Comparing Methods of Clinical Measurement: Reporting Standards for Bland and Altman Analysis

Srinivas Mantha, MD*†, Michael F. Roizen, MD†, Lee A. Fleisher, MD, FACC§,
Ronald Thisted, PhD‡, and Joseph Foss, MD†

*Department of Anesthesiology and Intensive Care, Nizam's Institute of Medical Sciences, Hyderabad, India; Departments of †Anesthesia and Critical Care and ‡Statistics and Health Studies, University of Chicago, Chicago, Illinois; and §Department of Anesthesiology, Johns Hopkins School of Medicine, Baltimore, Maryland

In this era of medical technology assessment and evidence-based medicine, evaluating new methods to measure physiologic variables is facilitated by standardization of reporting results. It has been proposed that assessing repeatability be followed by assessing agreement with an established technique. If the "limits of agreement" (mean bias ± 2SD) are not clinically important, then one could use two measurements interchangeably. Generalizability to larger populations is facilitated by reporting confidence intervals. We identified 44 studies that compared methods of clinical measurement published during 1996 to 1998 in seven anesthesia journals. Although 42 of 44 (95.4%) used the limits of agreement methodology for analysis, several inadequacies and inconsistencies in reporting the results were noted. Limits of agreement were defined *a priori* in 7.1%, repeatability was evaluated in 21.4%, and relationship (pattern) between difference and average was evaluated in 7.1%. Only one of the articles reported confidence intervals. A computer macro for the Minitab statistical package (State College, PA**)** is described to facilitate reporting of Bland and Altman analysis with confidence intervals. We propose standardization of nomenclature in clinical measurement comparison studies. **Implications:** A literature review of anesthesia journals revealed several inadequacies and inconsistencies in statistical reports of results of comparison studies with regard to interchangeability of measurement methods. We encourage journal editors to evaluate submissions on this subject carefully to ensure that their readers can draw valid conclusions about the value of new technologies.

(Anesth Analg 2000;90:593–602)

Validation of new technology for application to clinical medicine requires comparison with older techniques or assessment of outcomes. These processes, known as medical technology assessment and evidence-based medicine, have gained prominence through publication frequency (1,2). A standard nomenclature has evolved for reporting results after comparison of new methods to monitor physiologic variables with established ones. Thus, for example, the performance of a new monitor to measure cardiac output is compared with an established thermodilution technique.

Statistical evaluations of such comparison studies are not simple. The primary aim of comparison studies is to determine whether the two methods agree sufficiently to be used interchangeability. Because analysis with correlation and least squares linear regression (also known as calibration statistics) is fundamentally misleading, Bland and Altman favored a different statistical method for assessing agreement between two methods of measurement (3–5). Their analysis first calculates the difference in measurement values obtained by two methods on the same subject. The mean of such differences in a sample of subjects is the estimated bias (difference between methods), and the standard deviation (SD) of the differences measures random fluctuations around this mean. If the "limits of agreement" (mean difference ± 2SD) between two methods are not clinically important, one can use the two methods interchangeably. Another essential feature of the analysis is graphical representation of the data with between-method difference (*y* axis) plotted against the average (*x* axis). Such a graph allows one to evaluate any relationship between the measurement of error (difference) and the assumed true value (average). Because results obtained in a study furnish only the sample statistics, it is necessary for generalizability of results to other populations to

Supported by a grant from Clinical Practice Enhancement and Anesthesia Research Foundation, Chicago, IL.

Accepted for publication November 23, 1999.

Address correspondence and reprint requests to Srinivas Mantha, MD, 13/4 RT, LIGH, Barkatpura, Hyderabad 500027, India. Address e-mail to smantha@satyam.net.in.

0003-2999/00

report confidence intervals (CIs) (6,7). CIs show a range of values based on the observed data within which, with a specified probability, the population value lies. In Bland and Altman analysis (4), CIs for mean bias, mean bias − 2SD, and mean bias + 2SD are of particular interest. We reviewed the statistical reporting of measurement comparison studies published in the anesthesia literature according to Bland and Altman analysis.

Methods

We examined the table of contents of seven anesthesia journals (*Anesthesiology, Anesthesia & Analgesia, Journal of Cardiothoracic and Vascular Anesthesia, Journal of Clinical Anesthesia, British Journal of Anesthesia, Anesthesia,* and *Canadian Journal of Anesthesia*) published between January 1996 to December 1998. Articles with titles indicating evaluation of a new measurement technique were read. The primary goal was to identify comparison studies in which interchangeability of a new measurement technique with an established method. Animal studies were excluded. To ensure accurate data transcription, each eligible study was read at least twice by one author (SM) and graded by written criteria by using an extraction chart for each article. A second author's (JFF) opinion was taken in case of confusion regarding data transcription. From each study, data were retrieved based on written evaluation standards. Random audits to ensure accuracy of some data from each article were done by a third author (MFR).

We evaluated the comparison studies according to Bland and Altman methodology (3–4) for the following five items: repeatability, definition of limits of agreement, representation of *x* axis on Bland and Altman graph, evaluation of relationship (pattern) between difference (*y* axis data), and average (*x* axis data), and report of CIs. For repeatability assessment of each study, we first determined whether repeatability is feasible (or practical), and then we determined whether repeatability was evaluated. Repeatability is determined by taking repeated measurements on a series of patients and calculating the mean and SD of differences. According to the definition of repeatability coefficient given by the British Standards Institute, the mean difference must not be significantly different from zero, and 95% of the differences are expected to lie within the range from −2SD to + 2SD of the mean (4). When reviewing a study for limits of agreement, two aspects were evaluated. We determined whether the authors correctly defined the limits as "mean bias ± 2SD." In the methods section of each article, we looked for a statement defining maximum width for limits of agreement which would not impair medical care i.e., *a priori* definition of the limits. We determined the *x* axis of a Bland and Altman graph for each study because of the potential for authors to erroneously use the *x* axis to represent the values of the established method rather than the average values of the two methods. The relationship (correlation) between difference in measurement values and their average is evaluated to verify whether differences vary in any systematic manner over the range of measurement (3,4).

