An Introduction to Multilevel Modeling for Anesthesiologists

Dale Glaser, PhD,* and Randolph H. Hastings, MD, PhD†

In population-based research, subjects are frequently in clusters with shared features or demographic characteristics, such as age range, neighborhood, who they have for a physician, and common comorbidities. Classification into clusters also applies at broader levels. Physicians are classified by physician group or by practice site; hospitals can be characterized by size, location, or demographics. Hierarchical, nested structures pose unique challenges in the conduct of research. Data from nested structures may be interdependent because of similarities among subjects in a cluster, while nesting at multiple levels makes it difficult to know whether findings should be applied to the individual or to the larger group. Statistical tools, known variously as hierarchical linear modeling, multilevel modeling, mixed linear modeling, and other terms, have been developed in the education and social science fields to deal effectively with these issues. Our goal in this article is to review the implications of hierarchical, nested data organization and to provide a step-by-step tutorial of how multilevel modeling could be applied to a problem in anesthesia research using current, commercially available software. (Anesth Analg 2011;113:877–87)

HYPOTHESIS TESTING AND EXPERIMENTAL DESIGN

Scientific research frequently centers on testing hypotheses with controlled experiments. The scientist develops a hypothesis based on established knowledge and previous data, and then designs an experiment that will deliver information about the validity of the hypothesis. In its simplest form, a controlled experiment involves comparing an experimental group with a control group. The 2 groups differ in 1 factor important for the hypothesis, with all other factors and conditions held constant. The data for the 2 groups are compared by *t* test, Mann-Whitney *U* test, linear regression, or another form of statistical analysis appropriate for the data and the experimental design.

Although designing a controlled experiment is generally straightforward for bench laboratory research, the matter can be problematic for population-based human subjects research. Human subjects vary in innumerable ways that are impossible to control directly. They have a wide range of physical characteristics such as age, and they differ in genotype, ethnicity, socioeconomic group, and health status. They also vary in education, employment, residence location, environmental exposure, and other aspects of personal history. The uncontrolled sources of variability may require an investigator to increase sample size, balance subject characteristics as much as possible, and/or use a multivariate analysis for the purpose of accounting for the effects of covariates in the proposed model.

HIERARCHICAL STRUCTURES

In addition to these issues, subjects are frequently in clusters with shared features or demographic characteristics. For example, individuals may be nested by the area where they live, who they have as a physician, or their ethnic group. Nested structures are hierarchical because they exist at many levels; individuals can be nested in groups and groups of people can be clustered into larger structures. For a cross-sectional design, the hierarchy may split into a micro level, referring to individuals, and a macro level, referring to larger groups. Hierarchical structures are embedded in a wide array of disciplines, ranging across medicine, sociology, education, psychology, and business.[1] In education, a student is nested within class, which in turn is nested within school, then neighborhood. In anesthesiology research, patients are nested according to the anesthesiologist providing their care, anesthesiologists may be nested within their institution or hospital, and the hospital may be nested by size, location, profit status, or other factor.

CONSEQUENCES OF NESTING

It is well known from contextual studies that some degree of interrelationship characterizes nested structures within a given level and that the data may be interdependent. Interdependence may occur if the subjects belong to the same group, live in close vicinity to each other, share experiences, or are studied in the same short timeframe. For example, practice patterns may be homogeneous within an anesthesiology group and heterogeneous between groups.[2] Thus, patients under the care of one group of anesthesiologists at the same hospital could be expected to respond more closely to each other than to patients of another group at a different hospital because of similarities in how the anesthesia is delivered.

From the *School of Nursing, University of San Diego; and †Anesthesiology Service, VA San Diego Healthcare System, San Diego, California.

Accepted for publication March 3, 2011.

Supported by the Anesthesia Patient Safety Foundation.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.anesthesia-analgesia.org).

Dale Glaser, PhD, is currently affiliated with Glaser Consulting, San Diego, CA; and Randolph H. Hastings, MD, PhD, is currently affiliated with the University of California, San Diego, CA.

Reprints will not be available from the authors.

Address correspondence to Randolph H. Hastings, MD, PhD, VA San Diego Healthcare System, 3350 La Jolla Village Dr. 125, San Diego, CA 92161. Address e-mail to rhhastings@ucsd.edu.

DOI: 10.1213/ANE.0b013e3182198a01

Ignoring the consequence of hierarchical structures has been shown to result in underestimation of the standard errors, inflate rate of type I errors as a consequence, and lead to wrongful rejection of the null hypothesis when it is true. Nonindependence also distorts the estimated error variance, confidence intervals, *P* values, and effect sizes, potentially leading to an increase in type II errors and a concomitant decrease in power.[3] Whatever the direction of the error, a solution for interdependent data is to calculate a separate estimate of variance for each level of clustering.

Besides the interdependence errors, hierarchical structures also create problems in interpreting results. Suppose the researcher has dealt with the hierarchy by aggregating the data for individuals within each cluster, directing the analysis at the macro level. What conclusions can be drawn regarding the individuals in the study based on this analysis? The investigator is at risk of making what is called the ecological fallacy, using inferences derived from the group to reach conclusions about the individual. Conversely, disaggregating the data to evaluate results at the micro level may result in the atomistic fallacy, extending inferences from the individual to the group.[4,5]

MULTILEVEL MODEL ANALYSIS

Fortunately, the statistical theory for clustered, hierarchical data is well established and has a prominent role in dealing with these issues. Many descriptors and titles have been applied to designs that contain hierarchical structures: mixed linear models, multilevel linear models, mixed-effects models, random-effects models, random-coefficient regression models, and/or covariance components models.[6] An oft-used name, hierarchical linear modeling, is actually the title of a well-known software (HLM).[7] For purposes of this article, we refer to hierarchical designs as multilevel models (MLM), a common term used in a diverse range of disciplines.

