

# OPEN Metrology in Medicine: From Measurements to Decision, with Specific Reference to Anesthesia and Intensive Care

Pierre Squara, MD,\* Michael Imhoff, MD, PhD,† and Maurizio Cecconi, MD, PhD‡

Metrology is the science of measurements. Although of critical importance in medicine and especially in critical care, frequent confusion in terms and definitions impact either interphysician communications or understanding of manufacturers' and engineers' instructions and limitations when using devices. In this review, we first list the terms defined by the International Bureau of Weights and Measures regarding quantities and units, measurements, devices for measurement, properties of measuring devices, and measurement standards. The traditional tools for assessing the most important measurement quality criteria are also reviewed with clinical examples for diagnosis, alarm, and titration purposes, as well as for assessing the uncertainty of reference methods. (Anesth Analg 2015;120:66–75)

The International Bureau of Weights and Measures defines **metrology** as the **science** of **measurement**, embracing both experimental and theoretical determinations at any level of uncertainty in any field of science and technology.<sup>1</sup>

Measurements of physiological, biochemical, physical, and other patient-related variables are not only ubiquitous in intensive care medicine and beyond, but the results from such measurements also provide essential information for critical decision-making in clinical practice, as well as for research and technology development. Erroneous measurements can jeopardize patient safety and can expose the most critically ill patients to severe hazards. If physiological variables cannot be measured properly, then therapy-targeting changes in those variables cannot be adjusted properly.

Understanding metrological concepts and recognizing limitations and constraints help to interpret clinical study results and in technological assessment of new medical devices. It is crucially important that physicians share with other scientists the same understanding of objectives, terms, units, and criteria attributed to measurements. This is of particular importance in anesthesiology and intensive care medicine, owing to the large and increasing number of medical devices that are involved in clinical decision-making.

This review addresses the most frequent metrological misunderstandings encountered in critical care and their impact on daily practice. Following other authors,<sup>2</sup> it is also a plea for avoiding frequent ambiguity in the use of terms and aligning our terminology to the most recent common standards. Finally

and most importantly, we try to provide guidance for metrological concepts and their relevance for clinical decision-making.

## DEFINITIONS OF IMPORTANT CONCEPTS

The complete metrological list of terms divided into 5 main headings are Quantities and Units, Measurement, Devices for Measurement, Properties of Measuring Devices, and Measurement Standards (Etalons) and can be found in the Joint Committee for Guides in Metrology.<sup>1</sup> A summary is given in Table 1.

### Quantities and Units

A **quantity** is a **property** of a phenomenon, body, or substance, to which is attributed a **magnitude** that can be expressed as a **number** and a **reference**. A quantity is a scalar. However, a vector or a tensor, the components of which are quantities, is also considered to be a quantity. A reference can be a measurement unit, a measurement procedure, a reference material, or a combination of such. A quantity is characterized by a dimension, a unit, and a value. There are 7 base quantities on which international quantities are based. They are listed in Table 2 with 5 other quantities often used in medicine. The complete list can be obtained from the International Bureau of Weight And Measures.<sup>3</sup>

### Measurement

A **measurement** is a **process** of **experimentally** obtaining one or more **values** that can **reasonably** be **attributed** to a **quantity**. The **true value** of a quantity is unique at a specific time and always **unknowable**. Therefore, a measurement result is generally expressed as a single measured quantity value and a measurement **uncertainty**. The measurand is the quantity to be measured. A measurement method is based on a principle; a physical, chemical, or biological phenomenon serving as the basis of measurement. For example, the thermoelectric effect is a measurement method of temperature and the infrared spectroscopy of the HbO<sub>2</sub> concentration. A reference measurement procedure is a procedure accepted as providing measurement results that fit for their intended use. Although it has no international definition, a **"gold standard"** is supposed to be the best practically available reference method. The difference between a measurement and a reference value is characterized by different properties.

From the \*ICU Department, Réanimation CERIC, Clinique Ambroise Paré, Neuilly, France; †Abteilung für Medizinische Informatik, Biometrie und Epidemiologie, Ruhr-Universität Bochum, Bochum, Germany; and ‡Adult Critical Care Directorate and Anaesthesia, St. George's Hospital and St. George's University of London, London, United Kingdom.

Accepted for publication August 21, 2014.

Funding: None.

The authors declare no conflicts of interest.

Reprints will not be available from the authors.

Address correspondence to Pierre Squara, MD, ICU Department, Clinique Ambroise Paré, 27 bd Victor Hugo, 92200 Neuilly, France. Address e-mail to pierre.squara@orange.fr.

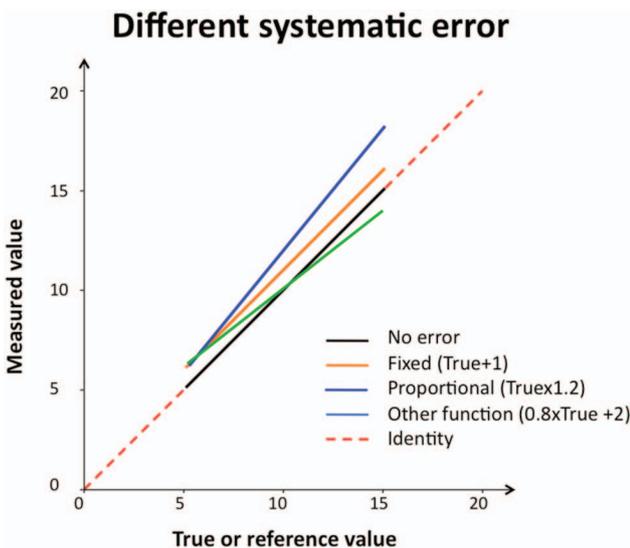
Copyright © 2014 International Anesthesia Research Society. This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License, where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially.