Bland and Altman (4) derived the following formulas for CIs needed in the analysis:

For 95% CIs, *t* is the critical value for a 5% two-sided test drawn from tables of *t* distribution with $n - 1$ degrees of freedom (*df*), where *n* is the sample size.

The formula for calculating CI for mean bias (mean difference = $\bar{d}$) is: $\bar{d} \pm \mathrm{t} \times \mathrm{SD}/\sqrt{n}$, where SD = standard deviation of differences.
The formula for calculating CI for limits of agreement ($\bar{d}$ - 2SD and $\bar{d}$ + 2SD) is

$$\text{CI for mean} - 2\text{SD} = (\bar{d} - 2\mathrm{SD}) \pm \mathrm{t} \times (\sqrt{3\mathrm{SD}^2/n})$$

$$\text{CI for mean} + 2\text{SD} = (\bar{d} + 2\mathrm{SD}) \pm \mathrm{t} \times (\sqrt{3\mathrm{SD}^2/n}).$$

For each study, we determined the following items: physiological variable assessed; the principle of the new monitoring method; the established method used for comparison; whether Bland and Altman analysis was used; whether repeatability was evaluated; whether definition of limits of agreement were made *a priori* (i.e., described in the methods); whether the *x* axis of the comparison graph represented the average values of two methods or the values of the established method; whether relationship (pattern) between measurement error and the average value was evaluated; and whether CIs were reported.

Finally, we also tried to infer the definitions of some terms peculiar to measurement, such as accuracy, precision, and parameter (8–10). However, we did not evaluate the studies based on the use of these terms.

Results

We identified 66 articles in which a new measurement method was evaluated. Three animal studies were excluded, as were 19 studies in which interchangeability was not the primary goal. In two other studies, conclusions were based on correlation regression analysis. These exclusions left 42 articles for further examination (11–52). In all these studies, Bland and Altman analysis was used to project the results. Table 1 lists the statistical reporting of measurement comparison studies in these studies. We noted the use of Bland and Altman plot (difference versus average) in 38 articles (90.5%). Data transcription for evaluation and summarization was possible from all the but two studies (27,51). In these two studies, the opinion of one of the coauthors (JFF) was sought to solve the problem.

Table 1. Statistical Reporting of Measurement Comparison Studies from 42 Articles in the Anesthesia Literature

	Total (*n* = 42)	*Anesthesiology* (*n* = 5)	*Anesthesia & Analgesia* (*n* = 7)	*Journal of Cardiothoracic and Vascular Anesthesia* (*n* = 6)	*Journal of Clinical Anesthesia* (*n* = 2)	*British Journal of Anaesthesia* (*n* = 6)	*Anaesthesia* (*n* = 8)	*Canadian Journal of Anaesthesia* (*n* = 8)
Repeatability evaluated (R)	9 (21.4)	2 (40)	2 (28.6)	1 (16.7)	0 (0)	1 (16.7)	1 (12.5)	2 (25)
Limits of agreement defined *a priori* (L)	3 (7.1)	0 (0)	1 (14.3)	0 (0)	0 (0)	1 (16.7)	0 (0)	1 (12.5)
x axis equals average value of two methods[a] (X)	36 (94.7)[b]	4 (80)	6 (85.7)	5 (100)[b]	1 (100)[b]	6 (100)	8 (100)	6 (100)[b]
Independence of measurement error and average evaluated (I)	4 (9.5)	2 (40)	1 (14.3)	0 (0)	0 (0)	0 (0)	0 (0)	1 (12.5)
Confidence intervals reported (C)	1 (2.4)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	1 (12.5)	0 (0)

Values are *n* (%).

[a] Correctly plotted the *x* axis of a Bland and Altman graph.

[b] The denominators were 38, 5, 1, and 6, respectively, for the articles with graphic representations.

Table 2 describes the methodology and reporting of the studies by physiological variables in chronological order of publication. Cardiac output was the most common physiological variable studied (12 of 42, 28.6%), and thermodilution technique was the most commonly used method for comparison. In 39 of 42 articles (92.9%), study subjects were patients (in intraoperative, postoperative, or critical care settings), whereas the rest were volunteers. According to our impressions, repeatability was feasible or practical in all but three studies (32,34,45). Irrespective of our impressions, repeatability was evaluated in only nine studies (21.4%). If the three studies are excluded, then repeatability reporting is 23.1%. In all but two studies (11,33), the limits of agreement were correctly represented as ‘mean bias ± 2SD’. But, the limits of agreement were defined *a priori* (described in methods) in only three studies: two studies measured blood pressure (29,30), and one measured cardiac output (20). Two methods were judged to produce identical results in cardiac output measurement that varied substantially from the established thermodilution method. Finally, CIs for Bland and Altman statistics were reported in one study (38). At least three of five quality criteria set in our methods were satisfied in only three studies (20,29,38).

Examination for definition of terms revealed that, in 23 studies (54.8%), the term “precision” was defined in different ways such as bias ± 2SD, bias ± 1SD, 1SD of bias itself, or 2SD of bias itself. In one study, the descriptive statistics in the sample were represented as bias ± precision (27). The term 95% CI was misused to indicate bias ± 2SD in three studies (7.1%). We noted the use of the term “accuracy” in 20 studies (47.6%). In these studies, the term was used when a new method of measurement showed a good agreement with an established one.

Discussion

Error quantification is an important component in the evaluation of new measurement techniques. Bland and Altman analysis is a statistical technique that quantifies error for repeatability and limits of agreement (3–4). Our study identified several inadequacies and inconsistencies in the statistical reporting of studies in which new measurement systems were evaluated, although 95% of the studies used Bland and Altman methodology for analysis.