MLM can analyze hierarchical structures of all kinds in one unified framework. Factors can be fixed (e.g., male versus female) or random (e.g., sampling of hospitals from a population of institutions). The design can be cross-sectional or follow a multiwave/longitudinal structure, whereby time is nested within the individual.[4] A distinct advantage of MLM is the ready ability to handle missing data or imbalanced designs. This is especially beneficial for longitudinal designs, where it is not uncommon to have missing time points. The default action for many non-MLM software programs is to delete all the data from an individual case if even one item is missing (i.e., listwise deletion). MLM will use all available data, even if there is only one wave of data for a given individual. However, different missing data techniques (e.g., full information maximum likelihood, multiple imputation) must still be pursued depending on why the data are missing (e.g., missing completely at random, missing at random), so it is paramount that the researcher attempt to discern the reason.[8–10]

Mixed linear models have been used in human subjects research for several decades,[5,6] but procedures were unwieldy until the advent of MLM software in the 1980s. The sociology and education fields embraced these advances readily and rapidly.[7] The adoption of MLM has progressed more slowly in medical specialties, including anesthesiology, but has become more prominent recently. For example, an article in this issue of *Anesthesia & Analgesia* describes an MLM-based analysis of how factors such as previous training experience, occupation, training equipment, and number of practice attempts affect development of laryngoscopy skill.[11] Within the past year, articles in other journals have described the use of MLM methods to study factors associated with anesthesiologist assistance with colonoscopy and factors affecting recovery from volatile anesthesia.[12,13] The analysis methods and the related concepts may not be familiar to many anesthesiologists. Thus, the goal for this article is to serve as a primer for the statistical and methodological advances for testing these relatively complex data and design structures. We will proceed by demonstrating a small MLM analysis in an anesthesia scenario.

MLM TUTORIAL USING SAMPLE DATA

Data Source

For illustrative purposes, we obtained data from a database often used in education research, the National Education Longitudinal Study of 1988 (NELS:88). The original database is publicly available (http://nces.ed.gov/surveys/nels88/) and contains aggregate data without personal identifying information. The data used for the statistical analysis in this tutorial are available as online supplemental data (see Supplemental Digital Content 1 and 2: Supplemental Table 1, http://links.lww.com/AA/A267, and Supplemental Table 2, http://links.lww.com/AA/A268).

Hypothetical Study

The identity of the variables in the original study is immaterial in conducting the MLM. Therefore, we will perform the analysis as if it were an anesthesiology study to make the text more interesting to the reader. Let us suppose that the hypothetical project is a study of factors that affect recovery from general anesthesia. The database includes 7185 subjects at the micro level who will be patients, clustered at the macro level among 160 anesthesiologists who provided their care. We will designate a continuous outcome variable as the quality of recovery (QR) score. A study of recovery from anesthesia would probably collect data for a large number of micro-level demographic factors, many patient-specific variables regarding health and physical condition, and information concerning the anesthetic and surgery. The MLM method is the same whether many or few variables are analyzed. To make the analysis easier to follow, we decided to include only 2 patient variables, "Preoperative Physical Status" score (Preop PS) for each patient and age. Preop PS is intended to be a hypothetical continuous version of the ASA classification. We used a random number generator set to a predetermined mean and SD of 56 ± 9 years to assign ages to the patients. We did not purposefully induce a correlation between age and outcome, but did allow age to vary across the anesthesiologist clustering variable, with means ranging from 49.7 to 59.1 years. Finally, we decided that the binary macro-level modifier would refer to the anesthesiologists' hospital setting, either a large metropolitan medical center or a nonmetropolitan hospital.

Model Construction

At this point, we have selected the relevant variables and excluded irrelevant ones, the first in an array of decisions to be made in performing MLM. The next step is to choose the proper statistical model. We selected a model with a continuous level outcome using an identity link function (implying normally distributed outcome and residuals) because it was appropriate for our data and problem. Also, the analysis for this type of outcome will be familiar to most readers who use multiple linear regression. Many of the same assumptions and issues that are central to regression models (e.g., normality of residuals, multicollinearity, homoscedasticity) also pertain to MLM.[14] After establishing the model to test, the investigator should consider whether logarithmic or similar data transformations for the outcome are needed to normalize the residuals, contemplate the need for nonlinear terms (frequently quadratic) to compensate a nonlinear relationship between the predictor and outcome, decide whether multiplicative interaction terms should be included within or across levels, and pick the algorithm to adjust for missing observations. Adjustment for missing data can be accomplished with full information maximum likelihood or restricted maximum likelihood algorithms, which are incorporated in many of the software packages. We picked a restricted maximum likelihood algorithms estimator because it adjusts for the uncertainty of the fixed effects, and has shown many favorable properties.[1] Our data needed no transformation because the residuals were normally distributed and we expected a linear relationship between predictor and outcome. We discuss interaction terms below.

Many software programs use listwise deletion as the default missing data technique. This technique is problematic unless it can be shown that the absent data are missing completely at random, the most restrictive assumption. If the data are missing at random, not as restrictive as missing completely at random, then multiple imputation or full information maximum likelihood are plausible alternatives. The least restrictive assumption would be that data are not missing at random. If this is the case, the mechanism for the missing data cannot be ignored and more complex methods, such as selection models or pattern mixture models, would be needed to account for not missing at random in the analysis. Further discussion is beyond the depth of this review. The reader may learn more about using maximum likelihood algorithms in reviews by Enders.[9,15]

In regression, continuous level predictors are mean centered so as to render the regression coefficients more interpretable and as well to decrease collinearity.[16] Centering is also an issue in mixed linear models. For purposes of this review, grand mean centering (score − grand mean) was conducted. The following citations provide more information regarding centering options.[17–19]

Software

For this article, the hypothesized model is tested by HLM 6.08 (Scientific Software, International, Lincolnwood, IL), an MLM dedicated software, SPSS 18.0.2 mixed linear model option (IBM, Chicago, IL), and Stata 11 (StataCorp LP, College Station, TX), a frequently used software in the medical sciences. SPSS, a more general-purpose software akin to SAS, is used primarily to obtain information-theoretic indices that are not included in the HLM output. These indices are discussed in a following section. Detailed outputs are provided and the fixed and random coefficients are interpreted.

In addition to these programs, several software packages can be used to test multilevel models. The rationale for choosing a specific package depends to an extent on personal preference and may also vary with the research discipline. For example, many investigators in the medically related sciences use MLwiN, whereas researchers in education and psychology frequently use HLM. Mixed linear modeling is also embedded in many familiar softwares such as SAS, SPSS, and Stata or in structural equation modeling software such as Mplus. Similar algorithms are used across softwares; thus, one software is not necessarily superior to the other. However, distinct differences may be found in terms of capability. For instance, HLM can be used for a variety of outcomes (e.g., multinomial, binary, counts, ordinal), whereas the mixed modeling option in SPSS (at least up to the current version) is restricted to testing outcomes that are continuous (i.e., assumedly interval level).

