med Decis Making. 2013 April ; 33(3): 396–406. doi:10.1177/0272989X12461855.

Identifying Individual Changes in Performance with Composite Quality Indicators while Accounting for Regression-to-the-Mean

Byron J. Gajewski^{1,2} and Nancy Dunton²

¹Department of Biostatistics, University of Kansas School of Medicine, Kansas City, KS 66160

²University of Kansas School of Nursing, Kansas City, KS 66160

Abstract

Almost a decade ago Morton & Torgerson (p. 1084) indicated that perceived medical benefits could be due to “...regression-to-the-mean.” Despite this caution, the regression-to-the-mean “... effects on the identification of changes in institutional performance do not seem to have been considered previously in any depth.” (Jones & Spiegelhalter, p. 1646). As a response, Jones & Spiegelhalter provide a methodology to adjust for regression-to-the-mean when modeling recent changes in institutional performance for *one variable quality indicators (QIs)*. Therefore, in our view, Jones & Spiegelhalter provide a breakthrough methodology for performance measures. At the same time, in the interests of parsimony, it is useful to aggregate individual QIs into a composite score. Our question is: Can we develop and demonstrate a methodology that *extends the ‘regression-to-the-mean’ literature* to allow for *composite quality indicators*? Using a latent variable modeling approach, we extend the methodology to the composite indicator case. We demonstrate the approach on four indicators collected by the National Database of Nursing Quality Indicators® (NDNQI®). A simulation study further demonstrates its “proof of concept.”

Keywords

provider profiling; individual changes; NDNQI; regression-to-the-mean; Bayesian Analysis

1. INTRODUCTION

One of the first policy responses to the IOM reports, [1] & [2], by governments and regulatory organizations was the requirement for health care providers, including nurses, be required to define, collect and monitor health care quality indicators (QI) relative to peers and track individual *improvement* over time. Several indicators may reflect the quality of care. In the interests of parsimony, it is useful to aggregate individual QIs into a composite score [3]. To that end, we are interested in developing a methodology for providing guidance as to when a provider has significantly improved or declined on overall quality as measured by *several* indicators.

There have been several methodological approaches to developing composite quality indicators. One example of measuring quality with several indicators includes a multilevel multidimensional latent variable model that examines the performance of a network of hospitals [4]. Another cross sectional approach is a model-based score developed using a latent variable model using Item Response Theory (IRT) [5]. An extension of [5] further discusses the merits of an IRT latent variable approach to modeling multivariate indicators

* Address for correspondence: Department of Biostatistics, University of Kansas School of Medicine, Mail Stop 1026, 3901 Rainbow Blvd., Kansas City, KS 66160, Phone: 913-588-1603, Fax: 913-588-0252, bgajewski@kumc.edu.

[6]. Serial dependence among latent variables is modeled through pairwise correlations [7]. We use a latent variable approach that *extends the ‘regression-to-the-mean’ literature* to allow for *multiple variables*.

Jones & Spiegelhalter [8] provide a methodology to adjust for ‘regression-to-the-mean’ when modeling recent changes in institutional performance for *one variable*. More on the history of regression-to-the-mean, including its connection to test-retest literature, is found in Jones and Spiegelhalter. The methodology relies on an ANCOVA type argument by predicting a particular institution’s performance using its time 1 outcome to predict what would have happened at time 2. The standardized difference between the observed and the prediction provides the test statistic for determining the institutions that have significant improvement or decline. Note that one can use an appropriate cut-point for this test statistic to adjust for false discovery rates [8].

We will examine quality on *individual acute care nursing units* (‘institution’) to track their own quality improvement. The individual nursing unit does not necessarily care about adjusting for all units across the nation. Therefore, if the unit wants a 0.05 Type I error for an indicator, they can use a usual two sided cut-point of $Z=1.96$. If units are interested in tracking several quality indicators at once, in our case there are four, they will likely be interested in controlling for an individual nursing unit’s type I error rate. To control for the error rate the unit can use a Bonferroni adjustment ($.05/4=.0125$) or alternatively, use the multivariate approach we propose here.

The research question we address is: Can we develop and demonstrate a methodology that *extends the ‘regression-to-the-mean’ literature* to allow for *composite quality indicators*? To account for multivariate data, we propose to take advantage of a single latent variable that may represent the nursing unit’s *quality*. Latent variable models (or Structural Equation Modeling) have become popular methods for multivariate healthcare data modeling (see, for example [9], [10]). They are most useful when one can define a finite number of constructs (usually smaller than number of variables) to summarize the multiple-dimensions. We will show that the construct unit *quality* adequately explains the variation when the appropriate distributions are used. For example, total nursing hours per patient day tends to resemble log-normal distribution, job enjoyment a normal distribution, fall rates a Poisson distribution, and pressure ulcers a binomial distribution. After these technical modeling details are established, we show that all temporal quality improvement (or decline) can be monitored on the construct unit *quality*. Not only does this approach control for Type I error and provide a parsimonious interpretation, it also adjusts for regression-to-the-mean.

2. METHODS: GENERAL MODELING AND DATA

In this section we lay the groundwork for adjusting for regression-to-the-mean of the construct unit *quality*. We will review the univariate version of the test, further introduce the American Nurses Association’s National Database of Nursing Quality Indicators® (NDNQI®) (see <http://www.nursingquality.org/>) NDNQI database, and show an example of a univariate test from NDNQI.