Repeatability is relevant in measurement comparison studies because poor repeatability (considerable variation in repeated measurements on the same subject) precludes the assessment of agreement between the two methods of measurement. Therefore, repeatability must be demonstrated before agreement between methods can be established.

A conclusion about interchangeability should not be based on mean bias alone but also should consider limits of agreement. For example, if a new instrument for noninvasive blood pressure measurement records systolic pressure as 120, 140, 110, 120, and 130 mm Hg in a sample of five subjects and the corresponding values obtained by direct arterial monitoring are 140, 110, 110, 100, and 160 mm Hg, respectively, then mean bias ± 2SD is 0 ± 51. This example illustrates that one can be misled in agreement evaluation if the conclusion is based on the mean bias alone disregarding the limits of agreement. This survey identified one study with such an error (28).

Table 2. Measurement Comparison Studies in Anesthesia Literature

First Author (reference), publication yr	Study methods, results, and authors' conclusion about interchangeability											Statistical reporting[a]				
	New method	Comparison method	Setting	Subvariable or condition	Patients (n)	Data sets (n)	Data sets/ patient or range	Units	Mean bias[b]	Limits of agreement bias −2SD to bias +2SD[b]	Inter-changeability	R	L	X	I	C
	Cardiac output															
Maslow (11), 1996	TEE Doppler of RVOT	Thermodilution	Cardiac surgery		38	38	1	L/min	0.01	−0.44 to 0.46	Y	–	–	Y	–	–
Bottiger (12), 1997	Continuous by thermal filament	Thermodilution	Liver transplantation		12	192	16	L/min	0.24	−3.34 to 3.81	Y	–	–	Y	–	–
Krishnamurthy (13), 1997	Esophageal Doppler	Continuous	CABG	Standard care	11	513	?	L/min	−0.7	−3.80 to 2.20	–	–	–	Y	–	–
				Close monitoring for wave forms	5	285	?		−0.14	−0.85 to −0.58	Y					
Greim (14), 1997	Continuous by thermal filament	Thermodilution	Liver transplantation	IFR > 1 L/h	14	270	17 to 20	L/min	1	−2.60 to 4.6	–	–	–	Y	–	–
				IFR < 1 L/h	14	270	17 to 20		0.2	−1.60 to 2.0	Y					
Imai (15), 1997	Pulse dye densitometry	Thermodilution	Cardiac/thoracic surgery	Entire group	22	191	3 to 6	L/min	0.16	−1.44 to 2.40	Y	–	–	Y	Y	–
				Output states												
				≤3.5 L/min	?	70	?	L/min	0.29	−0.88 to 1.46	–					
				3.5 to 6 L/min	?	99	?	L/min	0.104	−1.69 to 1.90	Y					
				>6 L/min	?	22	?	L/min	−0.032	−1.80 to 1.73	Y					
				Entire group	22	191	3 to 6	% error[c]	4.5	−34.70 to 43.70	Y					
				Output states												
				≤3.5 L/min	?	70		% error	9.3	−29.3 to 47.9	–					
				3.5 to 6 L/min	?	99		% error	2.2	−39.2 to 43.6	Y					
				>6 L/min	?	22		% error	0.5	−23.1 to 22.1	Y					
Aye (16), 1997	Visual estimation	Thermodilution	Cardiac surgery	By surgeon	35	1	1	L/min	0.46	−1.90 to 3.15	Y	–	–	?	–	–
				By anesthetist	35	1	1	L/min	0.59	−2.40 to 3.70	Y					
Lazor (17), 1997	Continuous	Thermodilution	Liver transplantation	Stat mode	29	108	?	L/min	−0.06	−12.8 to 1.16	Y	Y	–	Y	–	–
				Trend mode	29	108	?	L/min	−0.06	−1.04 to 0.92	Y					
Thangathurai (18), 1997	Continuous by bioimpedance	Thermodilution	Cancer surgery	Entire group	23	256	?	L/min	0.1	−1.90 to 2.10	Y	–	–	Y	–	–
				Original software	12	129	?	L/min	0.2	−2.20 to 2.60	Y					
				Revised software	11	127	?	L/min	−0.1	−1.70 to 1.50	Y					
Perrino (19), 1998	Multiplane TEE with Doppler	Thermodilution	Cardiac and noncardiac surgery		33	110	?	L/min	0.01	−1.11 to 1.13	Y	Y	–	Y	–	–
Seguin (20), 1998	Semicontinuous by thermal filament	Thermodilution	Critical care unit		15	87	6	L/min	−0.002	−1.48 to 1.47	–	Y	Y	Y	–	–
Colbert (21), 1998	Esophageal Doppler	Thermodilution	Liver transplantation		18	234	13	L/min	0.07	−4.20 to 4.32	–	–	–	Y	–	–
Gust (22), 1998	Transpulmonary thermodilution	Thermodilution	Critical care unit		75	375	?	L/min	0.456	−1.86 to 2.77	Y	–	–	NA	–	–
	Blood gases															
Zollinger (23), 1997	Intraarterial Paratrend 7	Laboratory blood gas analysis	Thoracoscopic surgery	Pao$_2$	23	1381	6	mm Hg	2.88	−69.16 to 72.24	Y	Y	–	Y	–	–
				Paco$_2$	23	138	6	mm Hg	2.36	−3.42 to 8.13	Y					
				pH	23	138	6	Units	−0.017	−0.082 to 0.048	Y					
Tobias (24), 1997	Transcutaneous and ET co$_2$	Laboratory blood gas analysis	Infants and toddlers on IPPV	Transcutaneous CO$_2$	25	100	not >5	mm Hg	−0.68	−5.38 to 4.02	Y	–	–	–	–	–
				ETco$_2$	25	100	not >5	mm Hg	−6.68	−16.7 to 3.34	–					