Table 1. Descriptive Statistics for Level 1 and Level 2 Variables in the Hypothetical Multilevel Modeling Analysis of Quality of Recovery in Anesthesiology Patients

Variables	Mean ± SD	Range (minimum–maximum)
Level 1 (n_i = 7185)		
Preop PS score	4.994 ± 0.661	1.34–7.85
Age, y	55.08 ± 9.05	19.7–89.0
QR score	12.748 ± 6.878	−2.83 to 24.99
	Frequency (*n*)	**Percentage**
Level 2 (n_j = 160)		
Nonmetro hospital	90	56.3
Large metro hospital	70	46.8
	Pearson correlation	**Significance**
Level 1 correlations		
Preop PS versus age	0.012	0.301
Preop PS versus QR	0.21	0.0001
Age versus QR	−0.038	0.001

PS = physical status; QR = quality of recovery.
Significance for correlations is 2-tailed.

Descriptive Statistics

A good practice is to examine the data before engaging in the detailed modeling work. Table 1 presents the descriptive statistics for the level 1 and level 2 variables in our problem. Level 1, the micro level, consists of the n_i = 7185 patients, whereas level 2, the macro level, refers to the n_j = 160 anesthesiologists. The descriptive statistics at level 1 include a mean ± SD QR score of 12.75 ± 6.88, an average age of 55.98 ± 9.05 years, and an average Preop PS score of 4.99 ± 0.66. The correlations of the age and Preop PS predictors with the QR outcome are −0.04 and 0.21, respectively. Hospital type was the only level 2 predictor. The anesthesiologists were split with 56.2% (n_j = 90) working at a nonmetro hospital and 43.8% (n_j = 70) at a large metro institution. As with regression models, collinearity (i.e., overlap or interdependence) of the predictors should be

```
HLM 6.08 Output:  Model M1, Unconditional Model

-----------------------------------------------------------------------------
                                    Standard              Approx.
   Fixed Effect         Coefficient  Error      T-ratio    d.f.     P-value
 -----------------------------------------------------------------------------
 For       INTRCPT1, B0
    INTRCPT2, G00          12.636972   0.243628    51.870       159    0.000
 -----------------------------------------------------------------------------


 Final estimation of variance components:
 -----------------------------------------------------------------------------
-
 Random Effect           Standard      Variance     df    Chi-square  P-value
                         Deviation     Component
 -----------------------------------------------------------------------------
-
 INTRCPT1,       U0        2.93501       8.61431   159    1660.23259    0.000
  level-1,       R         6.25686      39.14831
 -----------------------------------------------------------------------------
-


 Statistics for current covariance components model
 --------------------------------------------------
 Deviance                         = 47116.793469
 Number of estimated parameters = 2
```

Figure 1. Output from HLM 6.08 analyzing the unconditional means model for the hypothetical anesthesiology study, model M1. B0 is the mean patient quality of recovery score averaged across the 160 anesthesiologists. U_0 and R are the variances ascribed to between-patient and between-anesthesiologist differences. U_0 and R are used to calculate the intraclass coefficient, an indicator of whether multilevel modeling is appropriate for the data (see text).

preferably minimal. Accordingly, the Pearson correlation coefficient between age and PS score was $r = 0.012$, not significant in a 2-tailed test. The level at which collinearity becomes an issue is a matter of opinion. Cohen et al.[16] indicate that collinearity becomes problematic if $r^2 > 0.9$, but we would worry about instability in the solution with that cutoff level. We are generally comfortable if $r^2 < 0.25$.

Testing the Unconditional Means Model

It is often recommended that a sequence of models be examined for fit to ascertain the best model to select.[20] At the simplest level, one often commences by fitting data to the unconditional means model, a rudimentary model that omits predictors. Its primary objective is to investigate the extent of the heterogeneity between the clusters, thereby establishing the rationale for analyzing an MLM. A regression without predictors generates an equation with no slope and an intercept that is equal to the mean of the outcome variable. The unconditional model averages the outcome variable for the level 1 units (the patients in our problem) across the level 2 units (the anesthesiologists) and partitions the variance between level 1 and level 2. The between-cluster variance then represents the heterogeneity between the clusters at level 2. For the analysis of our dataset, the equation for the unconditional means model for our dataset is QR = B0, where B0 is the mean QR score. We refer to this analysis as model "M1." We provide the SPSS code for this model and subsequent models in the Appendix (see Supplemental Digital Content 3, http://links.lww.com/AA/A269).

Figure 1 displays the output produced by HLM 6.08 in analyzing model M1. It includes the following estimates:

1. The level 2 intercept for the QR score, B0. B0 is the mean QR for each anesthesiologist, the level 2 units, and G00 is the average intercept across the 160 anesthesiologists. A statistician would label the intercept, β_{0j}, with the subscripted "0" identifying the quantity as an intercept and the subscripted "j" linking it to level 2 (j anesthesiologists). The HLM output shows that the average intercept is 12.64 (SE = 0.24) and that it is significantly different from 0 ($P < 0.05$).
2. The random effect for QR at level 1, R. R is the variance estimate for the patients (level 1). In our example, $R = 39.15$.
3. The increment in intercept associated with level 2, U_0. U_0 is the variance component estimate mentioned in the previous paragraph that captures the variation in intercept, QR, between anesthesiologists at level 2. For our data, $U_0 = 8.61$.

What makes the unconditional means model particularly useful is the computation of the intraclass correlation coefficient (ICC). The reference to correlation arises because the ICC gives a measure of how homogeneous the data are within a level 2 cluster unit, i.e., how well the data within a unit correlate with each other, compared with between clusters.[21] The ICC value increases as heterogeneity increases. The ICC can be computed at any stage of modeling, but it is particularly informative for the unconditional means model in furnishing evidence of the extent of a clustering effect. In our analysis, the data in question are the QR values, "within cluster" refers to data for a given anesthesiologist, and the between-cluster comparison would examine QR values from different anesthesiologists. In model M1 with our data, the ICC is calculated by dividing the level 2 variance, U_0, by the total variance, $U_0 + R$, as follows, ICC = 8.61/(8.61 + 39.15) = 0.18. Thus, 18% of the variability is attributable to differences between anesthesiologists, rather than within patient differences. Although there is no invariable rule as to what constitutes a high ICC, many authorities in education and other disciplines that use MLM frequently opine that an ICC of 5% is substantive evidence of a clustering effect. On that

Table 2. Average Quality of Recovery Score for Selected Anesthesiologists in the Study, Showing Substantial Variation Across Anesthesiologists

Anesthesiologist ID	No. of patients	QR (mean ± SD)
1224	47	9.72 ± 7.59
1288	25	13.51 ± 7.02
1296	48	7.64 ± 5.35
1308	20	16.26 ± 6.11
1317	48	13.18 ± 5.46
1358	30	11.21 ± 5.88
1374	28	9.73 ± 8.36
1433	35	19.72 ± 3.88
1436	44	18.11 ± 4.55
1461	33	16.84 ± 6.95
1462	57	10.50 ± 6.31
1477	62	14.23 ± 7.15
1499	53	7.66 ± 6.34

QR = quality of recovery.

basis, MLM analysis seems warranted for our dataset. Conversely, there would not have been the same rationale for a multilevel approach had the ICC been substantively <0.05. Table 2, a sample from a few of the 160 anesthesiologists, shows distinct interanesthesiologist differences in the mean QR, furnishing support for the relatively high ICC and strengthening the case for proceeding with MLM.