This paper focuses on the development of composite quality indicators for hospital-based nursing care, using data from the NDNQI. Today, over 1800 hospitals with more than 17,000 nursing units participate in the database. Nursing care is provided on the individual unit or ward, and so the quality of care is assessed at the unit level. The types of quality indicators that NDNQI develops and monitors are reflective of Donabedian’s quality improvement model in which the structure and process of care influence the outcomes of care [11]. An example of a structural indicator is the number of nursing hours per patient

day. The number of nurses assigned to care for patients is determined by state policy (in some cases), professional standards, institutional resources, and arguably the general structural dedication to the quality of care in a unit. An example of a process measure is Registered Nurse (RN) job enjoyment. Two example patient outcomes include the fall rate and the hospital-acquired pressure ulcer rate. A fall is an “unplanned decent to the floor” and a pressure ulcer is also known as a “bed sore.” We define *quality* as a nursing unit’s *commitment* to quality. High staffing, happy staff, and safe patients reflect a commitment to high quality. We will focus the rest of this paper on these four quality indicators. NDNQI routinely demonstrates the reliability and validity of these indicators through carefully designed and analyzed studies (e.g. 12, 13, 14, 15, 16).

2.1 Univariate Adjustment for Regression-to-the-Mean (Jones & Spiegelhalter [8])

Using Jones & Spiegelhalter notation [8], let Y_{i1} and Y_{i2} be two realizations of a performance indicator for the i^{th} institution at time 1 and time 2 respectively. Under the null hypothesis of no change the ‘true’ unit performance is denoted by θ_i , and the distributions of $Y_{i1}|\theta_i$ and $Y_{i2}|\theta_i$ are conditionally independent, $N(\theta_i, \sigma^2)$. Via hierarchical model we can set $\theta_i = \mu + d_i$ where $d_i \sim N(0, \tau^2)$. Thus the marginal distribution of Y_{i1} and Y_{i2} are both $N(\mu, \tau^2 + \sigma^2)$ and intra-class (i.e. between period) correlation $\rho = \tau^2 / (\tau^2 + \sigma^2)$. Note that this approach assumes constant σ^2 , both across units and across time. Jones & Spiegelhalter also present a test statistic for the case of non-constant σ^2 . The alternative hypothesis is that θ_i does change over time. The adjusted test statistic, under the null hypothesis, is conditioned

on observing the institution’s first value: $Z^*| \{Y_{i1} = y_{i1}\} = (Y_{i2} - \hat{Y}_{i2}) / \sqrt{(1+\rho)\sigma^2}$, where $\hat{Y}_{i2} = \rho y_{i1} + (1 - \rho)\mu$, and $Y_{i2} - \hat{Y}_{i2}$ is the adjusted measure of individual change (i.e. adjusted for regression-to-the-mean). Jones & Spiegelhalter [8] indicate the test static is superior to the naïve version of $Z = (Y_{i2} - Y_{i1}) / \sqrt{2\sigma^2}$ because unlike Z , Z^* is uncorrelated with Y_{i1} so it adjusts for the classic “regression-to-the-mean.” While Jones & Spiegelhalter [8] mainly use a frequentist justification for this test based on the sampling distribution of $Y_{i2} | \{Y_{i1} = y_{i1}\}$, they briefly point out that the same test statistic has predictive distribution justification: “Equivalently, this tests the ‘surprisingness’ of Y_2 based on a simple random effects predictive distribution, whereas Z is the test statistic based on a fixed effects (independent) predictive distribution. [8, p. 1649]” Their paper has frequentist foundation, but we believe it can be conceptualized from a Bayesian point of view. More specifically, Y_{i2}^p (posterior predictive distribution given $Y_{i1} = y_{i1}$) has mean $\hat{Y}_{i2} = \rho y_{i1} + (1 - \rho)\mu$ and variance $(1 + \rho)\sigma^2$. A frequentist would plug in point estimates of ρ , μ , and σ^2 .

A *fully* Bayesian approach involves accounting for uncertainty in the population level parameters in each predictive distribution. The test statistic itself, Z^* , is not what would arise from a *fully* Bayesian approach. One way to accomplish a fully Bayesian approach is to use a software program tailored for such models (e.g. WinBUGS). This will allow σ^2 and μ to vary as well as ρ and τ^2 , although $\rho = \tau^2 / (\tau^2 + \sigma^2)$. We will accomplish such a fit in the univariate case after introducing the data. The univariate case motivates the multivariate model.

2.2 Data: NDNQI

The NDNQI reports that are generated for member units are stratified by bed size and unit type (as well as other hospital characteristics). We will focus on unit type in this article. The different unit types include: critical care, step down, medical, surgical, combined medical-surgical, and rehabilitation. Our example focuses on combined medical-surgical units aggregated up to the unit level for years 2006 and 2007. We focus on the NDNQI indicators: total nursing hours per patient day (TNHPPD), registered nurse job enjoyment (JE), fall rates, and pressure ulcer (PrU) rates.

The relative nursing effort in a unit is indicated as a ratio, total nursing hours and patient days (TNHPPD). Job enjoyment data are from a survey of registered nurses (RN). This indicator is an average of seven questions on a six-point Likert scale, ranging from (1) strongly disagree to (6) strongly agree. Example questions include: ‘nurses are satisfied with jobs’ and ‘find real enjoyment in their job.’ The number of falls per thousand patient days defines fall rates. The number of patients in a 24 hour period that have at least one hospital acquired pressure ulcer as a proportion of all patients assessed is the pressure ulcer rate. Note that as discussed in [17] we are not performing risk adjustment for our modeling, which is consistent with how NDNQI reports are presented to members. Table 1 indicates that across the two years average, TNHPPD is about eight hours per patient day. That converts to about three patients for every RN. The mean job enjoyment is very close to the mid-point (1–6 scale). The fall rates’ mean is about four falls for every 1,000 patient days. The mean percentage of patients with pressure ulcers is about 5.5%.