DOI: 10.1213/ANE.0000000000000477

**Table 1. Summary of Terms and Definitions**

Name	Simplified definition	Assessment
<b>Measurement</b>		
Measurement	Process of experimentally obtaining one or more value that can be attributed to a quantity.	
Measurand	Quantity to be measured.	
Measurement method	Method based on a principle; a physical, chemical, or biological phenomenon serving as basis of measurement.	
Reference procedure	Procedure accepted as providing measurement results that fit for their intended use.	
Accuracy	Closeness of agreement between a measurement value and a true value of the measurand.	Measurement error, percentage error
Trueness	Closeness of agreement between the average of an infinite number of replicate measurements and the true or reference value.	Systematic measurement error (bias = mean measured value – true value)
Precision	Closeness of agreement between replicate measurements on the same or similar objects under specified conditions.	Random measurement error (standard deviation, coefficient of variation)
Reproducibility	Precision under a set of conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects.	Random measurement error (standard deviation, coefficient of variation)
Repeatability	Precision under conditions that includes the same measurement procedure, same operators, same measuring system, same operation conditions and same location, and replicate measurements on the same or similar objects over a short period of time.	Random measurement error (standard deviation, coefficient of variation)
Measurement error	Measured quantity value minus a quantity value.	
Systematic measurement error	Component of measurement error that in replicate measurements remains constant or varies in a predictable manner.	Estimated by bias
Random measurement error	Component of measurement error that in replicate measurements remains constant or varies in unpredictable manner.	Standard deviation, coefficient of variation
Uncertainty	Parameter characterizing the dispersion of the quantity values being attributed to a measurand based on the information used.	Standard deviation, coefficient of variation
<b>Devices for measurements</b>		
Measuring instrument	Device used for making measurements, alone or in conjunction with one or more supplementary devices (measuring system).	
Transducer	Measuring instrument that provides an output quantity (most often an electric current) having a specific relation with an input quantity (most often a physiologic signal).	
Sensor	Element of a measuring system that is directly affected by a phenomenon carrying a quantity to be measured.	
Detector	Device or substance that indicates the presence of a phenomenon when a threshold value of an associated quantity is exceeded.	
<b>Properties of measuring devices</b>		
Indication	Quantity value provided by a measuring instrument.	
Measuring interval or measuring range	Set of values of quantities of the same kind that can be measured by a given instrument with specified instrumental uncertainty under defined conditions.	
Sensitivity	Quotient of the change in an indication and the corresponding change in a value of a measurand.	
Linearity	Capability of maintaining the sensitivity constant over a large measuring interval.	Correlation, Identity line
Selectivity	Property used with a specified measurement procedure, whereby it provides measured quantity values for one or more measurands such that the values of each measurand are independent of other measurands or other quantities in the phenomenon, body, or substance being investigated.	Correlation
Resolution	Smallest change in a quantity being measured that causes a perceptible change in the corresponding indication.	Least significant change
Discrimination threshold	Largest change in a value of a quantity being measured that causes no detectable change in the corresponding indication.	
Dead band	Maximum interval through which a value of a quantity being measured can be changed in both directions without producing a detectable change in the corresponding indication.	
Stability	Property of a measuring instrument, whereby its metrological properties remains constant in time.	
Instrumental drift	Continuous or incremental change over time of the indication because of change in metrological properties.	
Step response time	Delay between a physiological change and its indication.	
Maximum permissible error or limits of errors	Extreme value of measurement error, with respect to a known reference quantity value, permitted by specifications or regulations for a given measurement, measuring instrument, or measuring system.	
Tolerance	Magnitude of permissible variation of a quantity should not be used to designate the maximum permissible error.	
<b>Measurement standards (Etalons)</b>		
Calibration	Operation that, under specified conditions, establishes a relation between the quantity values provided by measurement standards and corresponding indications (with associated uncertainties) and in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.	
Correction	Compensation for an estimated systematic effect.	
Adjustment of a measuring systems	Set of operations (zero, offset, and span or gain adjustment) carried out on a measuring system so that it provides prescribed indications corresponding to given values of a quantity to be measured.	

Quantity	Dimension	Unit	Symbol
Length	L	meter	m
Mass	M	kilogram	kg
Time	T	second	s
Current	I	ampere	A
Temperature	Θ	kelvin	K
Amount of substance	N	mole	mol
Luminous intensity	J	candela	cd
Force	MLT <sup>-2</sup>	newton	N
Pressure	ML <sup>-1</sup> T <sup>-2</sup>	pascal	Pa
Work or energy	ML <sup>2</sup> T <sup>-2</sup>	joule	J
Power	ML <sup>2</sup> T <sup>-3</sup>	watt	W
Frequency	T <sup>-1</sup>	hertz	Hz



**Figure 1.** Schematic representation of different types of systematic errors. Typically, these lines are these observed when there is an issue in calibration (blue line), in amplification (red line), or when insufficient sensing is corrected by adding a constant (orange line).

**Measurement accuracy** is the closeness of agreement between a measured quantity value and a true quantity value of the measurand. The concept of accuracy is a quality and is not given a numerical value. A measurement is said to be more accurate when it offers a smaller measurement error. Therefore, a measurement error is qualifying a single measurement.