Testing the Research Question with More Complex Models

We are now ready to investigate whether Preop PS and age have an impact on the QR score using MLM methods. The procedure will be to fit the data to models that add in progression level 1 predictors, level 2 predictors, and possibly interaction terms. Deciding whether these models represent improvements and selecting the "best" model is somewhat subjective. However, information-theoretic indices, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), can be used to compare how the data fit 2 successive models.[22,23] A reduction of the AIC or BIC may indicate a more favorable result for the new model. In the unconditional means model for our data, the AIC = 47,121 and the BIC = 47,135. We will repeat the calculations with subsequent models, looking for reductions in the indices to determine whether adding the predictors provides a better fit to the data. The indices can be determined by the SPSS mixed linear model option, Mplus software, and SAS proc mixed.

Another option is to compare the deviance statistic, calculated as −2 times the difference in the logarithms of the fitted and full parameters in the model. The deviance statistic is calculated by HLM 6.08 (see outputs in Figs. 1 and 3 for examples), but AIC and BIC are not. If one uses the full information maximum likelihood estimator (as opposed to the default restricted maximum likelihood), one can take the differences of the deviances between models and compare the results with a χ^2 table. A significant finding would furnish support for the better-fitting model. The variance components and information-theoretic indices for each of the tested models are furnished in a table at the end of the analysis.

MLM Analysis of Level 1 Predictors

At this juncture, 2 level-1 predictors, patient age and PS score, are entered into the analysis as model M2. The equation for this model is $QR_{ij} = B_0 + B_1 \cdot (AGE_{ij}) + B_2 \cdot (PS_{ij}) + R$ where the ij subscript refers to the individual patient, B_1 and B_2 are the partial regression coefficients related to age and PS, respectively, and R is the level 1 residual. From the HLM output in Figure 2, both predictors are significant (controlling for other model covariates): age ($B_1 = -0.033 \pm 0.008$, mean ± SE, $P < 0.05$) and PS score ($B_2 = 2.20 \pm 0.108$, $P < 0.05$). Interpreting the results for age, we see there is a slight decrease in the outcome (0.03) concomitant with a 1-unit increase in age. We see for each unit change in the PS score, there is a 2.20 increase in the outcome. Stata 11.0 gives similar results when analyzing this model.

There is a slight increase in the variance component of the intercept in this model compared with the unconditional means model. U_0 increases from 8.61 to 8.69 (output not shown). However, R, the level 1 residual, decreases from 39.15 to 36.92. Based on this reduction, one can calculate the proportional reduction in residual variance, an estimate of effect size, using what is referred to as a pseudo R^2 statistic.[22] The computation is as follows: (39.15 − 36.92)/39.15 = 0.057; thus, incorporating the 2 level-1 predictors in the model leads to a 5.7% reduction in error. Further evidence is a decrease in the AIC from 47,121 for the unconditional means model to 46,719 for this model. Similarly, the BIC decreases from 47,135 to 46,733. Hence, adding age and Preop PS score as predictors appears to improve the model fit.

HLM 6.08 Output: Model M2 with Level 1 Predictors

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	12.636329	0.244704	51.639	159	0.000
For AGE slope, B1					
INTRCPT2, G10	-0.032926	0.007996	-4.118	7182	0.000
For PREOP_PS slope, B2					
INTRCPT2, G20	2.196667	0.108538	20.239	7182	0.000

Figure 2. HLM 6.08 analysis of the model with the level 1 predictors, model M2. Quality of recovery depends significantly on both physical status (PS) score and age.

A **Output from HLM 6.08, Model M3 with level 2 predictors**

```
Final estimation of fixed effects:
----------------------------------------------------------------------------
                                    Standard              Approx.
   Fixed Effect         Coefficient   Error      T-ratio   d.f.      P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          11.394602   0.293317    38.847       158    0.000
   HOSPITAL, G01           2.802168   0.439999     6.369       158    0.000
For      AGE slope, B1
   INTRCPT2, G10          -0.032769   0.007995    -4.099      7181    0.000
For PREOP_PS slope, B2
   INTRCPT2, G20           2.196643   0.108542    20.238      7181    0.000
----------------------------------------------------------------------------
Final estimation of variance components:
-----------------------------------------------------------------------------
Random Effect           Standard      Variance     df    Chi-square  P-value
                        Deviation     Component
-----------------------------------------------------------------------------
INTRCPT1,       U0        2.59981       6.75899   158    1377.76155    0.000
 level-1,       R         6.07684      36.92793
-----------------------------------------------------------------------------
Statistics for current covariance components model
--------------------------------------------------
Deviance                        = 46676.568369
Number of estimated parameters = 2
```

B **QR Score Averaged for non-metro and large metro hospitals**

Report

QRc_mean

Hospital	Mean	N	Std. Deviation
.00 non-metro	11.3895	90	2.83022
1.00 large metro	14.2038	70	2.74698
Total	12.6208	160	3.11765

C **Output from SPSS 18.0.2 with level 2 predictors**

Estimates of Covariance Parameters[a]

Parameter		Estimate	Std. Error	Wald Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Residual		36.927964	.623306	59.245	.000	35.726293	38.170054
Intercept [subject = Anes_ID]	Variance	6.758393	.868531	7.781	.000	5.253568	8.694259

a. Dependent Variable: QualRec Quality Recovery Score.

Figure 3. A, HLM 6.08 analysis of model with level 2 predictors, model M3. The type of hospital (metro versus nonmetro) has a significant effect on quality of recovery (QR), as noted by the G_{01} slope. The level 1 variance component, R, decreases compared with the value in the previous model as shown in Figure 2, suggesting an improvement in the fit to the data. B, QR scores differed between large metro and nonmetro hospitals by 2.8, as predicted by the G_{01} coefficient given in part A. C, SPSS 18.0.2 calculated the same intercept variance component and level 1 residual as HLM 6.08 and displayed additional parameters not reported by HLM 6.08.

