Tympanic temperature measurements: Are they reliable in the critically ill? A clinical study of measures of agreement*

John L. Moran, MBBS, FRACP, FJFICM, MD; John Victor Peter, MBBS, MD, DNB (Med) FRACP; Patricia J. Solomon, BSc, PhD; Bernadette Grealy, RN, RM, Intensive Care Cert, Dip App Sci-Nursing, BN; Tania Smith, RN; Wendy Ashforth, RN, BN; Megan Wake, RN, BN; Sandra L. Peake, BM BS, BSc (Hons), FJFICM, PhD; Aaron R. Peisach, MBBS, FRCA, FANZCA, FFICANZCA

Objective: **Accurate measurement of temperature is vital in the intensive care setting. A prospective trial was performed to compare the accuracy of tympanic, urinary, and axillary temperatures with that of pulmonary artery (PA) core temperature measurements.**

Design: **A total of 110 patients were enrolled in a prospective observational cohort study.**

Setting: **Multidisciplinary intensive care unit of a university teaching hospital.**

Patients: **The cohort was (mean ± SD) 65 ± 16 yrs of age, Acute Physiology and Chronic Health Evaluation (APACHE) II score was 25 ± 9, 58% of the patients were men, and 76% were mechanically ventilated. The accuracy of tympanic (averaged over both ears), axillary (averaged over both sides), and urinary temperatures was referenced (as mean difference, Δ degrees centigrade) to PA temperatures as standard in 6,703 recordings. Lin concordance correlation (p_c) and Bland–Altman 95% limits of agreement (degrees centigrade) described the relationship between paired measurements. Regression analysis (linear mixed model) assessed covariate confounding with respect to temperature modes and reliability formulated as an intraclass correlation coefficient.**

Measurements and Main Results: **Concordance of PA temperatures with tympanic, urinary, and axillary was 0.77, 0.92, and 0.83, respectively. Compared with PA temperatures, Δ (limits of agreement) were 0.36°C (−0.56°C, 1.28°C), −0.05°C (−0.69°C, 0.59°C), and 0.30°C (−0.42°C, 1.01°C) for tympanic, urinary, and axillary temperatures, respectively. Temperature measurement mode effect, estimated via regression analysis, was consistent with concordance and Δ (PA vs. urinary, $p = .98$). Patient age ($p = .03$), sedation score ($p = .0001$), and dialysis ($p = .0001$) had modest negative relations with temperature; quadratic relationships were identified with adrenaline and dobutamine. No interactions with particular temperature modes were identified ($p \geq .12$ for all comparisons) and no relationship was identified with either mean arterial pressure or APACHE II score ($p \geq .64$). The average temperature mode intraclass correlation coefficient for test–retest reliability was 0.72.**

Conclusion: **Agreement of tympanic with pulmonary temperature was inferior to that of urinary temperature, which, on overall assessment, seemed more likely to reflect PA core temperature. (Crit Care Med 2007; 35:155–164)**

KEY WORDS: critically ill; core temperature; ear-based thermometry; concordance; linear mixed model; intraclass correlation coefficient

Temperature measurement is an integral part of vital signs monitoring, along with measurement of respiratory rate, pulse rate, and blood pressure. Accurate assessment of temperature is essential in the intensive care unit (ICU), not only in therapeutic hypothermia (1) but also in other clinical situations in which alterations in temperature may indicate the presence of an infection, a systemic inflammatory response, deteriorating patient condition, or disorders of thermoregulatory function.

Peripheral and core temperature recordings are used to measure body temperatures in clinical practice. Peripheral temperature readings, measured in the outer 1.6 mm of skin or mucous membranes (2), are often considered unreliable because they are influenced by factors such as mouth breathing, temperatures of recently ingested food, and environmental temperature. Core temperature, on the other hand, is not influenced by external factors, more accurately reflects temperature of the vital organs, and is the preferred mode of measurement in the critically ill (3).

The optimal site of core temperature measurement is considered to be the pulmonary artery (PA) (4, 5), the routine use being limited by invasiveness, which restricts any ICU application to those patients requiring PA catheterization. Other modes of core temperature measurement include esophageal temperature measurement using a purpose-designed thermistor probe (6, 7) and rectal temperature measurement by means of a thermistor probe deep in the rectum (3, 8). In the last few decades, there has been considerable interest in

*See also p. 312.

From the Department of Intensive Care Medicine, The Queen Elizabeth Hospital, Woodville, South Australia, Australia (JLM, JVP, BG, TS, WA, MW, SLP, ARP); and the School of Mathematical Sciences, The University of Adelaide, Adelaide, South Australia, Australia (PJS).

The authors have not disclosed any potential conflicts of interest.

Supported, in part, by Unit Trust funds, Intensive Care Unit, The Queen Elizabeth Hospital.

DOI: 10.1097/01.CCM.0000250318.31453.CB

infrared ear-based thermometry, a method based on the principle of radiation of infrared energy by the tympanic membrane in proportion to its temperature. The speed, ease of use, and noninvasiveness has recommended its widespread application. Although commonly referred to as tympanic thermometry, it is quite different from direct tympanic thermometry (7, 9), a "good index of core temperature" (10), in which measurements are made using direct contact on the tympanic membrane by an electronic probe (10). A number of studies (4, 5, 11–20) have compared the various modes of thermometry (including axillary readings) with inconsistent results; the "best" method still being considered a "continuing question" (10). The majority of these studies have focused on the agreement (or lack thereof) between PA (core) temperature measurements and other modes using the method of differences, as recommended by Bland and Altman (21, 22). Three studies have further addressed the question of reliability (reproducibility) of different measurements (4, 13, 19); one study extended the observation time to 32 hrs (19), and three (4, 13, 14) also assessed confounding of temperature measurement by potential covariates. The patient subject number in these studies varied from 13 (20) to 128 (17) and the observer number from 2 (13) to 153 (11). Thus, any assessment of the continuing question of temperature measurement must address study and implicit patient heterogeneity.

There has recently been a renaissance of modeling approaches to the question of method comparison data, as opposed to more simple graphical and descriptive methods. These modeling approaches (23–26), adopted by Giuliano et al. (4), have used variance component analysis (27) to deal with the dual questions of reliability and agreement and to formally integrate covariate confounding into the (regression) analysis.

The current study, using a large number of both patients (n = 110) and on-duty clinical nurse observers (n = 90), with patient temperature recordings extending up to 5 days beyond ICU admission, explored the performance of four patient temperature measurement modes (pulmonary, ear-based tympanic, urinary, and axillary) to address the following questions: 1) what was the agreement between and repeatability of each method, using pulmonary temperatures as the "gold standard," 2) which covariates modified the performance of temperature measurements, and 3) what inferences were afforded by the different methods of analysis? In addressing these questions, we also sought to situate the current popular ear-based tympanic mode of temperature measurement and to reflect on untoward clinical consequences of potential bias in temperature measurement.

METHODS

Patients. All patients admitted to the ICU of a tertiary referral, university-affiliated hospital in Australia during a 7-month period were eligible for inclusion into the trial. Exclusion criteria were patients of <18 yrs of age, patients not willing to participate in the trial, and patients for whom insertion or reinsertion of a urinary catheter was not clinically indicated. This study was approved by the hospital's Ethics of Human Research Committee, and informed consent was obtained from the patient or the patient's closest relative.