Hatherhill (25), 1997	Intraarterial Paratrend 7	Laboratory blood gas analysis	Cyanotic heart disease children	Pao_2	10	100	10	mm Hg	0.3	−6.31 to 6.92	Y	–	–	Y	–	–
				$Paco_2$	10	100	10	mm Hg	−3.34	−8.97 to 2.28	Y					
				pH	10	100	10	Units	0.02	−0.04 to 0.08	Y					
Tobias (26), 1998	Intraarterial Paratrend 7	Laboratory blood gas analysis	Children in respiratory failure	$Paco_2$	4	17	3 to 6	mm Hg	3.0	0.2 to 5.8	Y	–	–	NA	–	–
				pH	4	17	3 to 6	Units	0.036	−0.004 to 0.076	Y					
Ishikawa (27), 1998	Intraarterial Paratrend 7	Laboratory blood gas analysis	One-lung ventilation	Entire group								–	–	NA	–	–
				Pao_2	12	84		mm Hg	−1	−41 to 39	–					
				% O_2 saturation	12	84	?	%	0.8	−20.8 to 22.4	–					
				$Paco_2$	12	84	?	mm Hg	0.9	−2.20 to 4.10	Y					
				pH	12	84	?	Units	0	−0.02 to 0.02	Y					
				At Pao_2 <150 mm Hg												
				Pao_2	?	?	?	mm Hg	−5	−22 to 12	–					
				% O_2 saturation	?	?	?	%	−2.1	−22.8 to 18.6	–					
	Blood pressure															
Green (28), 1996	Neonatal noninvasive module	Direct arterial measure of blood pressure	Major surgery	Systolic BP	18	1258	?	mm Hg	−9.1	−37.9 to 19.7	Y	–	–	Y	–	–
				Diastolic BP	18	1258	?	mm Hg	7.9	−15.5 to 31.3	Y					
				Mean BP	18	1258	?	mm Hg	0.7	−23.1 to 24.5	Y					
Kaufmann (29), 1996	Noninvasive oscillometric	Spacelabs Dinamap Marquette	Electroconvulsive therapy	Systolic BP[d]	12	182	?	mm Hg	1.6	−16.6 to 19.8	–	Y	Y	Y	–	–
				Diastolic BP[d]	12	182	?	mm Hg	7.3	−9.3 to 23.9	–					
				Mean BP[d]	12	182	?	mm Hg	2.8	−17 to 22.6	–					
				Systolic BP[e]	12	182	?	mm Hg	0.8	−22.8 to 24.4	–					
				Diastolic BP[e]	12	182	?	mm Hg	0.3	−19.1 to 19.7	–					
				Mean BP[e]	12	182	?	mm Hg	−1.7	−24.7 to 21.3	–					
Weiss (30), 1996	Continuous tonometric	Direct arterial	Surgery	Systolic	22	1375	?	mm Hg	−5.8	−34.2 to 22.6	–	–	Y	Y	–	–
				Diastolic	22	1375	?	mm Hg	7.2	−9.4 to 23.8	–					
				Mean	22	1375	?	mm Hg	3.9	−13.7 to 21.5	–					
	Cerebral oximetry															
Henson (31), 1998	Near infrared spectroscopy	Jugular venous oxygen saturation	Volunteers		30	360	?	%	3.8	−14 to 21.6	–	–	–	Y	Y	–
Buunk (32), 1998	Near infrared spectroscopy Gastric Mucosal CO_2	Jugular venous oxygen saturation	After cardiac resuscitation		10	176	16 to 18	%	−4	−23.8 to 15.7	–	–	–	Y	–	–
Creuter (33), 1997	Gas filled balloon	Saline-filled balloon	Ventilated patients		7	84	?	mm Hg	−0.1	−6.9 to 6.7	Y	Y	–	Y	–	–
Janssens (34), 1998	Gas filled balloon	Saline-filled balloon	ICU patients in shock		19	237	?	mm Hg	−0.3	−2.7 to 2.1	Y	–	–	Y	–	–
	Stroke volume															
Greim (35), 1996	Automated border detection by TEE	Thermodilution	Abdominal surgery	Multidisc formula	12	114	?	mL	−26	−68 to 16	–	–	–	Y	–	–
				Area length formula	12	114	?	mL	−28	−72 to 16	–					
Woltjer (36), 1996	Impedance cardiography	Thermodilution	After cardiac surgery	LSA with Kb Eq	37	37	1	mL	−27.9	−51 to −4	–	–	–	Y	–	–
				LSA with Sm Eq	37	37	1	mL	−2.7	−32 to 26.6	–					
				SSA with Kb Eq	37	37	1	mL	0.5	−16.6 to 17.6	Y					
				SSA with Sm Eq	37	37	1	mL	19.3	−13.9 to 52.5	–					
	Hemoglobin															
Jaeger (37), 1996	HemoCue	Laboratory hemoglobin value	Cardiac surgery	Arterial sample	12	48	4	g/dL	0.1	−0.2 to 0.4	Y	Y	–	Y	–	–
				Capillary sample	12	48	4	g/dL	0.8	−0.2 to 1.8	–					

Table 2. (*Continued*)