2.3 Example: TNHPPD

To illustrate Jones and Spiegelhalter’s method on our data, as well as its Bayesian analog discussed above, we examine the log transformed version of total nursing hours per patient day (TNHPPD): $Y_{i1} = \log(\text{TNHPPD}.1_i)$ and $Y_{i2} = \log(\text{TNHPPD}.2_i)$. We fitted a one-way random effects ANOVA model and estimated parameters using ANOVA expected mean squared error (Kuehl, 1994 [18], pp. 133–134) as $\hat{\mu} = 2.11$, $\hat{\sigma}^2 = 0.0032$, $\hat{\tau}^2 = 0.0240$, and $\hat{\rho} = 0.0240 / (0.0240 + 0.0032) = .8824$. Note that while Jones and Spiegelhalter found that the increase in power due to the adjustment was very small with ρ as large as 0.8, we simply illustrate the approach on TNHPPD. The value for ρ is much smaller in the multivariate example. We then flagged all of the units as significantly improving or declining that had $Z^* > 1.96$ or $Z^* < -1.96$ respectively. There are only two units out of 290 that are significantly decreasing. One dropped from 10.4 to 8.2 (Unit A) TNHPPD ($Z^* = -2.74$) while the other dropped from 6.8 down to 5.8 (Unit B) ($Z^* = -2.22$). This is depicted graphically on the log scale in Figure 1. Unit A started above the overall mean so part of its drop is explained by regression-to-the-mean. Unit B had only a one point drop but it is significant because it is trending opposite the overall mean. There are 11 units that had significant improvement. The unit with the largest improvement started with a rate of 6.9 and improved to 10.3 ($Z^* = 4.9$). As theory dictates $\text{corr}(Y_{i1}, Z^*) = -.035 \sim 0$ and $\text{corr}(Y_{i1}, Z) = -.288 < 0$; illustrating advantage of Z^* over Z . Note that as mentioned earlier, from NDNQI viewpoint, we are controlling for an individual nursing unit’s type I error rate. However, we warn here that from an overall perspective, for the 290 units, 1 in 20 (~15 units) are expected to be “significant” by chance alone. We also applied a Bayesian calculation via WinBUGS. We essentially placed flat priors on all unknown parameters and then calculated the posterior predictive distribution at each iteration of simulation to be $Y_{i2}^P \sim N(\rho y_{i1} + (1 - \rho)\mu, (1 + \rho)\sigma^2)$ and then calculated $N^{*P} = (Y_{i1} - Y_{i2}^P) / \text{sd}((Y_{i1} - Y_{i2}^P))$. Note that theoretically $\text{sd}((Y_{i1} - Y_{i2}^P)) > (1 + \rho)\sigma^2$ because the propagation of error from posterior distributions of ρ and σ^2 . But because of the relative large sample size, this propagation is negligible. We burned in 1,000 iterations and used the next 10,000 for inference. The Bayes estimates were very similar to the ANOVA-based estimates $E(\mu) = 2.11$, $E(\sigma^2) = 0.0033$, $E(\tau^2) = 0.0242$, and $E(\rho) = 0.8801$. The same units were significant under both approaches (Z^{*P} or Z^*). Note that in both approaches it is likely that the assumption of multivariate log-normal distribution is important. Examination of the marginal q-q plots indicates this assumption is valid.

3. METHODS: COMPOSITE INDICATOR (MULTIVARIATE) ADJUSTMENT FOR REGRESSION-TO-THE-MEAN

Almost a decade ago Morton & Torgerson [19] (p. 1084) indicated that perceived medical benefits could be "...due regression-to-the-mean." Despite this caution "...the phenomenon's effects on the identification of changes in institutional performance do not seem to have been considered previously in any depth." (Jones & Spiegelhalter, p. 1646). Therefore, in our view, Jones & Spiegelhalter provides a breakthrough methodology for performance measures. We expand this methodology for composite indicators.

The previous section illustrated the univariate approach for recognizing a unit's significant improvement or decline. Now that we have a goal to create a reflective-type model, we will use the same idea for recognition of significant units from a composite indicator (multivariate) approach via latent variable modeling.

3.1 Composite Indicators via a Multivariate Latent Variable Model

Building from the univariate model described above we begin the multivariate model by letting Y_{jit} be the j^{th} variable where $j=1,2,3,4$; i is the i^{th} unit where $i=1,2,3,\dots,290$; and $t=1,2$ where $t=1$ is from 2006 and $t=2$ is 2007. The sampling distribution of each is

$$Y_{jit} | \theta_{jit} \sim h_j(\theta_{jit}, \sigma_{jit}^2) \quad (1)$$

where θ_{jit} is a location parameter such that

$$g_j(\theta_{jit}) = (a_j + b_j f_{it} + d_{ji}) \quad (2)$$

and σ_{jit}^2 is the variance of the sampling distribution. Equation (2) is the single factor measurement model that allows for the i^{th} unit's t^{th} time point to correlate with latent variable f_{it} . The a_j represents the intercept for j^{th} variable. The b_j represents the respective regression coefficient (loading) between latent construct (f_{it}) and the transformed location parameter $g(\theta_{jit})$. The g_j is a link function to the measurement model via usual generalize linear model (glm) (e.g. [20]). The g_j link function can vary in form, such as logistic or exponential transformations. One clear advantage of this transformation is it forces the parameters to be in the correct space. For example, logistic keeps probabilities between 0 and 1 and exponential keeps fall rates positive. We allow the latent factor to correlate across time points using the same idea as in the univariate model with two normal distributions

$d_i \sim N(0, \sigma_d^2)$ and $f_{it} | d_i \sim N(d_i, 1 - \sigma_d^2)$. This latter constraint $(1 - \sigma_d^2)$ is imposed for identification purposes such that the unconditional $\sigma_f^2 = \text{var}(f_{it}) = 1$. Therefore, $\rho = \sigma_d^2$. It will be shown with data, that latent factor correlation does not explain enough of the correlation within measures across time. Therefore the correlations are further modeled with $d_{ji} \sim N(0, \sigma_{d_j}^2)$. The distributions defined in Equation (1) and the link in Equation (2) vary across the four variables.