Temperature Measurements. Bilateral tympanic and axillary temperatures were measured and recorded by nursing staff every 4 hrs for the first 72 hrs and then every 6 hrs for an additional 48 hrs. Temperature-sensing urinary or PA catheters were inserted when clinically indicated. Bilateral axillary temperatures were measured concurrently using glass mercury thermometers (Livingstone AS2190-1978 C), placed at the specified times and a reading taken after 5 mins. Ear-based temperatures were measured at the same time using Sherwood Medical First Temp (Nippon Sherwood Medical Industries, Tokyo, Japan) in both ears (degrees centigrade), using the "core" mode. The ear-tug method was used to straighten the external auditory canal by pulling the pinna in an upward and backward direction. A nursing in-service education was undertaken before commencement of the study to ensure uniformity in technique. For the purposes of further description in this study, "tympanic" temperature refers to infrared ear-based thermometry, unless otherwise specified. Urinary bladder temperatures were measured with a thermistor Foley catheter (Bard temperature-sensing urinary catheter, Bard Medical, Convington, GA) that was connected to the Spacelab monitor (Issaquah, WA) for continuous temperature measurements. Baxter PA catheters (Baxter Healthcare Corporation, Irvine, CA) were inserted when clinically indicated (for example, in patients with shock, requiring close monitoring of inotrope effect, or in patients admitted from surgical procedures with *in situ* PA catheters), their positions checked by chest radiography, and continuous temperature measurements were recorded. Temperatures at various time periods were recorded from single measurements. Ambient temperatures were measured twice daily (Digi-thermo; Actrol Ply Ltd, Blackburn, Victoria, Australia) to ensure that an ambient temperature in the ICU was maintained between 21°C and 22°C.

Demographic particulars were collected from all patients. In addition, information on ventilation, hemodynamics, and inotropic and sedative/analgesic agents were also recorded; level of sedation was scored: 0 (awake, no sedation), 1 (mild sedation, occasionally drowsy, easy to arouse), 2 (moderate sedation, often drowsy, easy to arouse), and 3 (severe sedation, somnolent, difficult to arouse), after Macintyre and Ready (28), and considered as an ordinal variable for analysis.

Statistical Analysis. 1. For pooled analysis of paired measurements, patient data were aggregated for each temperature measurement mode. The accuracy of the various temperature modes (tympanic [average of both ears], urine, and axillary [average of both sides]) was determined by the agreement (or lack thereof) with the "reference" PA temperature. Indices of agreement were also calculated for all other temperature mode combinations.

A. Using the method of differences, the Bland–Altman approach (21, 22), which is a data-scale assessment of agreement (29) with the "underlying" model formulated as a two-way analysis of variance (24), the following were calculated/generated:

i. Mean difference, standard (that is, PA) vs. test (equivalent to "fixed bias" or "offset") and 95% limits of agreement.

ii. The Bradley–Blackwood (30) omnibus test for mean values (bias) and variances (precision); nonsignificance implied concordance.

iii. Graphical display of difference (d_i) vs. mean, standard vs. test ($\bar{y}_i$), *not* the difference against standard method (22). The plot was used to inspect whether d_i and its variance was constant as a function of the average ($\bar{y}_i$), via the correlation of the difference vs. the average (equivalent to "proportional bias"); a value near zero implied concordance. In the presence of substantive correlation, temperature was log transformed, as recommended (21).

iv. As pointed out by Ludbrook (31, 32), fixed bias may be confounded by proportional bias using the method of differences, and separate agreement assessments were therefore generated using Deming regression, which assumes, for linear regression of method y vs. method x, measurement errors for both methods, as opposed to the dependent variable only for ordinary least-square regression (33–35). Deming regression was formulated as: $y = \alpha + \beta x$, with fixed bias indicated by the regression intercept (test of $\alpha = 0$) and proportional bias indicated by the regression slope parameter (test of $\beta = 1$).

B. Lin concordance correlation coefficient (CCC = ρ_c) (36, 37) (a parametric relationship-scale approach) was also calculated (29) for each comparison. The Pearson correlation coefficient (ρ) implies that data from two variables (y and x) with perfect correlation ($r = 1$) lies on a straight line, which, however, may not pass through the origin or have a slope

equal to unity. The CCC compares agreement between two sets of measurement by assessing the variation from the 45-degree line through the origin, the line of perfect concordance ($\rho_c = 1 -$ [expected squared perpendicular deviation from 45-degree line/expected squared perpendicular deviation from 45-degree line when y and x are uncorrelated] (25)). Thus, ρ_c may be considered as a product of precision (ρ) and a bias-correction factor, C_b, a measure of accuracy, and $-1 \leq \rho_c \leq 1$.

i. Muller and Buttner (38) and Dunn (25) consider that the CCC has "similarities" to the intraclass correlation coefficient (ICC) and is analogous to Cohen kappa statistic (for assessment of agreement between categorical ratings). The ICC, a measure of observer reliability (ranging from 0, no agreement, to +1, perfect agreement) is defined as the ratio of the variance of interest (there are various ICCs) to the total variance (39, 40). The simplest reliability (R) design derives from the one-way random effect analysis of variance (one-way nested variance components model) (26), $Y = u + s + e$ (measurements Y, "bias" u, subject effect s and [independent] random measurement error e); and

$$R = \frac{\hat{\sigma}_s^2}{\hat{\sigma}_s^2 + \hat{\sigma}_e^2},$$

where $\hat{\sigma}_s^2$ and $\hat{\sigma}_e^s$ are the corresponding variance estimates, whose sum is the variance of Y. "Optimal" levels of reliability have been suggested for an ICC of 0.7–0.75 (41, 42). With respect to the current context of temperature measurement, the ICC may be thought to reflect the "'closeness' of observations on the same subject relative to the closeness of observations on different subjects" (43).

ii. Similarly, Carrasco and Jover (23) have demonstrated that the ICC and the CCC are "the same measure of agreement estimated in two ways: by the variance components procedure and by the moment method," with observers as fixed effects. The propriety of the ICC in agreement studies has been questioned (44), but there is no consensus on this question (45).

C. The Lin and Bland–Altman approaches are complementary, indicating agreement (or lack of) on two different scales (29), and *both* assume (bivariate) normality of the data. Although described only 3 yrs after the approach of Bland–Altman, Lin CCC is "not very customary," but it has found recent endorsement (39, 46) and full implementation in general statistical software (29).

i. Normality of the distributions of the individual temperature modes was assessed graphically using kernel density plots (47). The kernel density plot is a modification of the histogram (a "smoothed" histogram), where densities are the continuous analogues of proportions (derivatives of the cumulative distribution function, so that areas under the density function read off as probabilities). The data are divided into intervals (which may overlap), and estimates of the density at the interval centres are produced; the "kernel" is the function (a number are available) that weights the observations by the distance from the center of the interval.