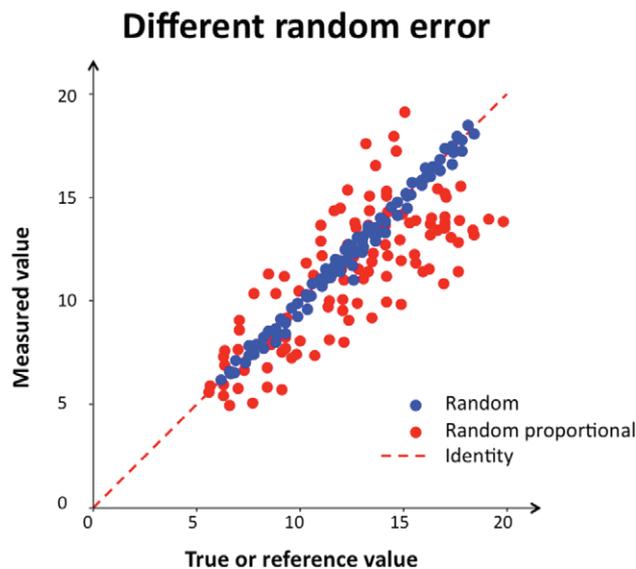
**Measurement trueness** is the closeness of agreement between the average of an infinite number of replicate measured quantity values and the true or a reference quantity value. The concept trueness is a quality and is not given a numerical value. Measurement trueness is inversely related to systematic measurement error but not to random measurement error. Since the mean random error is zero, the bias (average of measured value – reference value) is an estimate of the systematic measurement error. The traditional averaging of (measured value – reference value) is equivalent to the former formula only when there is a unique true (or reference) value. When there are different quantities of the measurand, the bias can be fixed, proportional, or distributed following specific functions (Fig. 1). Since a systematic error cannot be normally/randomly distributed, averaging (measured value – reference value) is therefore

an approximate representation of the averaged bias (systematic error).

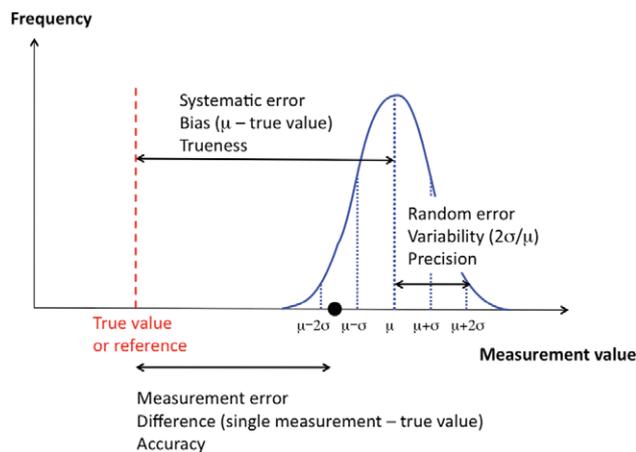
**Measurement precision** is the closeness of agreement between measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions. Measurement precision is related to random measurement error and usually expressed numerically by measures of imprecision, such as standard deviation ( $\sigma$ ), variance ( $\sigma^2$ ), or coefficient of variation ( $\sigma/\text{mean}$ ) and assuming a mean = zero (Fig. 2). Depending on the targeted confidence interval (68%, 95%, or 99%), 1, 2, or 3 $\sigma$  would be considered. The concept of precision is also used to define measurement repeatability and reproducibility. Repeatability is the precision under conditions that include the same measurement procedure, same operators, same measuring system, same operation conditions and same location, and replicate measurements on the same or similar objects over a short period of time. Reproducibility is the precision under a set of conditions that include different locations, operators, measuring systems, and replicate measurements on the same or similar objects.

**Measurement error.** We have seen in “Measurement Accuracy” that it is the difference between a unique measured quantity value and a reference quantity value. The measurement error can be systematic (bias) qualifying the untrueness (see “Measurement trueness”) or random qualifying the imprecision (see “Measurement precision”). When the term “measurement error” is used without further information, it combines systematic and random errors and qualifies the inaccuracy (Fig. 3). Although often used confusingly in the literature, accuracy should not be confounded with trueness and precision. Schematically, accuracy qualifies a single measurement, whereas trueness and precision are necessarily derived from repeated measurements. So, rigorously speaking, averaging different measurement errors is tricky. When it is done and/or when the percentage error is derived (measurement error/true value), generally the absolute values (or the root of the square values) should be used to avoid compensation of positive and negative errors. Otherwise, the averaged measurement error would become equivalent to an averaged systematic error (assuming that the averaged random error is zero) and accuracy and trueness would be confounded. It is important to understand these conceptual differences because they correspond to different mechanisms of errors and are corrected by different technological solutions. Even a review from the same authors published in 2009 was not explicit enough.<sup>4</sup> Basically, systematic errors indicate an issue in the signal processing that can be corrected by improving sensing, amplification, calibration, or by correction (Fig. 1). Random errors most often derive from a pollution of the physiologic signal (noise, interference, etc.) that can be minimized by optimizing the signal filtering or by averaging more measurements. Trueness and precision also have different practical consequences: trueness is necessary in reaching targets, precision is necessary to identify rapid changes of the quantity value.

**Measurement uncertainty** is a parameter characterizing the dispersion of the quantity values being attributed to a measurand based on the information used. This concept is broader than precision including uncertainty due to time drift, definitional uncertainty, and other uncertainties.



**Figure 2.** Schematic representation of 2 types of random errors. Blue plots, small random error ( $2\sigma = 0.6$ ) typically seen when an independent noise is properly filtered. In red plots, the variability in proportion to the measurement value indicating that the random noise is getting into the signal. In these examples, both data sets have correlation slopes on the identity line (black dotted line), indicating no systematic errors (x-axis and y-axis with arbitrary units).



**Figure 3.** Schematic representation of the different types of measurement errors with indication of the formula by which it is derived and the corresponding quality criteria.  $m$  = mean;  $\sigma$  = standard deviation of replicate measurements on the same object. In this figure, the variability is expressed by  $\pm 2\sigma/m$  to fit with a confidence interval of 95%.