Modeling the Level 2 Predictor

At this stage, hospital is evaluated as a level 2 predictor in a model along with the level 1 predictors, age and Preop PS, generating model M3. To account for the level 2 effects, the intercepts and slopes from the level 1 regression equations for the j = 160 anesthesiologists become outcomes that vary with the level 2 predictor, hospital type. The equation for intercept can be written, $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (hospital) + u_{0j}, where β_{0j} is the level 1 intercept for the jth anesthesiologist, γ_{00} is the regression intercept, γ_{01} is a slope for the effect of hospital type, (hospital) is a binary variable that can take 1 of 2 values depending on whether the hospital is large metro versus nonmetro, and u_{0j} is the residual. Similarly for level 1 slopes, the equation would be $\beta_{1j} = \gamma_{10} + \gamma_{11}$ (hospital) + u_{1j}, where β_{1j} is the slope of the age regression equation for the jth anesthesiologist and the other parameters are defined analogously to the parallel parameters for the intercept equation. Similar equations would be written for the PS slopes as outcomes, substituting a 2 in place of the 1 in the initial position in each of the subscripts used in the age slope equation.

The HLM output for model M3 is presented in Figure 3. (Gammas in the output are written as "Gs" and betas as "Bs".) The output shows that the 2 level-1 predictors are still significant: the age slope is $G_{10} = -0.033 \pm 0.008$, $P < 0.05$, and Preop PS slope is $G_{20} = 2.20 \pm 0.108$, $P < 0.05$. Moreover, the level 2 predictor is significant: $G_{01} = 2.8 \pm 0.44$, $P < 0.05$. The G_{01} coefficient signifies that there is a 2.8 units difference in the average QR between groups, with a higher intercept captured by the large metro hospitals. This

A **HLM 6.08 output: Model M4, QR vs. PS slope vary with anesthesiologist**

```
The outcome variable is  QUALREC
 Final estimation of fixed effects:
----------------------------------------------------------------------------
                                        Standard              Approx.
   Fixed Effect          Coefficient    Error      T-ratio    d.f.      P-value
----------------------------------------------------------------------------
For       INTRCPT1, B0
   INTRCPT2, G00          11.105992    0.289445    38.370       158     0.000
   HOSPITAL, G01           3.460769    0.422724     8.187       158     0.000
For      AGE slope, B1
   INTRCPT2, G10          -0.032162    0.008384    -3.836       159     0.000
For PREOP_PS slope, B2
   INTRCPT2, G20           2.219684    0.127283    17.439       159     0.000
----------------------------------------------------------------------------
Final estimation of variance components:
-----------------------------------------------------------------------------
Random Effect            Standard      Variance     df    Chi-square  P-value
                         Deviation     Component
-----------------------------------------------------------------------------
INTRCPT1,       U0        2.62033       6.86611   158    1375.56853    0.000
     AGE slope, U1        0.03131       0.00098   159     177.64772    0.148
PREOP_PS slope, U2        0.82141       0.67471   159     213.01097    0.003
 level-1,       R         6.04583      36.55210
-----------------------------------------------------------------------------
```

B **Corresponding output of SPSS 18.0.2**

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	1.797214	.796073	154.800	2.258	.025	.224645	3.369783
age	-.032199	.008357	136.510	-3.853	.000	-.048726	-.015673
PreOp_PS	2.219395	.127365	153.272	17.425	.000	1.967777	2.471014
Hospital	3.457434	.422845	151.940	8.177	.000	2.622018	4.292850

a. Dependent Variable: QualRec Quality Recovery Score.

C **Information-theoritic indices to assess model fit**

Information Criteria[a]

-2 Restricted Log Likelihood	46656.543
Akaike's Information Criterion (AIC)	46670.543
Hurvich and Tsai's Criterion (AICC)	46670.558
Bozdogan's Criterion (CAIC)	46725.697
Schwarz's Bayesian Criterion (BIC)	46718.697

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: QualRec Quality Recovery Score.

Figure 4. A, Output for model considering slopes for quality of recovery (QR) versus physical status (PS) and age within anesthesiologists to be random effects, model M4. The PS score slope appears to vary significantly with anesthesiologist, shown by variance component U2, but the age slope does not. B, Output generated by SPSS 18.0.2 for the same model. Parameters that are common with the HLM 6.08 output match exactly. C, Information-theoretic indices, the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), decrease with this model compared with the previous model, indicating a better fit to the data adding a stochastic component for anesthesiologist slope. The calculations were performed using SPSS, because HLM 6.08 does not provide the option to calculate AIC or BIC.

is verified in Figure 3B, where QR is14.20 for metro versus 11.39 for nonmetro hospitals.

In terms of the fit, we see a substantive decrease in the variance component of the intercept compared with the prior model from 8.69 to 6.76, even though the level 1 residual R increases by 0.01 from 36.92 to 36.93 (Fig. 3A). Because of the reduction for the intercept, the pseudo R^2 statistic would demonstrate that incorporating the hospital macro-level predictor into the model reduces the variance component by 22.2%. The AIC and BIC decrease appreciably with this model, now at 46,682 and 46,696, respectively, presenting further evidence of improved model fit.

For purposes of comparison, Figure 3C shows the output for the model random effects generated by SPSS 18.0.2 software. The U_0 and R values are the same as listed in Figure 3A, but other parameters that were not provided by HLM 6.08 are included, such as Wald z and the 95% confidence intervals.

Anesthesiologist as a Stochastic Predictor

One might want to investigate whether the relationship between the level 1 predictors and QR varied across different anesthesiologists. This can be accomplished by estimating the slopes, QR versus age or QR versus PS score, for patients under the care of each anesthesiologist and calculating the variance component attributable to the anesthesiologist. Instead of treating the slopes as fixed parameters, they are allowed to vary randomly across anesthesiologists. The slopes within anesthesiologists are being considered as random effects in what is often called a random coefficient model. A statistician would also state that the model added a stochastic or probabilistic property to the slopes.