2. Regression modeling using variance components of a linear mixed model (48). The model incorporated the different temperature measurement modes and potential modifying covariates as fixed effects, which "capture the influence of explanatory variables on the mean structure, exactly as in the standard linear model" (26). Patients, temperature measurement methods, and (clinical nurse) observers also entered as nested random effects (or levels), and random coefficients for time were allowed at the patient level (that is, individual patient "time slopes"); estimates of random-effects variables are expressed as standard deviations (that is, square-root of the variance). For each measurement mode, temperature over time was visualized graphically using a nonparametric multivariable scatterplot smoother (49). Potential heteroscedasticity, increase in variance with increase in temperature, was explored and appropriate compensation was made where indicated. Normality of residuals was assessed graphically. Multicollinearity was assessed using the variance inflation and condition number indices; where variance inflation of <10 and condition number less than "30 or more" are desirable (50). Competing models were adjudged by likelihood ratio tests and information criteria (51).

A. ICCs for various pairs of responses were calculated from the particular random effect variance components (patient, method, observers, and residual measurement error) of the linear mixed model (40, 52), in particular: between measurements for the same method on the same subject, between measurements for the same method on the same subject by the same observer, for measurements with different methods on the same subject, for measurements with different methods on the same subject by the same observer, and test–retest reliability of each method.

3. Stata (version 9.1 SE, 2005, StataCorp, College Station, TX) and S-PLUS (version 7, 2005, Insightful Corporation, Seattle, WA) statistical software was used.

RESULTS

During the study period, 421 patients were admitted to the ICU; 110 patients with temperature-sensing urinary catheters (n = 92) or PA catheters (n = 41) were enrolled. Insertion of temperature-sensing urinary catheters was not clinically indicated in 305 patients who already had urinary catheters in place at admission to the ICU or did not require one, three patients were unwilling to participate, and data were incomplete for three patients. The mean (SD) age of the study cohort was 64.6 (15.9) yrs, with mean (SD) Acute Physiology and Chronic Health Evaluation (APACHE) II score of 25.1 (8.7). Baseline characteristics of the patients are summarized in Table 1. A total of 2,165 tympanic, 2,118 axillary, 1,761 urinary, and 659 PA temperature measurements were recorded. Overall data are summarized in Table 2, and assessments of normality of temperature mode distributions via kernel density plots are seen in Figure 1. All temperature distributions were approximately normally distributed with a degree of (left) skewness and kurtosis.

Pooled Analysis. The relationship between PA and other modes of tempera-

Table 1. Baseline characteristics

Patients, n	110
Age in yrs, mean (SD)	64.6 (15.9)
Male sex, %	58.2
APACHE II, mean (SD)	25.07 (8.7)
SAPS II, mean (SD)	43.4 (19.2)
Number (%) of patients mechanically ventilated	75.5 (71)
Concurrent dialysis, %	4.4
Sedation score[a]	2 (0–3)
Narcotics,[b] %	52.6
Inotropes,[c] %	49
Adrenaline,[a] μg/min	9 (0.5–80)
Noradrenaline,[a] μg/min	10 (1–80)
Dopamine,[a] μg/min	5 (1–123)
Dobutamine,[a] μg/min	7.5 (2–15)
Patients with ICU diagnosis, n	
Cardiac	
Acute MI/cardiogenic shock/cardiac failure	7
Cardiac arrest	6
Respiratory	
Pneumonia/ARDS	11
COPD exacerbation	10
Others	5
Sepsis/septic shock	15
Postoperative	
Gastrointestinal surgery	14
Vascular	14
Drug overdose	6
Trauma	3
GIT hemorrhage	3
Intracerebral hemorrhage	3
Miscellaneous	13

APACHE II, Acute Physiology and Chronic Health Evaluation II score; SAPS II, Simplified Acute Physiology Score II; ICU, intensive care unit; MI, myocardial infarctions; ARDS, acute respiratory distress syndrome; COPD, chronic obstructive pulmonary disease; GIT, gastrointestinal tract.

[a]Data provided as median (range); [b]percentage of patients receiving parenteral narcotic during observation time; [c]percentage of patients receiving inotropic agents during observation time.

ture measurement was formally assessed across the entire temperature range (Table 3) and graphically displayed in the Bland–Altman panel plots of Figure 2, all displayed on the same scale. Tympanic temperatures showed only modest concordance with PA ($p_c = 0.77$), urinary ($p_c = 0.69$), and axillary ($p_c = 0.76$) temperatures. Although the average difference of PA vs. urinary temperature was small at −0.05, with a good concordance ($p_c = 0.92$), there was a modest correlation between the average difference and the mean, reflected in a significant fixed and proportional bias via Deming regression. PA-axillary comparisons performed surprisingly well across the agreement indices, with an offset that was comparable with that of PA-tympanic. The modest degree of correlation, difference vs. mean (Table 2, column "Correlation AVD Mean"), displayed by the (*dash-dot*) regression line in Figure 2 panels of PA vs. urinary, urinary vs. axillary, and tympanic vs. axillary temperature modes, was indicative of the proportional bias also identified by Deming regression. These relationships were not materially influenced by log-transformation of temperature, except for the proportional bias in the urinary-axillary comparison (Table 3). For each of the six comparisons, the hypothesis of joint equality of mean values and variances (Bradley–Blackwood test (30)) was rejected.

Linear Mixed Model. The results of the regression modeling of temperature modes for the full data set are seen in Table 4. Two models (1 and 2) are presented, with PA temperature mode as the comparator: model 1 was covariate unadjusted (except for time), and model 2 was covariate adjusted; there was statistical advantage for the covariate-adjusted model (likelihood ratio test, $p < .0001$). For both models, a random coefficient for time at the patient level (individual patient time slopes) had statistical advantage ($p < .0001$), although the fixed effect for time was nonsignificant ($p \geq .86$), as reflected in Figure 3, in which overall (mean) temperature profiles for each temperature mode were seen to be relatively stable. No disquieting heteroscedasticity was identified, and there was no multicollinearity (variance inflation, 1.1; condition number, 6.2). The fixed-effect comparisons between the temperature modes were, perhaps not surprisingly, consistent with those of the pooled comparisons (Table 3), both in magnitude and direction of difference. For both models, no statistical difference was demonstrated between the coefficient of effect for urinary temperature measurement and PA. The covariate associations were minimal (age) to modest (dialysis) in magnitude, with a negative relationship of age, sedation, and dialysis with temperature. The sedation "effect" (comparison with no sedation) seemed to be linear in its increments. Minimal to mild quadratic temperature associations were evident with adrenaline (dose ranges up to 80 μg/min) and dobutamine

Table 2. Overall summary of data

Temperature Mode	Observations, n	Mean (SD) Temp, °C	Temp Range, °C
Tympanic, right	2165	36.88 (0.96)	27.8–39.5
Tympanic, left	2160	36.93 (0.98)	
Axillary, right	2118	37.03 (0.79)	34.8–39.7
Axillary, left	2118	37.04 (0.79)	
Urinary	1761	37.32 (0.96)	26.8–40.4
Pulmonary	659	37.42 (0.82)	34.3–39.3

Temp, temperature.