There are 2 types of evaluation of the uncertainty of measurements: type A based on statistical analysis of measured values and type B based on other means such as certified references, authoritative published values, or personal experience.<sup>1</sup>

**Devices for Measurements**

A measuring instrument is a device used for making measurements alone or in conjunction with one or more supplementary devices (measuring system). A measuring instrument (gauge) is frequently a transducer, a device that provides an output quantity (most often an electric current) having a specific relation with an input quantity (most often a physiologic signal). The physiologic signal is collected by

a sensor defined as an element of a measuring system that is directly affected by a phenomenon, body, or substance carrying a quantity to be measured or less frequently by a detector defined as a device or substance that indicates the presence of a phenomenon, body, or substance when a threshold value of an associated quantity is exceeded.

**Properties of Measuring Devices**

An indication is the quantity value provided by a measuring instrument. The measuring interval or measuring range is the set of values of quantities of the same kind that can be measured by a given instrument with specified instrumental uncertainty under defined conditions. A measuring instrument/system is characterized by different properties or quality criteria.

**Sensitivity** is the quotient of the change in an indication and the corresponding change in a value of a measurand. The change considered in a value of a quantity being measured must be large compared with the resolution (see “Resolution”). The linearity, which has no international definition, usually refers to the capability of maintaining the sensitivity constant over the measuring interval. In other words, the linearity is also the capability of maintaining the bias (systematic error) constant over the measuring interval.

**Selectivity** is a property used with a specified measurement procedure, whereby it provides measured quantity values for one or more measurands such that the values of each measurand are independent of other measurands or other quantities in the phenomenon, body, or substance being investigated. For instance, a good selectivity implies measuring cardiac output independently to blood pressure in pulse contour systems, or measuring the concentration of 1 volatile anesthetic independent from the presence of other volatile anesthetic in respiratory gases.

**Resolution** is the smallest change in a quantity being measured that causes a perceptible change in the corresponding indication. The concept of resolution is linked to the discrimination threshold, which is the largest change in a value of a quantity being measured that causes no detectable change in the corresponding indication and to the dead band, which is the maximum interval through which a value of a quantity being measured can be changed in both directions without producing a detectable change in the corresponding indication. The least significant change derived as  $2\sqrt{2}$  standard error of the mean is a statistical approach of resolution.

**Stability** is the property of a measuring instrument, whereby its metrological properties remain constant in time. An instrumental drift is the continuous or incremental change over time of the indication because of change in metrological properties.

**Step response time** is the duration between the instant when an input quantity value of a measuring instrument is subjected to an abrupt change between 2 specified constant quantity values and the instant when a corresponding indication settles within specified limits around its final steady value.

**Maximum permissible measurement error or limits of errors.** These terms are defined as the extreme value of measurement error, with respect to a known reference quantity value, permitted by specifications or regulations for a given

measurement, measuring instrument, or measuring system. The term “tolerance,” which is the magnitude of permissible variation of a quantity, should not be used to designate the maximum permissible error. Tolerance includes the true value  $\pm$  the maximum permissible error. For example, if the maximum permissible error in weighting a patient is 1%, the tolerance for a 100 kg man will be  $100 \pm 1$  kg.

Thus, strictly speaking, accuracy, trueness, and precision are qualifying measurements, whereas sensitivity, selectivity, resolution, stability, and step response time are qualifying dynamic indications of devices. Their combination determines the instrumental bias, drift, and maximum permissible error. However, in daily practice, measurements and device indications are linked or confounded and it is not really spurious to apply to indications the terms used for qualifying measurements. Most often, especially in monitoring systems, where a physiologic signal is continuously analyzed, a given device indication (displayed value) results from averaging several measurements. The characteristics of device indications are linked to the precision of measurements. Averaging more measurements or filtering more the signal can improve precision as seen above. This will increase the resolution and allow decreasing the maximum permissible error, but these improvements will be offset by a decrease in sensibility and set response time. Therefore, when evaluating the quality of measurements of a physiologic signal and its relevance for decision-making, the analysis should not be restricted to 1 or 2 properties but must take into consideration the performance for all properties.

**Measurement Standards (Etalons)**

Any measurement requires a measurement standard (Etalon), which is the embodiment of the definition of a given quantity, with stated quantity value and associated measurement uncertainty, used as reference. This definition shows that the uncertainty with the standard contributes to the combined measurement uncertainty since values that result from the measurement process are in reality ratios between the measured values and the measurement standard, expressed in the same units. As an example, a weight of 80 kg means 80 times the value of a measurement standard of  $1 \text{ kg} \pm 3 \text{ }\mu\text{g}$ . The issues related to the practical realization of a measurement standard are not within the scope of this review. For instance, the international measurement standard of a meter is defined as the length of the path travelled by light in vacuum during a time interval of  $1/299,792,458$  of a second. The second is the duration of  $9,192,631,770$  periods of the radiation corresponding to the transition between the 2 hyperfine levels of the ground state of the caesium 133 atom at rest, at a temperature of  $0^\circ\text{K}$ .

A **measurement standard** is the prerequisite of any calibration, which is the operation that establishes a relation between the quantity value provided by a measurement standard and corresponding device indications, with associated uncertainties. A calibration may be expressed by a calibration diagram, calibration curve, or calibration table. It can be an additive or multiplicative correction. Calibration should not be confounded with the adjustment of a measuring system, sometimes called autocalibration, which is the set of operations (zero, offset, and span or

gain adjustment) performed on a measuring system so that it provides prescribed indications corresponding to given values of a quantity to be measured. Therefore in practice, the best performance will be obtained by a first calibration to determine the approximate magnitude of the adjustment needed, then the adjustment, then a final recalibration.