When this analysis is performed as model M4, all level 1 and level 2 predictors remain significant, as shown in Figure 4A in the section labeled "Fixed Effect." The output for the same parameters from SPSS 18.0.2 has a similar result (Fig. 4B). The Random Effect section of Figure 4A shows the

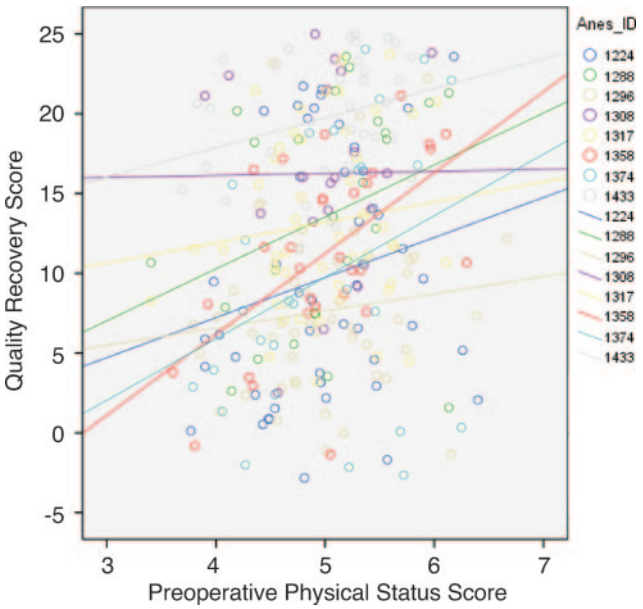

Figure 5. Plot of quality of recovery (QR) versus physical status (PS) score separated by individual anesthesiologists, as noted by the different colors. Slopes and intercepts vary with the anesthesiologist, as indicated by the multilevel modeling analysis shown in Figure 4.

variances. The anesthesiologist-dependent variance component for age, listed as U1, is not significant ($P = 0.148$), but significance was obtained for the variation in the preop score slope by anesthesiologist ($P = 0.003$). Interestingly, although the AIC somewhat decreases with the addition of the stochastic parameters for the slopes, the BIC worsens a bit (Fig. 4C). Given that the variance component for age did not impact fit, the model was retested analyzing random effects only for the PS score slope and not the age slope (model M5). Now we find a slight decrease in the AIC (46,666.48) and the BIC (46,694). The inference from this analysis might be that the impact of Preop PS score on patient QR depends to some extent on the anesthesiologist in this dataset, but the impact of age does not. The effect of modeling PS score as a stochastic parameter is illustrated by the pattern of random intercepts and slopes for the QR versus PS plots from a few of the anesthesiologists shown in Figure 5. Slopes and intercepts differ from one anesthesiologist to the next.

Modeling Interactions Across Levels

Model M6, the final model to be considered, involves adding a cross-level interaction to model M5. In this instance, we are interested in examining whether the level 1 predictor, PS score, interacts with the level 2 predictor, hospital. Here we gain further appreciation for the advantages of a multilevel approach through our ability to model the synergistic relationship between micro- and macro-level variables in one model. For model M6, a significant interaction was observed between PS score and hospital: $G21 = -1.34 \pm 0.233$, $P < 0.05$ (Fig. 6). However, the variance component for the PS slope is no longer significant ($P = 0.177$), although we see a slight reduction in the residual and variance component for the intercept. This demonstrates the dynamic nature of multilevel models. We also observe a decrease in the information-theoretic indices, indicating that adding the cross-level interaction may aid in improving model fit. Although only a small random sampling, we see in Figure 7 the nature of the interaction between hospital and Preop PS score; the nonmetro hospital evidences a stronger slope for the Preop PS/QR score relationship than does the metro hospital category. Interactions between age and hospital type have not been analyzed but could be estimated in a similar manner.

Choice of Model

Table 3 summarizes relevant parameters as we moved from one model analysis to the next. The AIC and BIC information-theoretic criteria and the variances generally decreased as we advanced between models. This pattern suggests that the added components improved model fit to the data and could identify important relationships. However, interpreting the implications of data modeling always requires an educated appraisal of the scientific questions. In some cases, changing the model may not lead to an obvious improvement in fit, or the indices may provide a mixed picture. Such situations require the investigator to exercise judgment or search for additional objective indications for how to proceed. Practical considerations would include whether changing the model led to a substantive change in conclusions or whether the data quality and quantity were sufficient to pursue the more complex model. Examining the data may prove helpful. For example, plotting QR score versus age data separated by anesthesiologist suggested that the age slope does not vary with anesthesiologist and provides an indication that elements of model M4 could be omitted. Evaluating a different model empirically, as we did with model M5, could accomplish the same goal. Analyzing the data from different perspectives could be useful. In this regard, comparing average QR scores for different anesthesiologists in Table 2 reinforced the impression given us by the unconditional means model M1 that MLM was warranted for our dataset. Finally, the findings of previous investigations could aid in making decisions.

One primary assumption in regression models is referred to as model specification. It entails including the proper variables and covariates in the postulated model and excluding irrelevant ones. In addition, the model specification assumption implies that proper mathematical modeling with nonlinear terms (quadratic and others) and multiplicative terms (interactions) has been used. The investigator has to balance efforts to optimize model fit against the risk of adding undue complexity, which may compromise model parsimony or result in overfitting the model.

Was MLM Valuable in This Analysis Over Least Squares Methods?

One might ask whether an ordinary least squares regression (OLS) would have led to the same inferences as MLM in this tutorial. An argument could be made for using multiple regression if one thought that the evidence for clustering were negligible, although a decision would still be needed whether to perform the regression at the micro (patient) or macro (anesthesiologist) level. We have repeated our data analysis with a least squares approach and present the outcomes along with the MLM results in Table 4 to illustrate the differences. The parameter estimates for

A **HLM Output: Model M6, Interactions between Level 1 and Level 2**

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. d.f.	P-value
For INTRCPT1, B0					
INTRCPT2, G00	11.395499	0.293138	38.874	158	0.000
HOSPITAL, G01	2.804294	0.439762	6.377	158	0.000
For AGE slope, B1					
INTRCPT2, G10	-0.032711	0.007969	-4.105	7180	0.000
For PREOP_PS slope, B2					
INTRCPT2, G20	2.807224	0.156041	17.990	158	0.000
HOSPITAL, G21	-1.339672	0.235320	-5.693	158	0.000

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	2.60008	6.76044	158	1389.74962	0.000
PREOP_PS slope,	U2	0.54400	0.29594	158	174.32325	0.177
level-1,	R	6.05058	36.60949			

B **SPSS Output**

Estimates of Fixed Effects[a]

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	-.822088	.865129	233.842	-.950	.343	-2.526531	.882356
age	-.032746	.007969	7051.750	-4.109	.000	-.048367	-.017124
PreOp_PS	2.807583	.154164	140.593	18.212	.000	2.502803	3.112362
Hospital	9.497549	1.128345	144.411	8.417	.000	7.267344	11.727754
PreOp_PS * Hospital	-1.340248	.232600	150.662	-5.762	.000	-1.799826	-.880669

a. Dependent Variable: QualRec Quality Recovery Score.