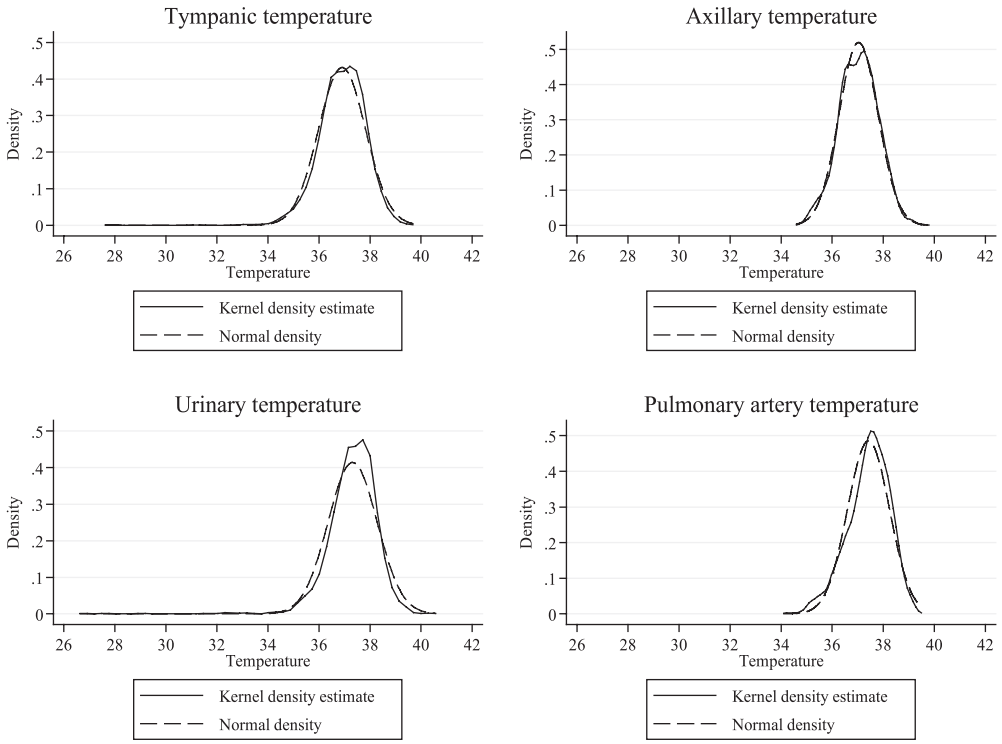

Figure 1. Kernel density estimates of temperatures. Density (*solid line*) is shown on the *y-axis* over temperature ranges on the *x-axis*. Superimposed normal density plot (*dashed line*).

Table 3. Comparison of different pooled temperature measurements

Assessments Temperature Modes	n	AVD (SD)	LOA	Correlation AVD Mean	B-B test p	CCC Precision Accuracy p_c	p	C_b	Deming Regression Fixed Bias (α)	Proportional Bias (β)
Pulmonary–tympanic	648	0.358 (0.469)	−0.560, 1.276	−0.072	.0001	.77	.841	0.914	1.777 (−0.114, 3.668)	0.962 (0.911, 1.013)
Pulmonary–urinary	355	−0.052 (0.327)	−0.694, 0.589	−0.20 (−0.204)[c]	.0001	.92	.923	0.995	2.752 (0.650, 4.854)[a,b]	0.925 (0.869, 0.981)[a,b]
Pulmonary–axillary	634	0.295 (0.367)	−0.424, 1.014	−0.008	.0001	.83	.889	0.933	0.433 (−0.876, 1.742)	0.96 (0.961, 1.031)
Urinary–tympanic	1735	0.447 (0.659)	−0.845, 1.739	0.031	.0001	.686	.761	0.901	−0.309 (−3.802, 3.185)	1.020 (0.926, 1.115)
Urinary–axillary	1701	0.322 (0.555)	−0.765, 1.409	0.185 (0.196) [c]	.0001	.712	.773	0.921	−4.348 (−8.848, 0.151)	1.126 (1.005, 1.247)[a]
Tympanic–axillary	2089	−0.097 (0.552)	−1.178, 0.984	0.154 (0.162) [c]	.0001	.761	.770	0.988	−3.972 (−4.008, −1.126)[a,b]	1.105 (1.064, 1.145)[a,b]

n, number of observations; AVD, average temperature difference; LOA, Bland--Altman 95% limits of agreement; B-B test, Bradley-Blackwood omnibus test of equality of mean values and variance; CCC p_c, Lin concordance correlation; precision p, Pearson correlation coefficient; accuracy C_b, bias correction ($0 < C_b \leq 1$) factor measuring deviation of best-fit line from line of identity (45 degrees); Deming regression line, $y = \alpha + \beta x$; fixed bias (test of $\alpha = 0$), regression intercept with 95% confidence interval; proportional bias (test of $\beta = 1$), regression slope parameter with 95% confidence interval.

[a]$p < .05$; [b]$p < .05$ for log-transformed variables; [c]correlation AVD mean for log-transformed variables.

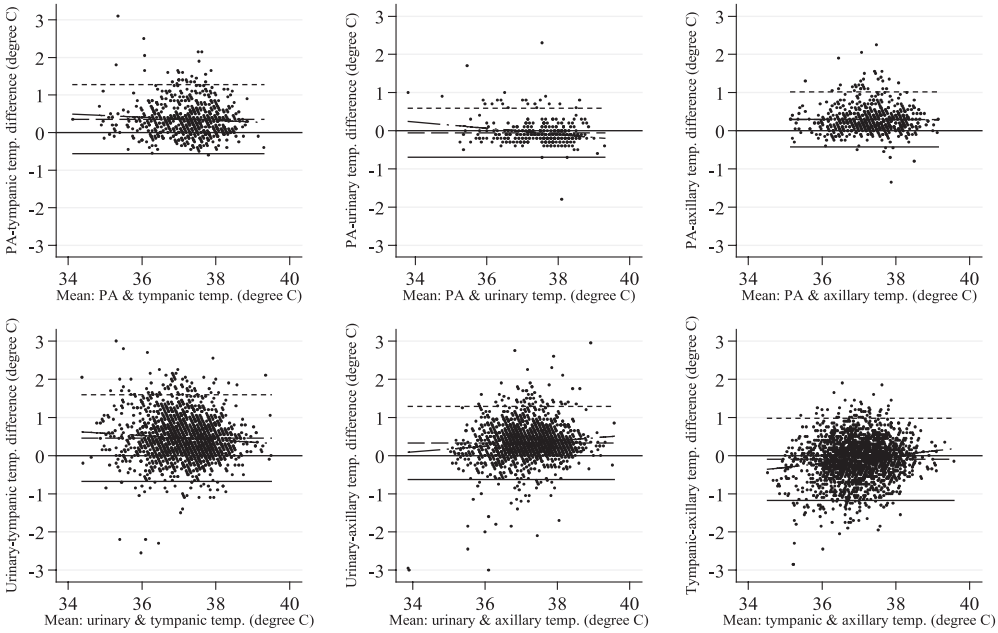

Figure 2. Bland–Altman plots of temperature mode differences vs. averages. The six panels for different combinations of temperature modes plot the temperature difference against the mean of the two particular temperatures. The 95% limits of agreement are indicated by the *upper short-dash line* and the *lower solid line*. The offset or bias between the two temperature modes is indicated by the *long-dash line*. The regression line (difference vs. mean) is indicated by the *long-dash–dot line*. *PA*, pulmonary artery; *temp*, temperature.