**FOR EACH OBJECTIVE, A SPECIFIC REQUIREMENT**

In medical practice, measurements usually have 3 main objectives: (1) diagnosis, by a timely comparison of the actual quantity value with its normal range, (2) alarm, by continuously checking if the quantity value is inside or outside a predetermined range, and (3) titration, by finely tuning a therapeutic action to bring a quantity value toward a targeted range. For each objective, the requirements would have different priorities.

In addition, the type of quantity to be measured also impacts measurement requirements. In physics, a fundamental distinction is made between intensive and extensive quantities. The magnitude of an intensive quantity is independent of its size (like density or pressure), while the magnitude of an extensive quantity is additive (like mass, volume, or energy). In medicine, there is no consensually recognized quantity distinction. On the basis of the preceding metrological considerations, we suggest making the following 3 distinctions. For each, the most popular statistical approaches are reviewed.

**Parameters**

The terms “variable” and “parameter” are often used interchangeably for the observed and measured instances in the medical literature. To avoid confusion, we recommend making a distinction. In the context of statistics, the word “variable” is used to refer to a measurable quantity value, characteristic, or attribute of an individual or a system (e.g., heart rate, arterial blood pressure, or tidal volume) measured by a sensor. A “parameter” is a value determining the properties of a function or a model describing the general behavior of a variable. In physiologic models, patient-specific values that remain constant over the time of measurements such as age, gender, weight, height, myocardial mass, lung capacity, etc. may be considered parameters because they adjust the physiologic model to the specific patient. Parameter values are collected or measured only once and the unique requirement is trueness. In case of untrueness, all functions or targets dependent of this parameter will be inappropriately estimated. As an example, in basal metabolic conditions, a cardiac output of  $2.3 \text{ L/min/m}^2$  can be considered as acceptable for an 80-year-old man. An error in age, such as 70 years instead of 80 years, may lead to an incorrect interpretation of a low cardiac output value.

**Regulated Variables**

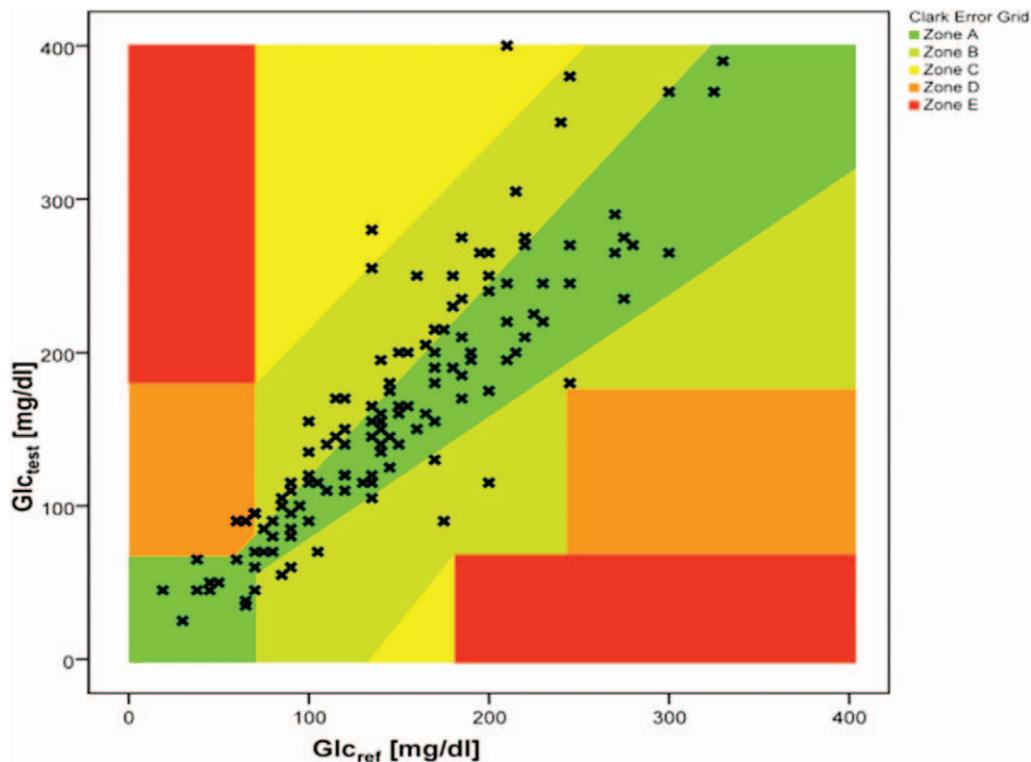
Among the physiologic variables, some of them are maintained in a narrow normal range, for instance, glycemia, hematocrit, core temperature, arterial blood oxygen saturation, etc. To avoid impacting the decision-making (including diagnosis, alarm, and titration), the measurements’ uncertainty should be negligible as compared with physiological

changes. For instance, if the true body core temperature is 38°C, the indications of a thermometer with no bias and precision ( $\pm\sigma$ ) of  $\pm 0.25$  °C will range from 37.25°C to 38.75°C in 99% of the cases ( $\pm 3\sigma$ ). The clinical impact of this imprecision will be very different if applied to a single measurement or to a monitoring system. For a single measurement, depending on the actually displayed values, the patient will be considered febrile or not, which will impact therapy and care. In contrast, if the thermometer is averaging several values continuously collected, the precision of the displayed value is no longer given by the standard deviation ( $\sigma$ ) but by the standard error of the mean ( $\sigma/\sqrt{N}$ ), where  $N$  is the number of averaged independent measurements (an estimation assuming that  $N$  is much smaller than the population sample size, therefore subject to eventual correction). The standard error of the mean is therefore  $\sigma/10$  after 100 measurements. Then, if we assume an independent measurement every second, the value averaged by the system and refreshed every 100 seconds will range from 37.925°C to 38.075°C, which can be considered as a clinically good performance. In reality, more measurements must be averaged than done in this schematic example since successive measurements of monitoring systems are often auto-correlated, then not fully independent.