For each of the four manifest variables, we need to define the variance components (σ_{jit}^2), sampling distribution (h_j), and link function (g_j).

3.1.1 Variable 1: Total Nursing Hours per Patient Day (TNHPPD)—As discussed in Section 2.3, the total nursing hours per patient day is justifiably modeled by a lognormal distribution. Therefore the variance, distribution, and link will be defined as

$\sigma_{1it}^2 \equiv \sigma_1^2; h_1(\theta_{1it}) \equiv LN(\theta_{1it}, \sigma_1^2)$; and $g_1(\theta_{1it}) \equiv \theta_{1it}$ respectively. The symbol “LN” stands for log-normal distribution. The average TNHPPD for the i^{th} unit is thus $\exp(\theta_{1it} + \sigma_1^2/2)$.

3.1.2 Variable 2: RN Job Enjoyment (JE)—Let Y_{2it} be the observed average score for unit i at time t and σ_{2it}^2 the standard deviation for n_i RNs in unit i . We assume that Y_{2it} is normally distributed; which is substantiated in [17]. We set

$\sigma_{2it}^2 \equiv \sigma_2^2/n_i; h_2(\theta_{2it}) \equiv N(\theta_{2it}, \sigma_{2it}^2)$; and $g_2(\theta_{2it}) \equiv \theta_{2it}$ respectively. Note that based on previous studies, $\sigma_2^2=0.726873$ [17]. Note that the variance of the sampling distribution is weighted by the number of RNs in each unit for the sampling time frame. The average RN job enjoyment for the i^{th} unit is thus θ_{2it} .

3.1.3 Variable 3: Fall Rate—For fall rates, we assume that Y_{3it} is the number of falls for the i^{th} year and w_{it} represents the respective number of patient days divided by 1,000. We assume that Y_{3it} follows a Poisson distribution and that θ_{3it} is the average fall rate (per 1,000 patient days) for the i^{th} unit and t^{th} year. Therefore, $Y_{3it}|\theta_{3it} \sim h_2(\theta_{3it}) \equiv \text{Poisson}(\theta_{3it} w_{it})$; which is substantiated in [17]. The link function is $g_3(\theta_{3it}) \equiv \ln(\theta_{3it})$ where we are setting $\sigma_{3it}^2 \equiv 1$. The average fall rate for the i^{th} unit is thus θ_{3it} .

3.1.4 Variable 4: Pressure Ulcer (PrU) Rate—The final variable involves hospital acquired pressure ulcers (PrU). Let Y_{4it} be the number of patients out of n_{it} who have a hospital acquired pressure ulcer. We assume that Y_{4it} is a binomial distribution with n_j trials and that θ_j is the average pressure ulcer rate for unit j . Therefore, $Y_{4it}|\theta_{4it} \sim \text{Bin}(\theta_{4it}, n_{it})$; also substantiated in [17]. The link function is $g_4(\theta_{4it}) \equiv \text{logit}(\theta_{4it})$ where we are setting $\sigma_{4it}^2 \equiv 1$. The proportion of pressure ulcers for the i^{th} unit is thus θ_{4it} .

3.1.5 Overall Model: Testing Improvement or Decline—The overall latent variable model is summarized in Figure 2 using standard graphical descriptions - ovals for latent variables and squares for manifest variables. The f 's represent the univariate latent variable, *quality*, that manifests itself as θ 's, after transformation, using the intercepts (a 's) and slopes (b 's). The f 's are allowed to vary through the variation in d 's (random effect of units). The manifest variables (Y 's) vary from the θ 's and sampling distributions. Before we can test for improvement in the i^{th} unit, all of the non-realized parameters in this model need estimation. We will accomplish this estimation via Bayesian modeling, which we will articulate further in Section 4. We will test the multivariate significant improvement or decline of a unit using the conditional distribution of the latent variable $f_{i2}^P | f_{i1} \sim N(\rho f_{i1}, (1+\rho)\sigma_f^2)$ where $\rho=\sigma_d^2$, and recall $\sigma_f^2=1$. Therefore our multivariate Bayesian test-statistic is $Z_f^* = (f_{i2} - f_{i2}^P) / \text{sd}(f_{i1} - f_{i2}^P)$, that properly adjusts for “regression-to-the-mean” in a similar manner to that of Jones & Spiegelhalter. While f_{i2} and f_{i1} are unknown (latent variable), like the other unknown parameters in the model, we are monitoring their posterior distributions in WinBUGS at each iteration.

Some may argue that a Jones & Spiegelhalter type correction does not do anything as it is applied to a latent variable as opposed to observed values. Since latent variables represent some underlying “true value”, it may seem that they are not subject to regression to the mean in the same way. However, pointed out to us by reviewer, this fear is unjustified in the context of the model for the latent variables. A simple model, one that sets $d_{it}=0$, is strictly analogous to the model given by Jones and Spiegelhalter for random variables in section 2.1 (basic assumptions) and section 2.2 of their paper. A simple latent variable model has $E(f_{it}|d_i) = E(f_{it2}|d_i) = d_i$ and $\text{Var}(f_{it1}|d_i) = \text{Var}(f_{it2}|d_i) = 1 - \sigma_d^2$. Further, in the population, $E(d_i) = 0$ and