	Study methods, results, and authors' conclusion about interchangeability											Statistical reporting[a]				
First Author (reference), publication yr	New method	Comparison method	Setting	Subvariable or condition	Patients (n)	Data sets (n)	Data sets/points or range	Units	Mean bias[b]	Limits of agreement bias −2SD to bias +2SD[b]	Inter-change-ability	R	L	X	I	C
Lardi (38), 1998	HemoCue	Laboratory hemoglobin value	Aortic surgery		13	52	4	g/dL	0.038	−0.37 to 0.45	Y	Y	–	Y	–	Y
	NMJ Monitoring															
McCluskey (39), 1997	Acceleromyography	Mechanomyography	Pediatric surgery		15	15	1	%	−25	−62 to 12	–	–	–	Y	–	–
Nakata (40), 1998	Acceleromyography	Electromyography	Surgery	Xenon usage[f]	11	11	1	%	−24	−59 to 11	–	–	–	NA	–	–
				Isoflurane usage[f]	17	17	1	%	−28	−61 to 5	–					
	Temperature															
Matsukawa (41), 1996	Infrared tympanic probe of different makes	Thermocouple	Volunteers	Genius	50	50	1	°C	0.34	−0.32 to 1	Y	–	–	Y	–	–
				Thermopit	50	50	1	°C	−0.73	−1.47 to 0.01	Y					
				Quickthermo	50	50	1	°C	−0.42	−1.12 to 0.28	Y					
				Thermoscan	50	50	1	°C	0.3	−0.4 to 1	Y					
Patel (42), 1996	Liquid crystal thermometry	Temperature measured at different sites	Surgery	Tympanic vs skin	40	477	?	°C	0.5	−1.48 to 2.52	–	–	–	Y	–	–
				Esophagus vs skin	40	289	?	°C	0.3	−1.64 to 2.32	–					
				Esophagus vs tympanic	40	329	?	°C	−0.1	−1.02 to 0.74	–					
	Blood volume															
Iijima (43), 1998	Pulse dye densitometry	Radioimmunoassay	Volunteers		11	11	1	% difference	3.99	−17.09 to 25.07	Y	–	–	–	–	–
	Mitral regurgitant flow															
Kolev (44), 1998	Proximal isovelocity surface area	Angiography	Cardiac surgery		33	33	1	mL	2.47	−16.9 to 21.9	Y	–	–	Y	–	–
	ST segment analysis															
Wajon (45), 1998	Merlin bedside monitor	ST segment analysis by cardiologist	After coronary Bypass surgery	Lead II	24	24	1	mm	0.1	−0.9 to 1.1	Y	–	–	Y	–	–
				V5	23	23	1	mm	−0.1	−1.9 to 1.7	Y					
	PCWP															
Nomura (46), 1997	Doppler transmitral flow variables	Pulmonary artery catheter	After CABG with EF <35%		16	16	1	mm Hg	0.55	−7.19 to 8.29	–	Y	–	Y	–	–
	Jugular venous O_2 saturation															
Trubiano (47), 1996	Fibreoptic catheter	Cooximetry	Cardiopulmonary bypass		20	100	5	%	?	−33 to 22	–	–	–	Y	Y	–
	Pulmonary artery blood gases															
Franklin (48), 1996	Fluorescent optode	Laboratory value of mixed venous sample	ICU patients	pH	14	96	4 to 6	Units	−0.004	−0.036 to 0.028	Y	–	–	Y	–	–
				Po_2	14	96	4 to 6	mm Hg	−1.9	−7.3 to 3.5	Y					
				Pco_2	14	96	4 to 6	mm Hg	−2.4	−6.8 to 2	Y					

	Heparin blood level															
Wahr (49), 1996	Electrochemical sensor	Hepcon	Cardiac surgery		24	162		U/mL	0.211	−0.75 to 1.16	Y	–	–	Y	–	–
	APTT															
Heap (50), 1997	Arterial sample	Venous sample	ICU patients	4.5 mL discard vol	92	92	1	s	1.24	−4.39 to 6.98	Y	–	–	Y	–	–
				16 mL discard vol	92	92	1	s	0.89	−3.25 to 5.03	Y					
	Plasma volume															
Sekimoto (51), 1997	Biexponential indocyanine green decay with 5–30-min regression time	Other regression times	Volunteers	Regression times						–		–	–	Y	–	–
				EOR to 30 min	15	15	1	mL/kg	−27.6	52.4 to −2.8	–					
				3 to 30 min	15	15	1	mL/kg	0.91	−2.41 to 4.23	Y					
				7 to 30 min	15	15	1	mL/kg	0.23	−1.73 to 2.2	Y					
	Central venous pressure															
Alzeer (52), 1998	Right iliac vein pressure	Right atrial pressure	Intensive care unit patients		26	26	1	mm Hg	−0.93	−1.77 to 1.93	Y	–	–	Y	Y	–

TEE = transesophageal echocardiography, RVOT = right ventricular outflow tract, CABG = coronary artery bypass graft surgery, IFR = peripheral intravenous infusion flow rates, IPPV = intermittent positive pressure ventilation, BP = blood pressure, ICU = intensive care unit, LSA with Kb Eq = lateral spot electrode array with Kubicek equation, LSA with Sb Eq = lateral spot electrode array with Sramek-Bernstein equation, SSA with Kb Eq = modified semicircular electrode array with Kubicek equation, SSA Sb Eq = modified semicircular electrode array with Sramek-Bernstein equation, NMJ = neuromuscular junction, PCWP = pulmonary capillary wedge pressure, EF = ejection fraction, APTT = activated partial thromboplastin time, vol = volume, EOR = end of recirculation.

Paratrend 7, Biomedical Sensor Ltd, Pfizer Hospital Products Group, High Wycombe, UK; Dinamap, Critikon, Tampa, FL; SpaceLabs, SpaceLabs Medical, Redmond, WA; Marquette, Marquette Electronics, Milwaukee, WI; HemoCue, Abbot Laboratories, San Jose, CA; Genius®, Sherwood IMS, IMC, Inc., CA; Thermopit®, Nipro Inc., Osaka, Japan; Quickthermo®, Omron Inc., Mie, Japan; Thermoscan®, Thermoscan, Inc., CA; Merlin, Hewlett-Packard, Australia Ltd. North Ryde, NSW; Hepcon, Medtronic Hemotec, Parker, CO.

[a] Statistical reporting for the entire study, not for each subvariable or condition. Refer to Table 1, Column 1 for R, L, X, I, and C.