When assessing the performance of a new technology in measuring regulated variables, a sufficiently large number of measurements should be compared with a reference method over an adequately wide range of values. We use the example of a new glucose blood analyzer compared with a reference. Although the correlation analysis (as seen in Fig. 1) can be helpful in assessing the systematic error (by comparing the correlation slope with the identity line), trueness and precision are inversely related to the sum of

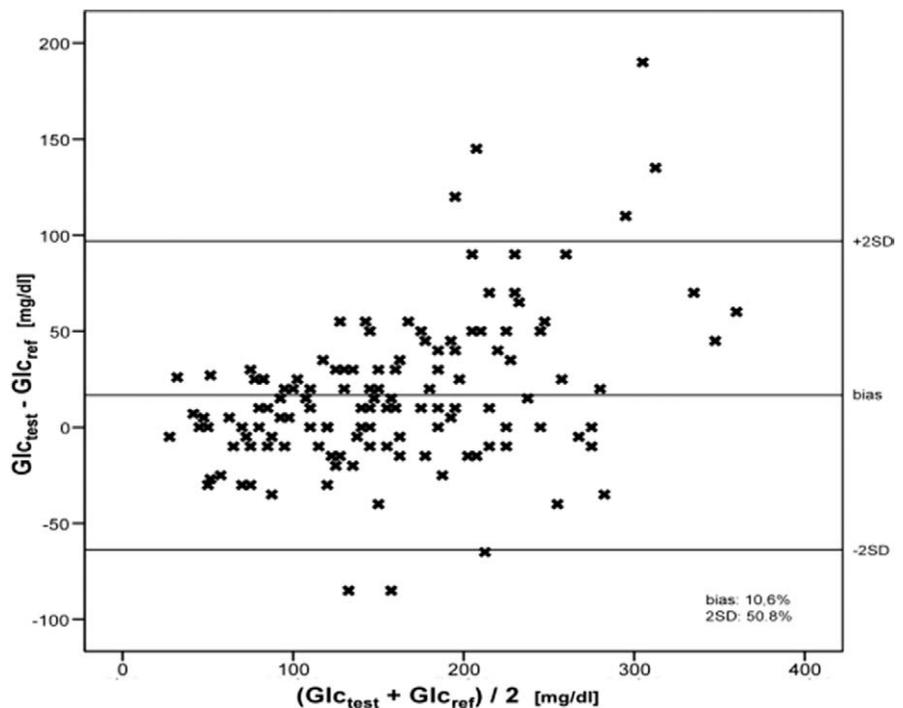
the residuals and are not easy to grasp. The Clark error grid analysis (Fig. 4) helps evaluating whether differences between 2 measurements would lead to differences in therapeutic decisions.<sup>5</sup>

Nevertheless, the most widely used standard is the Bland and Altman<sup>6</sup> representation that figures the variability of measurement errors. The  $\pm 2\sigma$  has been called the 95% limits of agreement (Fig. 5). When repeated measurements estimate the same quantity value, the mean error estimates the bias (systematic error) and the variability estimates the precision (random error). When different measurements are compared with their own reference of quantity values and pooled together, the mean error estimates an approximate “mean clinical bias” (see “Measurement Trueness”) and the variability combines quite confusingly: (1) the variability of systematic errors (which can be null in case of inappropriate zeroing, for example, proportional in case of inadequate calibration, or following complex, but not random functions (Fig. 1) and (2) the different types of random errors, which depend of the conditions of data acquisition. When investigating different patients, in different settings, different units, and using different devices, this overall precision is representative of the reproducibility and not indicative of the intrinsic precision of the studied technology (repeatability), which requires investigating the same or similar objects under specified conditions. That is why it has been suggested to assess the precision of a specific device as the variability of the measurement errors ( $\pm 2\sigma$ ) around a fixed quantity value, a reference trend line obtained by a reference method or a patient in steady state.<sup>4,7</sup> In addition, the Bland and Altman representation has been created to indicate if a method can be used interchangeably with another



**Figure 4.** Clark error grid for insulin dosing. Except for 2 measurements in zones C and D each, agreement between the 2 methods is acceptable for diabetes care in this example. (Reproduced from Ref. 20, with permission.)

**Figure 5. Bland-Altman plots.** Mean difference =  $17 \pm 40$  mg/dL. Limits of agreements ( $2\sigma$ ) of the measurement error/averaged value = 51%. (Reproduced from Ref. 20, with permission.)



one, when none of them can be considered as a reference. Therefore, when a studied method has to be compared with a widely accepted reference, the interpatient accuracy can be estimated by the frequency distribution of the absolute values of individual measurement errors.<sup>4,7</sup> In the example shown in Figure 5, the mean difference between the 2 devices is  $17 \pm 40$  mg/dL so that the estimated mean systematic error (mean estimated bias) is 17 mg/dL (11%) and the estimated combined (systematic + random) variability ( $\pm 2\sigma$ ) is 80 mg/dL (51%). Deriving the mean absolute measurement error gives  $30 \pm 32$  mg/dL with a percentage error =  $22 \pm 22\%$ . The frequency distribution indicates that 41% of the measurements have more than 20% error.