$\text{Var}(d_i) = \sigma_d^2$. So, unconditionally, we have $\text{Var}(f_{i1}) = \text{Var}(f_{i2}) = 1$ and $\text{Cov}(f_{i1}, f_{i2}) = \sigma_d^2$. So $\rho = \text{Corr}(f_{i1}, f_{i2}) = \sigma_d^2$. Given the assumption that the distributions are normal and that $0 \leq \rho \leq 1$, then $E(f_{i2}|f_{i1}) = \rho f_{i1} + (1 - \rho)\mu = \rho f_{i1}$, since $\mu = 0$. Therefore, $E(f_{i2}|f_{i1})$ lies between f_{i1} and $\mu = 0$. We will later show that the value of the latent variable on the second occasion (year 2007 for the data set used) has regressed toward the mean, precisely as in Jones and Spiegelhalter, p. 1648–1649, section 2.2. We have f_{i1} and f_{i2} analogous to Jones and Spiegelhalter's Y_{i1} and Y_{i2} and d_i analogous to Jones and Spiegelhalter's θ_i , etc. Further, even for the more complex latent variable model (e.g. $d_{ji} \neq 0$) this issue will be clarified in both the substantive example and the proof of concept simulation where we show the latent variable model does in fact suffer from regression-to-the-mean because estimates of $\text{Corr}(f_{i1}, f_{i2}, f_{i2}^P)$ is much closer to 0 than estimates of $\text{corr}(f_{i1}, f_{i2} - f_{i1})$.

3.2 Priors

The final step of the Bayesian model requires the prior distributions which we specify to be approximately “flat” or noninformative. The intercepts and slopes have the distributions $a_j \sim N(0, 100)$ and $b_j \sim N(0, 100)$. All of the variance components $\sigma_d^2, \sigma_{d_j}^2$, and σ_1^2 have a prior of $IG(0.01, 0.01)$, where IG is the inverse-gamma distribution.

4. METHODS: COMPUTATION, MODEL FIT & SIMULATION

WinBUGS [21] was used to fit a Bayesian estimation via Markov chain Monte Carlo (MCMC). Following a burn-in of 5,000 iterations, the posterior distributions were monitored over a further 10,000 iterations of the MCMC. Sensitivity analysis of the specification of the priors for the variance components [22] indicated negligible change in the posterior distributions. Therefore, we report only the posterior distribution using priors above. For all parameters in the model the 2.5%-tile, median, and 97.5%-tile were used to summarize the posterior distributions. For calculating significant improvement or decline in *quality*, we calculate the proportion of times that f_{i2} is larger than f_{i2}^P , calling it Pr_i . For reporting purposes Bayesian p-value = $\min(Pr_i, 1 - Pr_i)$, which reflects the one-sided Bayesian probability that there is significant improvement or decline.

Model assessment was accomplished through a relative fit as well as goodness of fit of the correlation matrix of two models. An alternative model did not include the variance components that modeled the direct error correlations between indicators (specifically $\sigma_{d_1}^2 = \sigma_{d_2}^2 = \sigma_{d_3}^2 = \sigma_{d_4}^2 = 0$ but $\sigma_d^2 > 0$). That is, all of the correlation among variables was modeled only through the quality latent variables (double arrows on right of Figure 2 are deleted). The second model was the model described in Section 3 that fits all variance components (specifically $\sigma_{d_1}^2 > 0, \sigma_{d_2}^2 > 0, \sigma_{d_3}^2 > 0, \sigma_{d_4}^2 > 0$ & $\sigma_d^2 > 0$). First, for relative fit we compare the deviance information criteria (DIC) [23] for both models. Second, we calculate the correlation matrix among all indicators across two time periods and compare it to correlation matrices generated from posterior draws from the two models.

Finally, as a “proof of concept,” we simulate data to understand the properties of our basic approach. We would like to (I) demonstrate that our approach properly estimates parameters from simulated data and (II) understand how well the approach detects true change in quality (power). Pseudo random numbers, using Matlab software, are used for all simulations. Our basic data set involves $n=100$ nursing units that is simulated using the model in Section 3 and corresponding median of the posterior distribution of the parameters listed in Table 2 (Section 5.2, below). Note that we are simulating a scenario in which there

are no temporal changes (i.e. $[f_{i1}, f_{i2}] \sim MVN\left(\begin{bmatrix} 0 & 0 \\ 1 & \rho \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$). We will fit the MCMC algorithm and compare posterior distributions to this truth, demonstrating part (I).

Part (II) adds some complexity to the basic dataset by randomly selecting a unit (index it as $i=100$) and taking it out of the sample. We replace the data from that unit with data using the following two-way factor design. The factors involve the assumption regarding the quality at time 1 (f_i) and the other the difference in quality from time 1 to time 2 (f_2-f_i). The first factor will vary the starting point of where the unit's quality begins, either at the mean ($f_i=0$) or far from the mean ($f_i=3$). The second factor will vary in how much the quality of the unit will increase across time ($f_2-f_i=0.0, 0.50, 1.00, \dots, 5.0$). This second simulation will involve 5 iterations for each combination of the two factors. Thus, separate simulations will be run for $2 \times 11 = 22$ combination of factors.

The main outcome in this simulation will be *p-value* _{i} described above, which will be calculated for each combination. A multiple regression, the quality at time 1 (binary) and the difference in quality from time 1 and time 2 (continuous) interaction with quality at time 1 will be estimated, after a logit transformation, to understand how *p-value* varies across the two factors. More specifically, $\text{logit}(p\text{-value}_i) = B_0 + B_1 x_i + B_2 (1-x_i)(f_{2i}-f_{1i}) + B_3 x_i(f_{2i}-f_{1i}) + e_i$, where $x_i=0$ if $f_{1i}=0$ and $x_i=1$ if $f_{1i}=3$. After estimated, we will plot $\text{logit}^{-1}(\hat{B}_0 + \hat{B}_1 x_i + \hat{B}_2 (1-x_i)(f_{1i}-f_{2i}) + \hat{B}_3 x_i(f_{1i}-f_{2i}))$ as a function of x_i and $(f_{2i}-f_{1i})$.