[b] Mean bias and limits of agreement are presented as originally reported by the authors when measurement value of comparison method was subtracted from the new one. Otherwise, the bias and limits of agreement were corrected for uniformity of presentation. For example, Maslow et al. (11) obtained −0.01, −0.46, and 0.44 L/min for mean bias, mean bias −2SD, and mean bias +2SD, respectively, when the new method value was subtracted from the old one. The values were corrected to 0.01, −0.44, and 0.46, respectively. The other references for which such correction was made are 12, 13, 17, 19, 20, 28, 33, 34, 40, 41, 45, 51, and 52. Similarly, pressure expressed in kPa units is converted to mm Hg for data in References 23 and 25, and hemoglobin value in g/L is corrected to g/dL for data in Reference 37.

[c] Implies an absolute difference in cardiac output: 100 × (difference in cardiac output between two methods)/(mean of cardiac output by two methods).

[d] Comparison between SpaceLabs and Dinamap monitors.

[e] Comparison between Marquette and Dinamap monitors.

[f] Results when T1/Tc ratio of train-of-four is 0.15 by acceleromyography. Similar results are obtained by the authors with T1/Tc ratios of 0.2 and 0.25 under both types of anesthetic agents.

– = No, Y = yes, NA = not applicable, ? = not defined or not clear.

Ideally, the limits of agreement need to be defined *a priori* in the methods, and such a definition was given in only three studies (20,29,30). The American National Standards of the Association for the Advancement of Medical Instrumentation recommend that maximal bias of noninvasive arterial pressure, obtained from at least 85 patients, should not exceed 5 mm Hg ± 8 SD from a noninvasive reference method (53). The British Hypertension Society considered the above criterion too liberal and proposed an alternative grading system according to the percentage of readings ≤ 5, ≤ 10, ≤ 15 mm Hg from a noninvasive reference method (54). Unfortunately, both these criteria are not readily applicable in perioperative settings because these guidelines were planned for evaluating blood pressure instruments used in outpatient clinics. In perioperative settings, an invasive reference standard is usual. One cardiac output study defined the limits of agreement *a priori* as ±1 L/min (20). Although not described in methods, two studies used valid criteria for limits of agreement while evaluating results (23,39). The intraarterial blood gas monitoring study (23) used published guidelines (55) to evaluate its results. The limits of agreement for blood gas measurements are as follows: Po_2 range, 30.4 to 152 mm Hg; Pco_2 range, 20.5 to 80.56 mm Hg; the limits must be ±4.6 mm Hg of the reference. In another study in which an intraoperative hemoglobin monitor was evaluated (39), the limits were empirically defined as ±1 g/dL from the laboratory reference method. Defining the limits of agreement for different physiologic variables may be a difficult aspect in designing the measurement comparison studies, especially in perioperative and critical care settings, because action limits (clinically important) depend upon the clinical scenario and the status of other related variables. Nevertheless, an attempt must be made to define such limits at a minimum after pooling data from other studies. Alternatively, a delphi survey (opinion from experts) may be used to design the study. Without *a priori* setting of limits, widely discrepant limits of agreement have been chosen (Table 2). Such varying limits seem too difficult to accept in practice and may mislead clinicians who are inexperienced in technology of evidence-based analysis.

The *x* axis of the Bland and Altman analysis should ideally be represented by the average of measurement values obtained by two different methods because true value is unknown. Bland and Altman proved mathematically that the *x* axis must represent the average values of two methods (5). Three studies used values obtained by the established method alone on the *x* axis.

The plot of difference against average in Bland and Altman analysis also allows us to investigate any possible relationship (correlation) between measurement error (difference between two methods) and the assumed true value (average value of two methods). Bland and Altman's suggestions are subject to the assumption that there is no pattern in the plot of difference versus average (3,4). The correlation coefficient could be tested against the null hypothesis of $r = 0$ for a formal test of independence. Ideally, such independence should also be demonstrated during a repeatability experiment for each of the two methods. In other words, it is important to ensure that within-subject repeatability is not associated with the size of measurements. Otherwise, results of subsequent analysis might be misleading (3).

Although the computational scheme for CIs for Bland and Altman statistics is easy to comprehend, the algebraic calculations are tedious for repeated use. We devised a macro (see Appendix 1) for Minitab (Release 10 and above; Minitab Inc., State College, PA) to facilitate such computation and present it graphically. Minitab is statistical software that can be used for medical applications (56).

Finally, standardization of nomenclature is an important issue in scientific writing. It is common to find the terms "accuracy" and "precision" in measurement comparison studies (8–10). Accuracy is defined as closeness of a measurement to its true value, and the term is used when a method is compared with an external standard. In practice, one is rarely comparing a measurement with the true value because a "gold-standard" method need not necessarily give the true value. Therefore it may be preferable to avoid the word accuracy in these contexts, and use of the term "agreement" may be preferable (D. G. Altman and M. J. Bland, written communication, 1999). Precision refers to closeness of values on repeated measurements obtained by the same method, i.e., a measure of repeatability. Confusion may arise with the use of the term "precision" because of another definition found in statistical literature. A statistical dictionary (57) defines it as follows: "precision of an estimator is its tendency to have its values cluster closely about the mean of its sampling distribution." Thus, precision is related inversely to the variance of this sampling distribution—the smaller the variance, the greater is the precision. In fact, Bland and Altman used the term "precision" in the context of reporting CIs (4). In our survey of articles, "precision" was the most common incorrectly defined term and was used in contexts other than repeatability or reporting CIs. Therefore, in measurement comparison studies, avoiding the term "precision" and using the term "repeatability" may seem reasonable. If used, the term must clearly be defined (D. G. Altman, written communication, 1999). In medical literature, it is also common to find the word "parameter" used for "variable," as in "We measured the following parameters: temperature, arterial blood pressure, pulse oximetry, end-tidal carbon dioxide and cardiac output." In statistical literature, the term "variable" refers to quantities that vary from individual to individual. The term "parameter" refers to quantities defining a theoretical model (58) and is used to indicate numerical characteristics of a population that are analogous to the numerical characteristics of a sample (statistics). The unknown population parameter is estimated from a sample of values of a variable. Therefore substitution of the specific statistical term "parameter" for "variable" must be avoided.