(dose ranges up to 15 μg/min) (likelihood ratio test, $p = .0001$) but not with mean arterial pressure (likelihood ratio test, $p = .16$). No significant interactions were demonstrated between the individual temperature measurement modes and 1) time in which the temperatures were recorded ($p \geq .65$), 2) inotropic agents (adrenaline, noradrenaline, dobutamine, and dopamine; $p \geq .30$ for all comparisons), 3) mean arterial pressure ($p \geq .65$ for all comparisons), 4) mechanical ventilation ($p \geq .43$ for all comparisons), and 5) the covariates age, sedation score, and dialysis ($p \geq .12$, $\geq .13$, and $\geq .16$, respectively, for all comparisons). There was no demonstrable effect of severity of illness as measured by the admission APACHE II score ($p = .64$). ICC calculations are given in Table 5 for both models; test–retest reliability of the methods was acceptable.

DISCUSSION

The current investigation compared four modes of temperature measurement in the critically ill, using PA temperature as the comparator, and contrasted two approaches to the assessment of agreement between these different measurement modes. The principal findings were the lack of difference between PA and urinary temperatures compared with PA and tympanic, the significant effect of covariates on temperature, and the relatively modest ICCs for response pairs.

Studies of Temperature Measurement Modes. As noted in the introduction to this article, a number of studies have addressed the comparison of temperature measurement modes, with a variable fo-

Table 4. Variance components regression modelling: Covariate unadjusted and adjusted

Model	1 Covariate Unadjusted	95% CI (Lower, Upper)	p	2 Covariate Adjusted	95% CI (Lower, Upper)	p
n	6,399			6,298		
Correlation structure	Unstructured			Unstructured		
Parameter/variable	Estimate			Estimate		
Fixed effects: PA comparator						
Tympanic temperature	−0.403	−0.497, −0.309	.0001	−0.416	−0.512, −0.321	.0001
Urinary temperature	0.012	−0.087, 0.112	.810	0.001	−0.100, 0.102	.981
Axillary temperature	−0.288	−0.382, −0.194	.0001	−0.299	−0.396, −0.204	.0001
Alternate comparisons						
Urinary vs. tympanic	−0.415	−0.482, −0.348	.0001	−0.418	−0.486, −0.349	.0001
Urinary vs. axillary	−0.301	−0.368, −0.223	.0001	−0.301	−0.369, −0.232	.0001
Tympanic vs. axillary	0.114	0.052, 0.177	.0001	0.117	0.053, 0.182	.0001
Covariates						
Age				−0.008	−0.015, −0.0006	.035
Time	0.006	−0.004, 0.005	.798	0.002	−0.003, 0.006	.439
Sedation score 1				−0.073	−0.137, −0.008	.028
Sedation score 2				−0.112	−0.193, −0.031	.007
Sedation score 3				−0.205	−0.294, −0.117	.0001
Adrenaline				0.022	0.015, 0.029	.0001
Adrenaline square				−0.0005	−0.0006, −0.0003	.0001
Dobutamine				−0.098	−0.164, −0.033	.003
Dobutamine square				0.006	0.005, 0.011	.033
Dialysis				−0.660	−0.819, −0.501	.0001
Constant	37.3	37.1, 37.5	.0001	37.4	37.2, 37.5	.0001
Random effects						
Patient						
SD (time)	0.020	0.017, 0.027		0.020	0.016, 0.025	
SD (intercept)	0.805	0.694, 0.933		0.765	0.659, 0.889	
Method						
SD (intercept)	0.137	0.098, 0.190		0.146	0.108, 0.197	
Observer						
SD (intercept)	0.391	0.370, 0.414		0.383	0.362, 0.406	
Residual error	0.492	0.481, 0.505		0.491	0.480, 0.504	

Unstructured, unstructured error correlation structure; estimate, parameter point estimate; 95% CI, lower and upper 95% confidence interval; PA, pulmonary artery temperature; age, age (centered) in years; SD, standard deviation of the random effects for each nested level; SD (time), standard deviation of the random patient-time slopes; SD (intercept), standard deviation (square-root of the variance) of the intercept parameters for patient, method, and observer random effects, which are formulated from a normal distribution with mean centered at zero; residual, residual error; sedation score, levels 1, 2, and 3 compared with 0 (see "METHODS" section).

For models with same n = 6,298, likelihood ratio test favored model 2 vs. model 1. $p < .0001$. The scale of the parameters was degrees centigrade. The inotrope effect (adrenaline and dobutamine) was per unit (μg) increase in dose.

cus on tympanic measurements. In one of the earliest, in 15 critically ill patients, Nierman (19) observed that tympanic measurements did not give consistent and reliable bedside measurements of core body temperature. Compared with PA readings, tympanic measurements showed a mean (SD) difference of −0.38°C (0.42°C) and −0.04°C (0.27°C) for urinary measurements. A similar degree of mean difference, −0.42°C between PA and tympanic temperatures, was found by Klein et al. (17) in 128 adult surgical intensive care patients, although the authors suggested on the basis of this that tympanic temperatures were an "appropriate substitute" for PA recordings. The study of Erickson and Kirklin (13) was somewhat at variance with these results. In a convenience sampling of 38 ICU patients, they observed that tympanic temperatures had overall mean (SD) differences of only 0.07°C (0.41°C) and 0.03°C (0.23°C) with pulmonary and urinary temperatures. However, readings were taken during a 4-hr period with stable temperatures in 42% of patients and temperature changes of ≤0.1°C in the remaining patients. It is conceivable that the absence of marked fluctuations in temperatures may have influenced their results. Similar minimal degrees of bias (−0.06°C to −0.13°C), PA vs. tympanic, were also found in the pediatric experience of Romano et al (53). However, a study on cardiac surgery patients showed that rapid changes in core temperatures (as measured by PA catheter) resulted in a lagging behind of other temperatures (notably rectal and tympanic); only esophageal measurements in this context closely approximated changes in pulmonary temperatures (54). Urinary temperatures were not measured in this study. Giuliano et al. (4), comparing a single set of 102 pulmonary with 102 tympanic and oral temperature measurements in 102 ICU patients, observed mean (SD) differences between PA and tympanic temperatures of 0.11°C (0.57°C), although tympanic measurements were associated with the greatest variability. Serial measurements of temperature were not undertaken, and the cohort did not include hypothermic and hyperthermic subjects. In a subsequent study by the same group (15), up to three sets of data (812 measurements) were collected from 72 subjects. The primary objective of their study was to demonstrate that oral measurements had less variability than tympanic measurements when compared with measurements obtained via a pulmonary catheter. Although only 47 of 203 data points were outside the tolerance region for oral cath-