5. RESULTS

5.1 Model fit

The full proposed model has a DIC=8130.2 (all variance components) which is much smaller than the alternative model of DIC=11310.1 (some variance components set to 0). Call Rn the sample correlation matrix from the observed indicators (i.e., correlation matrix from $\exp(Y_{1i}), Y_{2i}, Y_{3i}/w_i, Y_{4i}/n_i$ in Table 1) and Rp and RpA a posterior predictive correlations (one iteration each) from full model and alternative respectively, which comes from posterior predictive data from the respective models. The mean absolute error (MAE)

$$MAE_{np} = \sum_{i=1}^8 \sum_{j=1}^8 |Rn_{ij} - Rp_{ij}| / (8^2 - 8)$$

was calculated as $MAE_{npA} = \sum_{i=1}^8 \sum_{j=1}^8 |Rn_{ij} - RpA_{ij}| / (8^2 - 8)$ respectively. As with the DIC, the full proposed model's MAE was lower than the alternative model's ($MAE_{np}=0.05$ and $MAE_{npA}=0.22$). The fit analysis clearly indicates modeling the extra variance components is well worth the effort.

5.2 Estimation of full Model Parameters (Univariate & Multivariate)

The model's intercepts (a 's) across indicators have median values and intervals, after appropriate back-transformation to the original scale of quality indicators (Table 2), that either cover or are close to covering the mean observed values (Table 1). This indicates a reasonable fit and is an appropriate comparison since the median value of $a_j + b_j f_{it} + d_{jt}$ is a_j .

Regarding the relationship between indicators and quality, all slopes (b) are significant since the posterior probability that the slopes are bigger than 0 is greater/less than 0.95; $Pr(b>0)>.95$, (TNHPPD and JE) or $Pr(b<0)>.95$ (Fall & PrU rates). Taking the ratio of posterior mean and standard deviation indicates the relative absolute strength of the association between quality and: JE ($|Z|=7.6$), PrU rates ($|Z|=7.1$), Fall Rates ($|Z|=5.5$), and TNHPPD ($|Z|=2.6$).

Thus JE and PrU rates are the most influential factors for the empirical definition of quality in nursing units.

The temporal associations across time are all significant ($\sigma_{d_1}^2, \sigma_{d_2}^2, \sigma_{d_3}^2$, and $\sigma_{d_4}^2$) as indicated in the analysis of the alternative model in Section 5.1. The correlation across time for the quality variable runs from small to large since its 95% credible interval is .14–.65, indicating a clear need for adjustment for regression to the mean.

5.3 Multivariate Profiling: Testing Improvement or Decline

As theory dictates $\text{corr}(\hat{f}_{i1}, \hat{f}_{i2} - \hat{f}_{i2}^p) = -0.23$ which is much closer to 0 than $\text{corr}(\hat{f}_{i1}, \hat{f}_{i2} - \hat{f}_{i2}) = -0.67$ where $\hat{x} = E(x)$; illustrating advantage of Z^* over Z , thus the regression adjustment is an improvement. Figure 3 shows the unit (call it Unit A) with the largest adjusted increase in quality ($p=.0195$) from 2006 to 2007. Depicted in the plot is the predicted regression up towards mean quality (0); what is subtracted is this value from the estimated quality at 2007.

Figure 3 also shows the unit (Call it Unit B) with the largest adjusted decrease in quality ($p=.0644$) from 2006 to 2007. Also depicted in the plot is the predicted regression down towards mean quality (0).

5.4 Simulation

Table 3 indicates that for all of the parameters except one (a_3), the 95% Credible Intervals cover the true parameter. This suggests “proof of concept” for our approach. Details regarding the estimates for quality also have promising properties. First, as theory dictates $\text{corr}(\hat{f}_{i1}, \hat{f}_{i2} - \hat{f}_{i2}^p) = -0.11$ which is much closer to 0 than $\text{corr}(\hat{f}_{i1}, \hat{f}_{i2} - \hat{f}_{i1}) = -0.55$ where $\hat{x} = E(x)$; illustrating advantage of Z^* over Z . Second, there were 0 values of Pr_i larger than 0.95 and 1 value smaller than 0.05 ($Pr_i=0.0383$). Since the simulation came from no change, this last value is considered a false positive.

The results for part (II) of the simulation are summarized in Figure 4 (the inverse logit of the expected value from the multiple regression, $\text{logit}^{-1}(\hat{B}_0 + \hat{B}_1 x_i + \hat{B}_2 (1 - x_i) (f_{1i} - f_{2i}) + \hat{B}_3 x_i (f_{1i} - f_{2i}))$). Neither regression line has a significant intercept different from 0 ($\hat{B}_0 + \hat{B}_1 = -.37$, $se=.335$, $p=.274$ & $\hat{B}_0 = -.032$, $se=.277$, $p=.910$) nor are the intercepts different from each other ($\hat{B}_1 = .34$, $se=.435$, $p=.440$). However, the $f_i=3$ slope is significantly negative ($\hat{B}_3 = -1.37$, $se=.108$, $p<.001$) and the $f_i=0$ slope is significantly higher than $f_i=3$ slope ($\hat{B}_2 - \hat{B}_3 = .54$, $se=.143$, $p<.001$). This demonstrates that it is faster to detect significant change when a unit's quality starts away from the mean and continues “against the grain.”