Moreover, requirements for measurement performance may depend on the clinical setting. Since the therapeutic window for tight glucose control is narrower for critical care than for ambulatory diabetes care, the Clark error grid as seen in Figure 4 is not sufficient for the assessment of glucose monitoring technologies. For example, Ellmerer et al<sup>8</sup> have proposed a modification of the Clark error grid (Fig. 6) with a higher resolution and narrower boundaries that are more adequate for tight glucose control and are based on established protocols.<sup>9,10</sup>

### Adaptive Variables

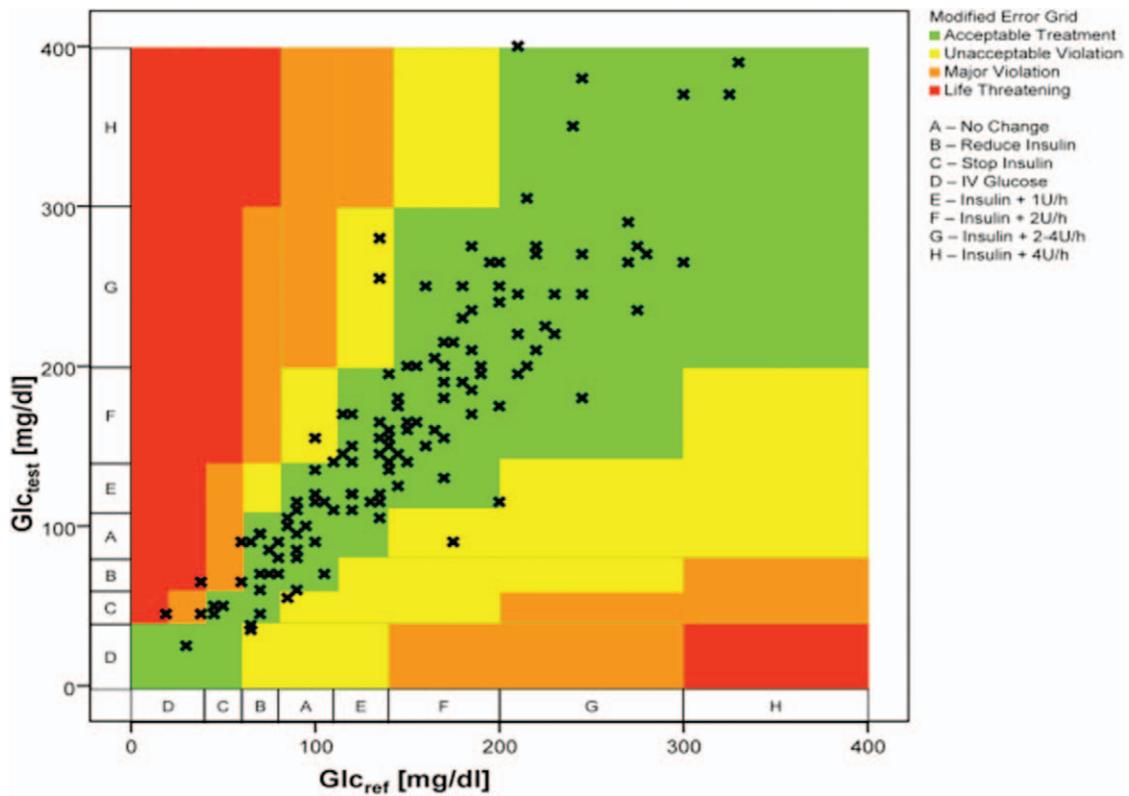
The last category is represented by adaptive variables. They do have statistical normal range, but the adequate value in a specific pathologic situation may be out of normal range since they are regulated to maintain homeostatic equilibriums. Any change in 1 variable of the equilibrium must be compensated by others. This is the case, for example, for cardiac output, blood pressure, mixed venous oxygen saturation, minute ventilation, etc. Their usual range is large, overlapping normal and adaptive fluctuations. Therefore, trueness is a secondary objective since it is a priori unknown if a given value is adequate or not for maintaining the homeostasis. In contrast, any change in such variables and the magnitude of

the change are always indicative of a physiologic event in the regulation loop. At best, these variables should be continuously monitored and the measurement properties of critical importance are therefore step time response and precision so that any change in the indication is real and represents a change in the quantity value.<sup>11</sup> In both situations, precision must be good over different conditions (reproducibility) and over the requested measuring interval (stability, linearity). The combination of precision and linearity can be illustrated by plotting the changes of the test device indications following therapeutic challenges versus the changes in the reference values. Alternatively, the polar plots described by Critchley et al<sup>12</sup> may be used with distance from the center as mean change and the angle with the horizontal axis, as agreement.<sup>13</sup>

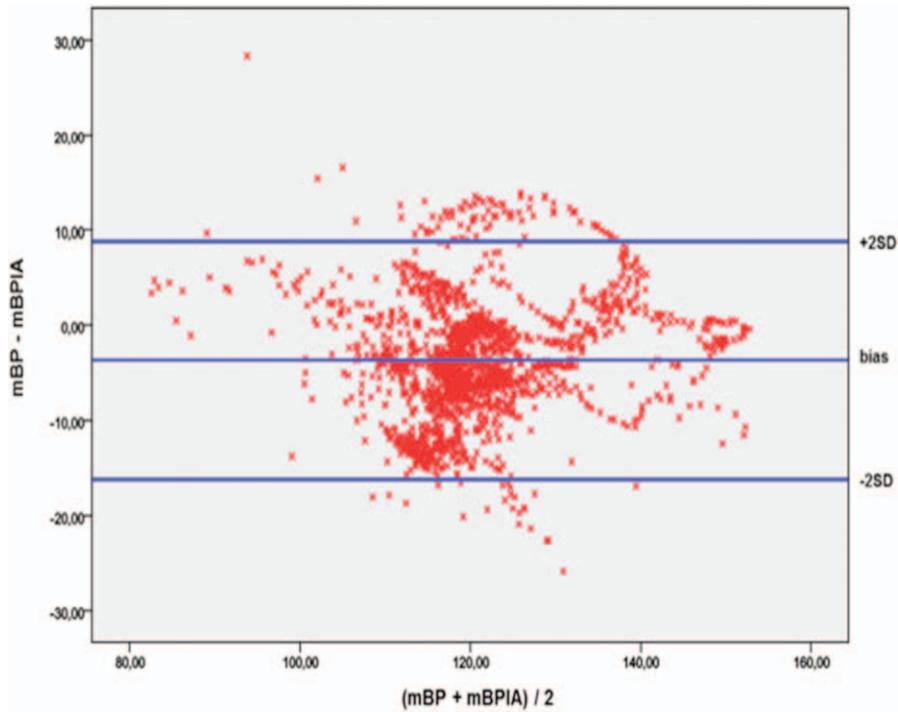
Relevant titration of life-sustaining therapies also depends mostly on precision and step response time of measurements if relative changes are used. For example, catecholamines are commonly titrated against changes in lactate, blood pressure, flow, heart rate, and urinary output. However, trueness is also required when specific threshold values are considered. This is of special importance when these measurements trigger automatic therapeutic guidance and/or closed loop systems.<sup>14-18</sup>