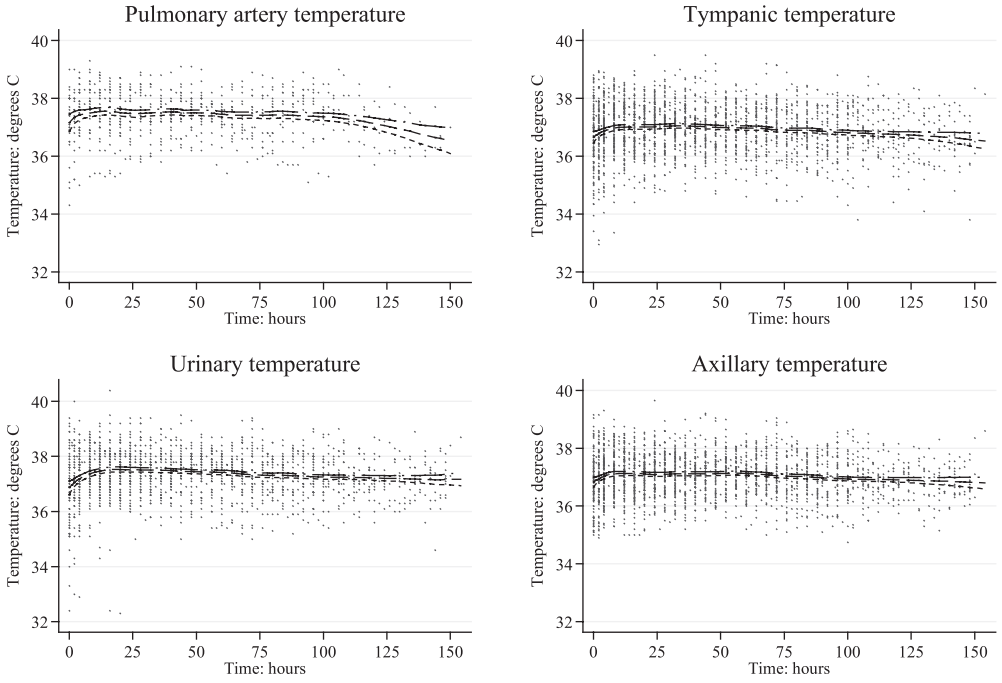

Figure 3. Temperature change over time for each measurement mode. Scatterplot panels of each temperature mode over time in hours. The mean response (*long-dash line*) with 95% confidence interval (*long-dash–dot* and *short-dash lines*) is determined by a nonparametric running line smoother.

Table 5. Intraclass correlation coefficients for various models

Error Correlation Structure Intraclass Correlation Coefficient	Model 1 Covariate Unadjusted Unstructured	Model 2 Covariate Adjusted Unstructured
Between measurements for same method on same subject	0.733	0.714
Between measurements for the same method on the same subject by the same observer	0.771	0.757
For measurements with different methods on the same subject	0.712	0.689
For measurements with different methods on the same subject by the same observer	0.767	0.752
Test–retest reliability of each method	0.727	0.707

eters, 75 of 203 data points were outside the tolerance region for tympanic measurements. They also observed that the degree of variability was less for febrile patients as compared with afebrile patients. Again, the major limitation, identified by the authors, was the absence of markedly hypothermic patients.

The conclusion to be drawn from this review is the uncertain status of tympanic measurement as a valid estimate of core temperature, as reflected by PA recordings. In the current study, using the method of differences, the optimal temperature gradient was that of PA-urinary (average difference, −0.052; 95% LOA, −0.69, 0.59), which was also reflected in the CCC estimate of 0.92 and the coefficient estimate (0.018, compared with PA) from the linear mixed model (Table 4). Similar results with respect to PA-urinary temperature differences have been noted (5, 12, 16, 18). This being said, all pooled temperature comparisons rejected the joint hypothesis of equal mean values (bias) and variances (precision). Thus, inference regarding the "optimum" temperature comparison (in the current study, PA-urinary) as being necessarily applicable to other samples of subjects must also be limited, a point consistently reiterated by Dunn (25, 33, 55); similar cautions seem applicable to the results of the studies surveyed above. Furthermore, the current study was unable to address any effect of performance variation between, for example, different infrared ear-thermometry models, in which circuit algorithms presumably differ.

The relatively poor performance of tympanic measurements in this study may have reflected a number of factors. First, in alterations of regional blood flow accompanying critical illness, tympanic membranes may behave as an extension of the skin or the mucous membrane in the critically ill, and the peripheral vasoconstriction that occurs with inotropes and some forms of shock may occur in the tympanic membrane. Second, Amoateng-Adjepong et al. (11) reported greater variability of readings when tympanic measurements were performed by nurses in routine clinical practice. Before commencement of the current study, in-service education was undertaken, although "training" may have little impact on the performance of tympanic temperature measurement (56). Third, no specific detailed examination of the ears was undertaken during the study to rule out local factors in the ear (e.g., cerumen, perforation) that may have influenced tympanic temperature measurements.

Although the agreement indices for PA-axillary measurements compared surprisingly well with those of PA-tympanic

(Table 3), there may have been an intrinsic bias in the axillary measurements due to the limitation of time (5 mins) after which these temperatures were recorded. Erickson and Meyer recorded times of 3–13 mins for axillary temperatures (electronic thermometer) to reach maximum (54% at 8 mins and 94% at 12 mins) (14). Furthermore, there is obvious limitation of application of PA-axillary measurements to jurisdictions where glass mercury thermometers are unavailable.

Confounding Factors. In the three studies (4, 13, 14) that addressed the question of covariate effect, no significant confounding influences were demonstrated. However, only Giuliano et al. (4) used a methodology comparable with that of the current study, which also had the advantage of a considerably larger number of observations. This latter point has again been stressed by Dunn and Roberts (33), who have suggested minimum sample sizes of ≥200. The negative associations with temperature of age, sedation score, and the presence of dialysis were intuitively reasonable. The lack of effect of mechanical ventilation, with the requirement for initial sedation, was somewhat surprising, although Erickson and Meyer (14) reported confounding between oral temperatures only and mechanical ventilation. With respect to the presence of inotropic agents, these "effects" may have been surrogates for the complex interplay between patient temperature and pathophysiology and the dose-dependent α and β effects (comprising both vasoconstriction and vasodilatation and thermogenesis) of both inotropic agents (adrenaline being prescribed at much higher rates). In the absence of significant interaction of covariate effects with temperature measurement modes, the former may be general associations of temperature in the critically ill, with no implied directional causality.

Categorization of temperature measurements into different ranges has been attempted to study the performance of measurement modes on temperature range, a strategy which is equivalent to the formal assessment of proportional bias. Erickson and Kirklin (13) observed that the "accuracy" of each of the methods (tympanic, bladder, oral and axillary) "varied with the level[s] . . . [seven] . . . of pulmonary artery temperature," with axillary temperatures being "highly variable." In a subsequent study (14), with temperatures categorized into four ranges, similar differences were found, but the PA-tympanic temperature difference was significant with only one of the tested instruments. Although seemingly reasonable from a clinical point of view (20), such a strategy is problematic: 1) shorter analytic ranges are known to lead to reduced values of any correlation coefficient (44), the magnitude of correlation coefficients being dependent on the extent of the analytic range, imprecision, and inaccuracy (systematic bias) (46, 57); 2) in general, cut-point analyses are associated with an increase in type 1 error, overestimation of effect at each of the cut-point levels, and the conceptual problem of sudden marked changes in effect at the various levels (58); and 3) as pointed out by Bland and Altman, plotting the difference (or bias) against the standard (the procedure adopted in Erickson and Meyer (14) and Erickson and Kirklin (13), above), rather than the average of test and standard, will "show a relation, whether there is a true association between difference and magnitude or not" (22). As opposed to these studies, we do not report performance indices based on categorization of temperatures.