In this era of evidence-based medicine, standardization of statistical reporting of studies facilitates easy appraisal of published material. This survey has identified several inadequacies and inconsistencies in statistical reporting of measurement comparison studies. Such inadequacies render the validity of the conclusions in each of the articles in doubt. We encourage journal editors to evaluate submissions on this subject carefully to ensure that their readers can draw valid conclusions about the value of new technologies.

Appendix 1

The macro files in Minitab use the default extension MAC. For example, this macro can be baa.mac. It must be stored in the macros subdirectory (in the Windows version) or a folder (the Macintosh version) under the main Mintab directory or folder. The macro is invoked by the following command: %baa c4 c6, if the measurement values for the two methods are entered in Columns 4 and 6 of Mintab's worksheet. After the macro is invoked, the user is asked whether the graph should be plotted with confidence intervals or just with mean bias, bias −2SD, and bias + 2SD. After the appropriate response (yes or no) from the user, the macro performs the required calculations. The text output of the macro includes confidence intervals no matter which graphical output is chosen. The macro is available for downloading from our Web site: http://mantha.uchicago.edu.

The authors thank Sally Kozlik for editorial assistance.

References

1. Fleisher LA, Mantha S, Roizen MF. Medical technology assessment: an overview. Anesth Analg 1998;87:1271–82.
2. Pace NL. Technology assessment of anesthesia monitors. J Clin Monit 1992;8:142–6.
3. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17.
4. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–10.

5. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346:1085–7.
6. Gardner MJ, Altman DG. Estimating with confidence. Br Med J 1988;296:1210–1.
7. Mantha S. Scientific approach to presenting and summarizing data [letter]. Anesth Analg 1992;75:469–70.
8. Westgard JO, Hunt MR. Use and interpretation of common statistical tests in method-comparison studies. Clin Chem 1973; 19:49–57.
9. Fisher LD, van Belle G: Biostatistics: a methodology for the health sciences. New York: John Wiley, 1993.
10. LaMantia KR, O'Connor T, Barash PG. Comparing methods of measurement: an alternative approach. Anesthesiology 1990;72: 781–3.
11. Maslow A, Comunale ME, Haering JM, Watkins J. Pulsed wave Doppler measurement of cardiac output from the right ventricular outflow tract. Anesth Analg 1996;83:466–71.
12. Bottiger BW, Sinner B, Motsch J, et al. Continuous versus intermittent thermodilution cardiac output measurement during orthotopic liver transplantation. Anaesthesia 1997;52:207–14.
13. Krishnamurthy B, McMurray TJ, McClean E. The peri-operative use of the oesophageal Doppler monitor in patients undergoing coronary artery revascularisation: a comparison with the continuous cardiac output monitor. Anaesthesia 1997;52:624–9.
14. Greim CA, Roewer N, Thiel H, et al. Continuous cardiac output monitoring during adult liver transplantation: thermal filament technique versus bolus thermodilution. Anesth Analg 1997;85: 483–8.
15. Imai T, Takahashi K, Fukura H, Morishita Y. Measurement of cardiac output by pulse dye densitometry using indocyanine green. Anesthesiology 1997;87:816–22.
16. Aye T, Milne B, Ballantyne M. Cardiac output estimation by visual inspection vs thermodilution during cardiac surgery. Can J Anaesth 1997;44:126–30.
17. Lazor MA, Pierce ET, Stanley GD, et al. Evaluation of the accuracy and response time of STAT-mode continuous cardiac output. J Cardiothorac Vasc Anesth 1997;11:432–6.
18. Thangathurai D, Charbonnet C, Roessler P, et al. Continuous intraoperative noninvasive cardiac output monitoring using a new thoracic bioimpedance device. J Cardiothorac Vasc Anesth 1997;11:440–4.
19. Perrino AC Jr, Harris SN, Luther MA. Intraoperative determination of cardiac output using multiplane transesophageal echocardiography: a comparison to thermodilution. Anesthesiology 1998;89:350–7.
20. Seguin P, Colcanap O, Le Rouzo A, et al. Evaluation of a new semi-continuous cardiac output system in the intensive care unit. Can J Anaesth 1998;45:578–83.
21. Colbert S, O'Hanlon DM, Duranteau J, Ecoffey C. Cardiac output during liver transplantation. Can J Anaesth 1998;45:133–8.
22. Gust R, Gottschalk A, Bauer H, et al. Cardiac output measurement by transpulmonary versus conventional thermodilution technique in intensive care patients after coronary artery bypass grafting. J Cardiothorac Vasc Anesth 1998;12:519–22.
23. Zollinger A, Spahn DR, Singer T, et al. Accuracy and clinical performance of a continuous intra-arterial blood-gas monitoring system during thoracoscopic surgery. Br J Anaesth 1997;79: 47–52.
24. Tobias JD, Meyer DJ. Noninvasive monitoring of carbon dioxide during respiratory failure in toddlers and infants: end-tidal versus transcutaneous carbon dioxide. Anesth Analg 1997;85: 55–8.
25. Hatherill M, Tibby SM, Durward A, et al. Continuous intra-arterial blood-gas monitoring in infants and children with cyanotic heart disease. Br J Anaesth 1997;79:665–7.
26. Tobias JD, Meyer DJ, Helikson MA. Monitoring of a pH and PCO_2 in children using the Paratrend 7 in a peripheral vein. Can J Anaesth 1998;45:81–3.
27. Ishikawa S, Makita K, Nakazawa K, Amaha K. Continuous intra-arterial blood gas monitoring during oesophagectomy. Can J Anaesth 1998;45:273–6.
28. Green DW. Use of a neonatal noninvasive blood pressure module on adult patients. Anaesthesia 1996;51:1129–32.
29. Kaufmann MA, Pargger H, Drop LJ. Oscillometric blood pressure measurements by different devices are not interchangeable. Anesth Analg 1996;82:377–81.
30. Weiss BM, Spahn DR, Rahmig H, et al. Radial artery tonometry: moderately accurate but unpredictable technique of continuous non-invasive arterial pressure measurement. Br J Anaesth 1996; 76:405–11.
31. Henson LC, Calalang C, Temp JA, Ward DS. Accuracy of a cerebral oximeter in healthy volunteers under conditions of isocapnic hypoxia. Anesthesiology 1998;88:58–65.
32. Buunk G, van der Hoeven JG, Meinders AE. A comparison of near-infrared spectroscopy and jugular bulb oximetry in comatose patients resuscitated from a cardiac arrest. Anaesthesia 1998;53:13–9.
33. Creteur J, De Backer D, Vincent JL. Monitoring gastric mucosal carbon dioxide pressure using gas tonometry: in vitro and in vivo validation studies. Anesthesiology 1997;87:504–10.
34. Janssens U, Graf J, Koch KC, Hanrath P. Gastric tonometry: *in vivo* comparison of saline and air tonometry in patients with cardiogenic shock. Br J Anaesth 1998;81:676–80.
35. Greim CA, Roewer N, Laux G, Schulte am Esch J. On-line estimation of left ventricular stroke volume using transoesophageal echocardiography and acoustic quantification. Br J Anaesth 1996;77:365–9.
36. Woltjer HH, Bogaard HJ, Scheffer GJ, et al. Standardization of noninvasive impedance cardiography for assessment of stroke volume: comparison with thermodilution. Br J Anaesth 1996;77: 748–52.
37. Jaeger M, Ashbury T, Adams M, Duncan P. Perioperative on-site haemoglobin determination: as accurate as laboratory values? Can J Anaesth 1996;43:795–8.
38. Lardi AM, Hirst C, Mortimer AJ, McCollum CN. Evaluation of the HemoCue for measuring intra-operative haemoglobin concentrations: a comparison with the Coulter Max-M. Anaesthesia 1998;53:349–52.
39. McCluskey A, Meakin G, Hopkinson JM, Baker RD. A comparison of acceleromyography and mechanomyography for determination of the dose-response curve of rocuronium in children. Anaesthesia 1997;52:345–9.
40. Nakata Y, Goto T, Saito H, et al. Comparison of acceleromyography and electromyography in vecuronium-induced neuromuscular blockade with xenon or sevoflurane anesthesia. J Clin Anesth 1998;10:200–3.
41. Matsukawa T, Ozaki M, Hanagata K, et al. A comparison of four infrared tympanic thermometers with tympanic membrane temperatures measured by thermocouples. Can J Anaesth 1996; 43:1224–8.
42. Patel N, Smith CE, Pinchak AC, Hagen JF. Comparison of esophageal, tympanic, and forehead skin temperatures in adult patients. J Clin Anesth 1996;8:462–8.
43. Iijima T, Iwao Y, Sankawa H. Circulating blood volume measured by pulse dye-densitometry: comparison with (131)I-HSA analysis. Anesthesiology 1998;89:1329–35.
44. Kolev N, Brase R, Wolner E, Zimpfer M. Quantification of mitral regurgitant flow using proximal isovelocity surface area method: a transesophageal echocardiography perioperative study. J Cardiothorac Vasc Anesth 1998;12:22–6.
45. Wajon P, Lindsay G. Detection of postoperative myocardial ischemia by bedside ST-segment analysis in coronary artery bypass graft patients. J Cardiothorac Vasc Anesth 1998;12: 620–4.
46. Nomura M, Hillel Z, Shih H, et al. The association between Doppler transmitral flow variable measured by transesophageal echocardiography and pulmonary capillary wedge pressure. Anesth Analg 1997;84:491–6.
47. Trubiano P, Heyer EJ, Adams DC, et al. Jugular venous bulb oxyhemoglobin saturation during cardiac surgery: accuracy and reliability using a continuous monitor. Anesth Analg 1996;82: 964–8.