6. DISCUSSION & CONCLUSION

The NDNQI[®] is used for monitoring indicators of nursing quality, such as, total nursing hours per patient day, average registered nurse (RN) job enjoyment, fall rates, and pressure ulcer rates. From an NDNQI data repository, data are summarized and published in reports that enable nursing units to compare results of their units for each of the quality indicators with comparable units in other hospitals. There are two disadvantages to this reporting scheme. First, the sampling variability is not explicit. Second, the correlation among indicators is not reflected in the reporting. In previous work [17] we proposed a Bayesian report card system that involved reporting credible intervals as well as smoothed point estimates for a single time-point. Jones & Spiegelhalter [8] provide a methodology to adjust for ‘regression-to-the-mean’ when modeling recent changes in institutional performance for *one variable* at two time-points. We extend this model to allow for the four indicators above

to be correlated via a latent variable model across time, while at the same time adjusting for regression-to-the-mean. Generalizing this approach for other indicators, allowing multiple constructs, and automating the approach to handle large databases would be welcome improvements and topics suitable for future research.

In this paper we argued that the observed indicators are 'reflective' (not 'causal') because they are measured by the latent variable *quality*. A classical view of quality manifesting itself as reflective of structure, process, and outcomes is conceptually attractive.

However some may believe that there are theoretical reasons to question the case study model structure as specified. In Section 2 we argued on treating the indicators as reflective rather than causal of quality of care. However, one may disagree with our choice in treating total nursing hours per patient day (TNHPPD) as a reflective indicator of quality. We are specifically saying that improvement in the latent variable "quality of care" causes a change in TNHPPD. One may argue that it may be that an increase in TNHPPD would lead to an improvement in the quality (e.g. more nurses leading to fewer falls).

However, fall rates and pressure ulcers seem to be clear reflective indicators. One may argue that job enjoyment is less clear than fall and pressure ulcer rates but more than TNHPPD. It seems reasonable that working on a quality unit influences job enjoyment.

So what is the point of this part of the discussion? Two major points:

1. *There are alternative models to the pure reflective approach.* For example, a model with TNHPPD as a causal indicator and the others as reflective is an option to be considered. Other possible approaches include a cross-lag effect from TNHPPD at time 1 on quality at time 2. To empirically test these different models, one may want to consider performed the vanishing tetrad tests described in Bollen, Lennox & Dahly.
2. *Alternative models may lead to new approaches to adjusting for regression-to-the-mean.* If a model that is a mix of a causal indicator and reflective, then we would need to slightly adjust the overall approach we present. A major strength of this paper is that we considered the reflective model with four variables and demonstrated it with a case study and simulation. At the same time, a major limitation of the current paper is we did not develop the methodology for other types of composite indicators. This too is a topic for future research.

Further reading on the issue of causal and reflective can be found in for example in Fayers, Hand, Bjordal, & Groenvold [24] as well as Bollen [25].

Recent policy indicates a need for methodology that simultaneously: (1) builds composite indicators and (2) adjust for regression-to-the-mean. Our approach addresses this for reflective models and can be applied very generally to perform performance measurement. After careful composite indicator development and decisions, the methodology can be used now.

Acknowledgments

This work was supported in part by the National Database of Nursing Quality Indicators® (NDNQI®), from a contract with the American Nurses Association® and Frontiers: The Heartland Institute for Clinical and Translational Research CTSA UL1RR033179 which is now UL1TR000001 (first author). We thank editor, associate editor, and reviewers for comments that greatly improve our paper. Specifically one reviewer comment's improved focus, the second provided valuable detailed corrections to notation and text, and the third contributed greatly to the conceptualization of choosing reflective variables.

REFERENCES

1. Institute of Medicine. *To Err Is Human: Building a Safer Health System*. Washington, DC: The National Academies Press; 2001.
2. Institute of Medicine. *Keeping Patients Safe: Transforming the Work Environment of Nurses*. Washington, DC: The National Academies Press; 2004.
3. Shwartz M, Burgess JF, Berlowitz D. Benefit-of-the-doubt approaches for calculating a composite measure of quality. *Health Services and Outcomes Research Methodology*. 2009; 9(4):234–251.
4. Landrum MB, Normand SLT, Rosenheck RA. Selection of related multivariate means: monitoring psychiatric care in the Department of Veterans Affairs. *Journal of the American Statistical Association*. 2003; 98(461):7–16.
5. Teixeira-Pinto A, Normand SLT. Statistical methodology for classifying units on the basis of multiple-related measures. *Statistics in Medicine*. 2008; 27:1329–1350. [PubMed: 18181221]
6. Chen TT, Lai MS, Lin IC, Chung KP. Exploring and comparing the characteristics of nonlatent and latent composite scores: implications for pay-for-performance incentive design. *Medical Decision Making*. in press.
7. Daniels MJ, Normand SLT. Longitudinal profiling of health care units based on continuous and discrete patient outcomes. *Biostatistics*. 2006; 7(1):1–15. [PubMed: 15917373]
8. Jones HE, Spiegelhalter DJ. Accounting for regression-to-the-mean in tests for recent changes in institutional performance: analysis and power. *Statistics in Medicine*. 2009; 28:1645–1667. [PubMed: 19358144]
9. Bollen KA, Lennox RD, Dahly DL. Practical application of the vanishing tetrad test for causal indicator measurement models: An example from health-related quality of life. *Statistics in Medicine*. 2009; 28(10):1524–1536. [PubMed: 19266502]
10. Cai JH, Song XY, Hser YI. A Bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. *Statistics in Medicine*. 2010; 29(18):1861–1874. [PubMed: 20680980]
11. Donabedian A. The quality of care: How can it be assessed? *Journal of the American Medical Association*. 1988; 260:1743–1748. [PubMed: 3045356]
12. Hart S, Bergquist S, Gajewski B, Dunton N. Reliability testing of the National Database of Nursing Quality Indicator's pressure ulcer indicator. *Journal of Nursing Care Quality*. 2006; 21(3):256–265. [PubMed: 16816607]
13. Gajewski BJ, Hart S, Bergquist S, Dunton N. Inter-rater reliability of pressure ulcer staging: ordinal Probit Bayesian hierarchical model that allows for uncertain rater response. *Statistics in Medicine*. 2007; 26(25):4602–4618. [PubMed: 17393413]
14. Gajewski BJ, Boyle DK, Miller P, Oberhelman F, Dunton N. A Multilevel confirmatory factor analysis of the practice environment scale (PES): a case study. *Nursing Research*. 2010; 59(2):147–153. [PubMed: 20216017]
15. Simon M, Yankovskyy E, Klaus S, Gajewski B, Dunton N. Midnight census revisited: Reliability of patient day measurement and its impact on nursing quality indicators. *International Journal of Nursing Studies*. in press.
16. Simon M, Klaus S, Gajewski B, Dunton N. Falls reliability study. in preparation.
17. Gajewski BJ, Mahnen JD, Dunton N. Improving quality indicator report cards through Bayesian modeling. *BMC Medical Research Methodology*. 2008; 8:77. <http://www.biomedcentral.com/1471-2288/8/77>. [PubMed: 19017399]
18. Kuehl, RO. *Statistical Principles of Research Design and Analysis*. Belmont, CA: Duxbury Press; 1994.
19. Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. *British Medical Journal*. 2003; 326:1083–1084. [PubMed: 12750214]
20. Skrondal, A.; Rabe-Hesketh, S. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman and Hall CRC; Washington DC: 2004.
21. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. *Statistician*. 1994; 43:169–177.