Methodologic Concerns. Any statistical analysis must address the three components of agreement: the degree of linear relationship between (two) measurements, differences in mean values (location shift) and in variances (scale shift); the "more an individual measure addresses these three components, the better it evaluates agreement" (46). The comprehensive analytic approach of the current study was similar to that of recent recommendations (59, 60). We sought to extend graphical and summary measure techniques, " . . . often the end-point of methods for analysis in method comparison" (33), with complementary approaches: a focus on both fixed and proportional bias, the use of the CCC and Deming regression, and the extension to regression modeling to formally incorporate covariate confounding. This being said, the basis for analysis via Bland–Altman plots, the CCC and Deming regression, that of "pooling" ("Statistical Analysis," *1*, above), was problematic in that individual patient measurements were not independent. The linear mixed model, by definition, accommodated such dependency and was the apposite approach, although the inferences obtained from the "simple" approaches were generally consonant with those of the linear mixed model.

Inconsistencies may be expected to attend this multiplicity of testing, in particular, that Deming regression suggested somewhat better performance of the PA-tympanic comparison than either the method of differences or the CCC and the uniform rejection of the joint hypothesis of equal mean values and variances by the Bradley–Blackwood test. Similarly, the degree of proportional bias identified in a number of comparisons in Table 3 was not reflected by the lack of heteroscedasticity in the regression analysis.

That the magnitude of the ICCs (Table 5) tended to approximate the lower range of "optimal" levels also deserves comment. First, as pointed out by Bland and Altman (44), low values of the ICC may reflect low variability between subjects, not lack of agreement between methods. Second, the (mild) decrease in the ICCs with covariate adjustment reflected the decrease in variability produced by this adjustment (seen in Table 4 in the decrease of the "Patient, SD (intercept)" from 0.805 to 0.765), a finding similar to that of Carrasco and Jover (23). Third, the regression model was explicitly formulated to examine sources of variability (33), with random effects for patients, methods, and observers. To this extent, it addressed a quite different question than the comparison of two (or more) specific methods, fundamental to the Bland–Altman approach (45); rather, patients/methods/observers were treated as random samples from populations. This assumption of "randomness" was "part of the hypothesis which is being tested" (61) and did not imply "strict" random sampling (62). Fourth, no significant (fixed effect) temperature–time change was demonstrated, and an unstructured error correlation structure was employed; thus, the interpretation of the ICCs was that they represented the relative proportions of variability explained over the time period of the study. This may have been nonoptimal, and specific serial error correlation structures (for instance, the familiar autoregressive [AR1] correlation) may have been apposite. The problem with incorporating such a correlation structure is to explicate what was dependence (serial correlation) and what was a systematic effect (time as fixed effect). The inclusion of autocorrelation in the model would potentially attribute the systematic time effect (and therefore the test–retest reliability *as defined*) to dependence and, without further modification of the (mathematical) form of test–

retest reliability, bias its calculation. This analytic issue was not pursued, although it was noted that Vangeneugden et al. (26) demonstrated a decline in reliability over time with repeated measurements using linear mixed model methodology and an exponential Gaussian serial process, with a complex derivation of reliability. Similarly, formally modeling the variance across the measurement methods (by allowing the variance to differ across the methods using the "varIdent" function in the S-PLUS linear mixed model module *lme* (48)) had statistical advantage ($p < .001$), with appreciably larger variances of the tympanic and urinary vs. axillary methods (with respect to the PA, fixed at 1) in both the unadjusted (tympanic, 1.44; urinary, 1.38; axillary, 1.06) and covariate-adjusted (tympanic, 1.51; urinary, 1.44; axillary, 1.11) models (Table 3, "LOA"). Such decreased axillary temperature variability may have reflected lack of full equilibration (see "Studies of Temperature Measurement Modes," above), long-term staff familiarity, or rounding of values with a nonelectronic device.

However, the inference obtained from the ICCs was to question the reliance placed on (single) measurements of temperature over time obtained under clinical conditions in the critically ill. In this study, residual error variance (related to precision; see "Residual error" in Table 4) exceeded that of the observers (related to accuracy; see "Observer, SD (intercept)," Table 4), and thus, "lack of agreement" was mainly due to lack of (measurement) precision (23).

CONCLUSIONS

We would conclude that the place of tympanic membrane measurements as accurate reflections of core temperature in the critically ill is not established. The use of urinary catheters mandates urinary core measurements as the most reasonable alternative to PA core temperatures in critically ill patients. Linear mixed modeling of temperature profiles offers advantage in interpretation.