48. Franklin ML, Peruzzi WT, Moen SG, Shapiro BA. Evaluation of an on-demand, *ex vivo* bedside blood gas monitor on pulmonary artery blood gas determinations. Anesth Analg 1996;83:500–4.
49. Wahr JA, Yun JH, Yang VC, et al. A new method of measuring heparin levels in whole blood by protamine titration using a heparin-responsive electrochemical sensor. J Cardiothorac Vasc Anesth 1996;10:447–50.
50. Heap MJ, Ridley SA, Hodson K, Martos FJ. Are coagulation studies on blood sampled from arterial lines valid? Anaesthesia 1997;52:640–5.
51. Sekimoto M, Fukui M, Fujita K. Plasma volume estimation using indocyanine green with biexponential regression analysis of the decay curves. Anaesthesia 1997;52:1166–72.
52. Alzeer A, Arora S, Ansari Z, et al. Central venous pressure from common iliac vein reflects right atrial pressure. Can J Anaesth 1998;45:798–801.
53. Proposed standard for electronic or automated sphygmomanometers. Arlington, VA: Association for the Advancement of Medical Instrumentation, 1992.
54. O'Brien E, Petrie J, Littler W, et al. The British Hypertension Society protocol for the evaluation of automated and semiautomated blood pressure measuring devices with special reference to ambulatory systems. J Hypertens 1990;8:607–19.
55. Blood gas/pH analyzers. Health Devices 1995;24:208–43.
56. Minitab reference manual. Release 10 for Windows. State College, PA: Minitab, 1994.
57. Freund JE, Williams FJ. Dictionary/outline of basic statistics. New York: Dover Publications, 1966.
58. Altman DG, Bland JM. Statistics notes: variables and parameters. Br Med J 1999;318:1667.