22. Gelman A. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*. 2006; 1:515–534.
23. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society B*. 2002; 64:583–640.
24. Fayers PM, Hand DJ, Bjordal K, Groenvold M. Causal indicators in quality of life research. *Quality of Life Research*. 1997; 6(5):393–406. [PubMed: 9290306]
25. Bollen KA. Evaluating effect, composite, and causal indicators in structural equation models. *MIS Quarterly*. 2011; 35(2):359–372.

Figure 1. Similar to plot in Jones & Spiegelhalter demonstrating the effect of baseline deviation from the mean on the adjusted change, for nursing units A and B.

Figure 2.
Overall latent variable model.

Figure 3. Multivariate 'Quality' demonstrating the effect of baseline deviation from the mean on the adjusted change for nursing units A (improving) and B (declining).

Figure 4. Average p -values for the $n=100$ simulation study for variations in starting point (f_1) and difference ($f_2 - f_1$). The bottom line represents $p=0.05$ level.

Table 1

Med-surg summary statistics, n=290 units with full data from 2006 (time 1) to 2007 (time 2).

Measure		Year	Min	Max	Mean	Std
TNHPPD	$\exp(Y_{11})$	2006	5.50	13.94	8.24	1.40
	$\exp(Y_{12})$	2007	5.66	13.84	8.41	1.39
JE	Y_{21}	2006	2.00	4.62	3.49	0.49
	Y_{22}	2007	1.70	5.11	3.55	0.51
Fall Rate	$Y_{31}/w1$	2006	0.00	8.45	4.01	1.53
	$Y_{32}/w2$	2007	0.19	8.37	4.01	1.45
PrU Rate	$Y_{41}/n1$	2006	0.00	0.46	0.06	0.05
	$Y_{42}/n2$	2007	0.00	0.33	0.05	0.04

Table 2

Posterior distribution summaries for parameters from multivariate model of medical-surgical (n=290) units with full data from 2006 (time 1) to 2007 (time 2). (see Table 1).

Parameter		2.50%	median	97.50%	mean observed
$\exp(a_1)$	TNHPPD	8.06	8.21	8.36	8.33
a_2	JE	3.48	3.53	3.58	3.52
$\exp(a_3)$	Fall Rate	3.65	3.80	3.96	4.01
$\text{logit}^{-1}(a_4)$	PrU Rate	0.04	0.05	0.05	0.06
b_1	TNHPPD	0.0046	0.0156	0.0282	
b_2	JE	0.1792	0.2370	0.3023	
b_3	Fall Rate	-0.1641	-0.1221	-0.0787	
b_4	PrU Rate	-0.4851	-0.3758	-0.2766	
σ_1^2		0.0027	0.0032	0.0037	
$\sigma_{d_1}^2$		0.0198	0.0236	0.0283	
$\sigma_{d_2}^2$		0.0926	0.1276	0.1654	
$\sigma_{d_3}^2$		0.0737	0.0931	0.1169	
$\sigma_{d_4}^2$		0.2772	0.3718	0.4876	
$\sigma_d^2=\rho$		0.1453	0.4404	0.6459	

Table 3

Posterior distribution summaries for parameters from multivariate model of *simulated* medical-surgical ($n=100$) units with data from 2006 (time 1) to 2007 (time 2).

Parameter	Truth	2.50%	median	97.50%
a_1	2.105 ⁺	2.103	2.133	2.163
a_2	3.532 ⁺	3.477	3.562	3.655
a_3	1.335 ⁻	1.180	1.252	1.332
a_4	-3.035 ⁺	-3.263	-3.082	-2.905
b_1	0.016 ⁺	-0.001	0.020	0.043
b_2	0.239 ⁺	0.209	0.303	0.399
b_3	-0.127 ⁺	-0.245	-0.178	-0.120
b_4	-0.366 ⁺	-0.583	-0.397	-0.233
σ_1^2	0.003 ⁺	0.003	0.005	0.006
$\sigma_{d_1}^2$	0.024 ⁺	0.014	0.019	0.027
$\sigma_{d_2}^2$	0.128 ⁺	0.051	0.106	0.173
$\sigma_{d_3}^2$	0.094 ⁺	0.067	0.100	0.146
$\sigma_{d_4}^2$	0.374 ⁺	0.337	0.520	0.783
σ_d^2	0.426 ⁺	0.038	0.501	0.726

Note: '+' says that 95% CrI covers truth & '-' does not.