REFERENCES

1. Bernard SA, Buist M: Induced hypothermia in critical care medicine: A review. *Crit Care Med* 2003; 31:2041–2051
2. Togawa T: Body temperature measurement. *Clin Phys Physiol Meas* 1985; 6:83–108
3. Fulbrook P: Core temperature measurement: A comparison of rectal, axillary and pulmonary artery blood temperature. *Intensive Crit Care Nurs* 1993; 9:217–225
4. Giuliano KKR, Scott SSRBC, Elliot SRBM, et al: Temperature measurement in critically ill orally intubated adults: A comparison of pulmonary artery core, tympanic, and oral methods. *Crit Care Med* 1999; 27:2188–2193
5. Lefrant JY, Muller L, de La Coussaye JE, et al: Temperature measurement in intensive care patients: Comparison of urinary bladder, oesophageal, rectal, axillary, and inguinal methods versus pulmonary artery core method. *Intensive Care Med* 2003; 29: 414–418
6. Patel N, Smith CE, Pinchak AC, et al: Comparison of esophageal, tympanic, and forehead skin temperatures in adult patients. *J Clin Anesth* 1996; 8:462–468
7. Shiraki K, Konda N, Sagawa S: Esophageal and tympanic temperature responses to core blood temperature changes during hyperthermia. *J Appl Physiol* 1986; 61:98–102
8. Rabinowitz RP, Cookson ST, Wasserman SS, et al: Effects of anatomic site, oral stimulation, and body position on estimates of body temperature. *Arch Intern Med* 1996; 156: 777–780
9. Shiraki K, Sagawa S, Tajima F, et al: Independence of brain and tympanic temperatures in an unanesthetized human. *J Appl Physiol* 1988; 65:482–486
10. Erickson RS: The continuing question of how best to measure body temperature. *Crit Care Med* 1999; 27:2307–2310
11. Amoateng-Adjepong Y, Del MJ, Manthous CA: Accuracy of an infrared tympanic thermometer. *Chest* 1999; 115:1002–1005
12. Earp JK, Finlayson DC: Relationship between urinary bladder and pulmonary artery temperatures: A preliminary study. *Heart Lung* 1991; 20:265–270
13. Erickson RS, Kirklin SK: Comparison of ear-based, bladder, oral, and axillary methods for core temperature measurement. *Crit Care Med* 1993; 21:1528–1534
14. Erickson RS, Meyer LT: Accuracy of infrared ear thermometry and other temperature methods in adults. *Am J Crit Care* 1994; 3:40–54
15. Giuliano KK, Giuliano AJ, Scott SS, et al: Temperature measurement in critically ill adults: A comparison of tympanic and oral methods. *Am J Crit Care* 2000; 9:254–261
16. Harasawa K, Kemmotsu O, Mayumi T, et al: Comparison of tympanic, esophageal and blood temperatures during mild hypothermic cardiopulmonary bypass: A study using an infrared emission detection tympanic thermometer. *J Clin Monit* 1997; 13:19–24
17. Klein DG, Mitchell C, Petrinec A, et al: A comparison of pulmonary artery, rectal, and tympanic membrane temperature measurement in the ICU. *Heart Lung* 1993; 22: 435–441
18. Mravinac CM, Dracup K, Clochesy JM: Urinary bladder and rectal temperature monitoring during clinical hypothermia. *Nurs Res* 1989; 38:73–76
19. Nierman DM: Core temperature measurement in the intensive care unit. *Crit Care Med* 1991; 19:818–823
20. Schmitz T, Bair N, Falk M, et al: A comparison of five methods of temperature measurement in febrile intensive care patients. *Am J Crit Care* 1995; 4:286–292
21. Bland JM, Altman DG: Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 1:307–310
22. Bland JM, Altman DG: Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet* 1995; 346:1085–1087
23. Carrasco JL, Jover L: Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* 2003; 59:849–858
24. Carstensen B: Comparing and predicting between several methods of measurement. *Biostatistics* 2004; 5:399–413
25. Dunn G: Statistical Evaluation of Measurement Errors: Design and Analysis of Reliability Studies. London, Hodder Arnold, 2004
26. Vangeneugden T, Laenen A, Geys H, et al: Applying linear mixed models to estimate reliability in clinical trial data with repeated measurements. *Control Clin Trials* 2004; 25: 13–30
27. Cox DR, Solomon PJ: Components of Variance. Boca Raton, FL, Chapman and Hall/ CRC, 2003
28. Macintyre PE, Ready LB: Acute Pain Management: A Practical Guide. London, WB Saunders, 2001
29. Steichen TJ, Cox NJ: sg84: Concordance correlation coefficient. *Stata Tech Bull Reprints* 1999; Sg84:137–145
30. Blackwood LG, Bradley EL: An omnibus test for comparing two measuring devices. *J Qual Technol* 1991; 23:12–16
31. Ludbrook J: Comparing methods of measurement. *Clin Exp Pharmacol Physiol* 1997; 24:193–203
32. Ludbrook J: Statistical techniques for comparing measurers and methods of measurement: A critical review. *Clin Exp Pharmacol Physiol* 2002; 29:527–536
33. Dunn G, Roberts C: Modelling method comparison data. *Stat Methods Med Res* 1999; 8:161–179
34. Linnet K: Evaluation of regression procedures for methods comparison studies. *Clin Chem* 1993; 39:424–432
35. Marchenko Y: Deming: Stata Module. Available at: http://www.stata.com/users/jmarchenko/deming. Accessed December 2005
36. Lin LI: A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989; 45:255–268
37. Lin LI, Hedayat AS, Yang M: Statistical methods in assessing agreement: Models, issues and tools. *J Am Stat Assoc* 2002; 97:257–270
38. Muller R, Buttner P: A critical discussion of intraclass correlation coefficients. *Stat Med* 1994; 13:2465–2476
39. Schuck P: Assessing reproducibility for inter-

val data in health-related quality of life questionnaires: Which coefficient should be used? *Qual Life Res* 2004; 13:571–586
40. Shrout PE, Fleiss JL: Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* 2005; 86:420–428
41. Arts DGT, De Keizer NF, Vroom MB, et al: Reliability and accuracy of Sequential Organ Failure Assessment (SOFA) scoring. *Crit Care Med* 2005; 33:1988–1993
42. Lee J, Koh D, Ong CN: Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput Biol Med* 1989; 19:61–70
43. Rabe-Hesketh S, Skrondal A: Linear variance component models. *In:* Multilevel and Longitudinal Modeling Using Stata. Rabe-Hesketh S, Skrondal A (Eds). College Station, TX, Stata Press, 2005, pp 1–30
44. Bland JM, Altman DG: A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990; 20:337–340
45. Rankin G, Stokes M: Reliability of assessment tools in rehabilitation: An illustration of appropriate statistical analyses. *Clin Rehabil* 1998; 12:187–199
46. Sanchez MM, Binkowitz BS: Guidelines for measurement validation in clinical trial design. *J Biopharm Stat* 1999; 9:417–438
47. Wilcox RR: Kernel density estimators: An approach to understanding how groups differ. *Understanding Stat* 2004; 3:333–348
48. Pinheiro JC, Bates DM: Extending the basic linear mixed-effects model. *In:* Mixed-Effects Models in S and S-Plus. Pinheiro JC, Bates DM (Eds). Rensselaer, NY, Springer-Verlag, 2000, pp 201–270
49. Royston P, Cox NJ: A multivariable scatter plot smoother. *Stata J* 2005; 5:405–412
50. Belsley DA: Conditioning Diagnostics, Collinearity and Weak Data in Regression. New York, John Wiley and Sons, 1991
51. Kuha J: AIC and BIC: Comparisons of assumptions and performance. *Sociol Methods Res* 2005; 33:188–229
52. Rabe-Hesketh S, Skrondal A: Multilevel and Longitudinal Modeling Using Stata. College Station, TX, Stata Press, 2005
53. Romano MJ, Fortenberry JD, Autrey E, et al: Infrared tympanic thermometry in the pediatric intensive care unit. *Crit Care Med* 1993; 21:1181–1185
54. Robinson J, Charlton J, Seal R, et al: Oesophageal, rectal, axillary, tympanic and pulmonary artery temperatures during cardiac surgery. *Can J Anaesth* 1998; 45:317–323
55. Dunn G: Design and analysis of reliability studies. *Stat Methods Med Res* 1992; 1:123–157
56. Petersen MH, Hauge HN: Can training improve the results with infrared tympanic thermometers? *Acta Anaesthesiol Scand* 1997; 41:1066–1070
57. Lin LI, Chinchilli V: Rejoinder to the Letter to the Editor from Atkinson and Nevill. *Biometrics* 1997; 53:777–778
58. Royston P, Altman DG, Sauerbrei W: Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006;25: 127–141
59. Bartko JJ: Measures of agreement: A single procedure. *Stat Med* 1994; 13:737–745
60. Luiz RR, Szklo M: More than one statistical strategy to assess agreement of quantitative measurements may usefully be reported. *J Clin Epidemiol* 2005; 58:215–216
61. Kendall MG: A theory of randomness. *Biometrika* 1941; 32:1–15
62. Serlin RC, Wampold BE, Levin JR: Should providers of treatment be regarded as a random factor? If it ain't broke, don't "fix" it: A comment on Siemer and Joormann (2003). *Psychol Methods* 2003; 8:524–534