C **Model Fit**

Information Criteria[a]

-2 Restricted Log Likelihood	46629.571
Akaike's Information Criterion (AIC)	46637.571
Hurvich and Tsai's Criterion (AICC)	46637.577
Bozdogan's Criterion (CAIC)	46669.087
Schwarz's Bayesian Criterion (BIC)	46665.087

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: QualRec Quality Recovery Score.

Figure 6. Modeling interactions between the level 1 physical status (PS) score and the level 2 hospital variable, model M6. The quality of recovery (QR) versus Preop PS score slope varied significantly with hospital type. A, HLM 6.08 output; B, SPSS 18.0.2 output for the same model; C, Information theoretic indices to evaluate fit.

both the micro- and the macro-level predictors are fairly similar whether calculated by HLM or OLS at the individual patient level (n = 7185). The only major difference is in standard errors for the level 2 hospital predictor. Thus, both MLM and OLS approaches would find that QR score varied inversely with age and directly with PS score and both would conclude that QR was significantly higher, on average, at large metro hospitals. However, MLM and OLS would deliver marked differences if the data were modeled at the anesthesiologist level (j = 160) for Preop PS score. Thus, OLS at the macro level would miss the important relationship between PS score and QR if the data were aggregated, a result of the ecological fallacy mentioned earlier. Furthermore, OLS analysis at the micro or macro level would not identify the cross-level interaction between hospital and the QR-PS slope (model M6) nor would OLS

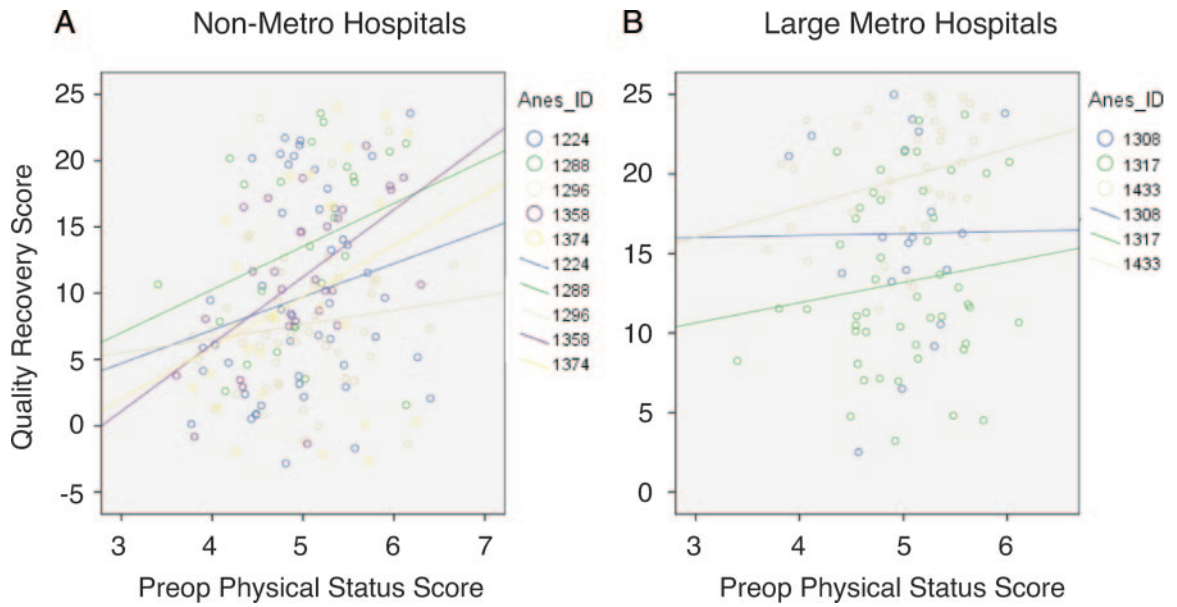

Figure 7. Plots of quality of recovery versus Preop physical status (PS) score for patients at nonmetro hospitals (A) or large metro hospitals (B). Slopes appear greater at the nonmetro hospitals.

Table 3. Summary of Models

Parameter	M1: unconditional means	M2: M1 + fixed level 1 predictors	M3: M2 + fixed level 2 predictors	M4: M3 + random level 1 slopes	M5: M3 + random slope for PS	M6: M5 + cross-level interaction
Intercept (B_{0j})	12.64*	12.64*	11.39*	11.11*	11.11*	11.40*
Age slope (B_{1j})		−0.033*	−0.033*	−0.032*	−0.032*	−0.033*
PS slope (B_{1j})		2.20*	2.20*	2.22*	2.22*	2.81*
Hospital (G_{11})			2.80*	3.46*	3.47*	2.80*
PS × hospital						−1.34*
Level 1 residual (R_{ij})	39.15	36.92	36.93	36.55	36.63	36.61
Variance component for intercept (U_0)	8.61*	8.69*	6.76*	6.87*	6.87*	6.76*
Variance component for age slope (U_1)				0.00098		
Variance component for PS slope (U_1)				0.675*	0.68*	0.30
AIC	47120	46718	46682	46670	46666	46637
BIC	47134	46732	46696	46718	46694	46665

PS = physical status; AIC = Akaike Information Criterion; BIC = Bayesian Information Criterion.
* $P < 0.05$.

Table 4. Comparison of Results Obtained by Multilevel Modeling or by Ordinary Least Squares Regression at the Micro or Macro Level

	MLM (n = 7185)			OLS: micro level, patients (n = 7185)			OLS: macro level, anesthesiologists (n = 160)		
	B	SE	P value	B	SE	P value	B	SE	P value
Intercept	11.390	0.293	<0.001	2.056	0.756	0.007	691.113	3983.058	0.862
Age	−0.033	0.008	<0.001	−0.030	0.009	<0.001	0.166	0.161	0.304
Preop PS	2.197	0.109	<0.001	2.196	0.117	<0.001	−137.940	797.826	0.863
Hospital	2.802	0.440	<0.001	2.803	0.155	<0.001	2.826	0.447	0.000

MLM = multilevel modeling; OLS = ordinary least squares regression; PS = physical status.

reveal that the QR-PS slope varied with anesthesiologist (models M4 and M5). Thus, the advantage of MLM was to incorporate and accommodate the effect of clustering, a benefit of HLM noted earlier in this article. If our data had included repeated measures, the MLM would also allow incorporation of a working correlation matrix (autoregressive or compound symmetry) to nest time within the patient, who then may be nested within the anesthesiologist.

Summary

The above primer only begins to identify the breadth of models that MLM is equipped to test. A host of nonlinear

models can be tested via a hierarchical generalized modeling approach including models with binary outcomes (using a logit link function), counts (Poisson model with log link), or ordinal outcomes. Also, models with >2 levels can be tested, such as a 3-level model in which students are nested within classes nested within schools. Models with >2 levels can still accommodate predictors at each level. Another variation on the MLM approach is in longitudinal studies[18,19] whereby time is nested within the individual. MLM analysis of longitudinal studies provides much flexibility in how to model time, as opposed to the fixed nature (i.e., wave 1, wave 2, etc.) inherent in many software packages. The interested reader could pursue the topic of nonlinear modeling with a number of good reviews,[24–27] including a primer on use of MLM in longitudinal studies.[28] Another guide to MLM with illustrations from the same database we analyzed, the NELS:88 study, can be found in a recent review by Peugh.[29]

DISCLOSURES

Name: Dale Glaser, PhD.
Contribution: This author helped design the study, conduct the study, analyze the data, and write the manuscript.
Attestation: Dale Glaser approved the final manuscript.
Name: Randolph H. Hastings, MD, PhD.
Contribution: This author helped design the study and write the manuscript.
Attestation: Randolph H. Hastings approved the final manuscript.

ACKNOWLEDGMENTS

Randolph H. Hastings statement on FAER funding: I applied for a Parker B. Francis/FAER Young Investigator Award in 1991 to continue studies on the mechanisms of alveolar protein clearance that I'd begun with my mentor, Michael Matthay. It was a wonderful surprise and a honor when I was selected for one of the grants. Looking back, the FAER support was instrumental in initiating my professional academic career at a number of levels. The favorable review and response boosted my self-confidence, letting me know that other scientists thought my career was promising and my project worthwhile. It also strengthened my stature as a physician-scientist within the department, justifying the nonclinical time I'd received. The research support gave me independence and was particularly generous, allowing me to complete the project and set-up the basic equipment I needed in my lab for future work. I published 3 papers, solved the question I had posed for the grant and set myself up with data and a publication record to seek larger scale funding. The basic research finding was that protein in pulmonary edema funding was cleared by passive diffusion out of the air spaces and not by active transcellular endocytosis as had been suspected before my work. I have subsequently received Veterans Affairs Merit Awards, National Institutes of Health funding and numerous foundation grants. My research direction has broadened and shifted considerably, but much of my success I owe to the start I received through FAER.

REFERENCES

1. Raudenbush S, Bryk AS. Hierarchical Linear Models. 2nd ed. Thousand Oaks, CA: SAGE Publications, 2002
2. Kenny DA, Judd CM. A general procedure for the estimation of interdependence. Psychol Bull 1996;119:138–48
3. Kenny DA, Mannetti L, Pierro A, Livi S, Kashy DA. The statistical analysis of data from small groups. J Pers Soc Psychol 2002;83:126–37
4. Hox J. Multilevel Analysis. Mahwah, NJ: Lawrence Erlbaum Associates, 2002
5. Kozlowski SW, Klein KJ. A Multilevel Approach to Theory and Research in Organizations. San Francisco: Jossey-Bass, 2000
6. Heck RH, Thomas SL. An Introduction to Multilevel Modeling Techniques. Mahwah, NJ: Lawrence Erlbaum Associates, 2000
7. Raudenbush SW, Bryk AS, Cheong YF, Congdon RT Jr. HLM6: Hierarchical Linear and Nonlinear Models. Lincolnwood, IL: Scientific Software International, 2004
8. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77
9. Enders CK. Applied Missing Data Analysis. New York: Guilford Press, 2010
10. Little RJ, Rubin DB. Statistical Analysis with Missing Data. New York: Wiley, 1987
11. Wong W, Kedarisetty S, Delson N, Glaser D, Moitoza J, Davis DP, Hastings RH. The effect of cross-training with adjustable airway model anatomies on laryngoscopy skill transfer. Anesth Analg 2011;113:862–8
12. Alharbi O, Rabeneck L, Paszat LF, Wijeysundera DN, Sutradhar R, Yun L, Vinden CM, Tinmouth J. A population-based analysis of outpatient colonoscopy in adults assisted by an anesthesiologist. Anesthesiology 2009;111:734–40
13. Stuttmann R, Jakubetz J, Schultz K, Schafer C, Langer S, Ullmann U, Hilbert P. Recovery index, attentiveness and state of memory after xenon or isoflurane anaesthesia: a randomized controlled trial. BMC Anesthesiol 2010;10:5
14. Bickel R. Multilevel Analysis for Applied Research. New York: Guilford Press, 2007
15. Enders CK. A primer on the use of modern missing-data methods in psychosomatic medicine research. Psychosom Med 2006;68:427–36
16. Cohen J, Cohen P, West SG, Aiken LS. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Mahwah, NJ: Lawrence Erlbaum Associates, 2003
17. Hofmann D, Gavin M. Centering decisions in hierarchical linear models: implications for research in organizations. J Manage 1998;24:623–41
18. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963–74
19. Cudeck R, Harring JR. Developing a Random Coefficient Model for Nonlinear Repeated Measures Data. New York: Taylor & Francis, 2010
20. Singer JD, Willet JB. Applied Longitudinal Data Analysis. Oxford: Oxford University Press, 2003
21. Snijders T, Bosker R. Multilevel Analysis. London: SAGE Publications, 1999
22. Roberts JK, Monaco JP, Stovall H, Foster V. Explained Variance in Multilevel Models. New York: Taylor & Francis, 2011
23. Akaike H. Factor analysis and AIC. Psychometrika 1987; 52:317–32
24. Rodriguez G, Goldman N. An assessment of estimation procedures for multilevel models with binary responses. J R Stat Soc A 1995;158:73–89
25. Goldstein H, Jon R. Improved approximations for multilevel models with binary responses. J R Stat Soc A 1996;159:505–13
26. Goldstein H. Nonlinear multilevel models, with an application to discrete response data. Biometrika 1991;78:45–51
27. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc 1993;88:9–25
28. Holden JE, Kelley K, Agarwal R. Analyzing change: a primer on multilevel models with applications to nephrology. Am J Nephrol 2008;28:792–801
29. Peugh JL. A practical guide to multilevel modeling. J Sch Psychol 2010;48:85–112