### UNCERTAINTY DUE TO THE REFERENCE

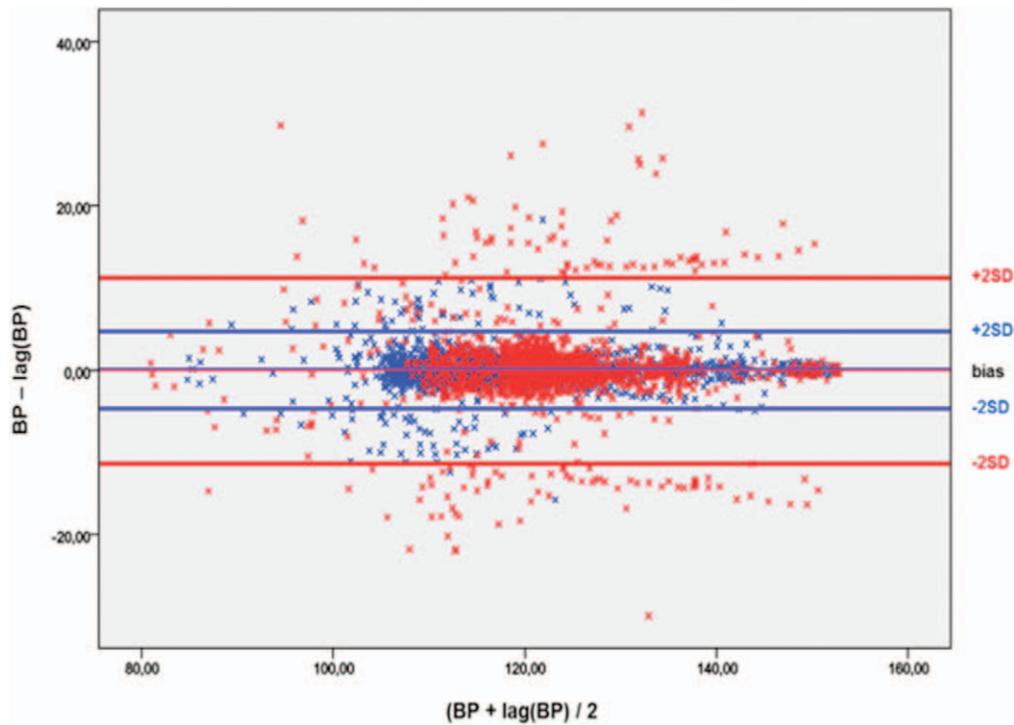
A new measurement method is usually validated by comparison to an existing, established reference method, even a gold standard. We have seen in the Measurements Standards section that even etalons have an uncertainty. In clinical practice, the uncertainty of the reference contributes to the measurement error and is often too far off to be negligible. Figure 7 gives an example of a Bland-Altman plot where a new, noninvasive continuous blood pressure monitor is not interchangeable with the invasive blood pressure, that is, the assumed gold



**Figure 6.** Modified error grid for the same data as in Figure 4. This error grid shows that the test device would have resulted in unacceptable, serious protocol violation, or even life-threatening therapy decisions in 10 cases, which may be considered unacceptable for clinical use in the intensive care unit. (Reproduced from Ref. 20, with permission.)



**Figure 7.** Bland-Altman plot of mean arterial pressure measured by a noninvasive method (mBP) and the invasive method (mBPIA). Difference =  $-3.7 \pm 6.2$  mm Hg,  $\pm 10.3\%$ .



**Figure 8.** Same data as in Figure 6. Overlaid Bland-Altman plots of repeated measurements for each measurement technique (blue mBP; red = mBPIA). mBP: CR = 6.53, mBPIA: CR = 15.65

standard. Nonphysiological variability of the reference may markedly influence the observed interdevice differences. In this situation, the Bland-Altman plot also is useful to assess selectively the repeatability of each method by comparing 2 measurements on the same subject using the same method on a series of subjects.<sup>19</sup> Since a unique method is used, there is no systematic error and the mean difference of repeated measurements should be zero. Therefore, the Coefficient of Repeatability (CR) can be calculated as 2 times the standard deviation of the differences between the 2 measurements ( $d_2$  and  $d_1$ ) of the same quantity:

$$CR = 1.96 \times \sqrt{\frac{\sum (d_2 - d_1)^2}{n - 1}}$$

Figure 8 shows the same data as in Figure 7, but now assessing the repeatability of each measurement method independently. The results from this example actually indicate that the invasive blood pressure monitoring contributes more to the intermethod variability than the new studied method of noninvasive blood pressure monitoring.

### CONCLUSIONS

A basic understanding of metrology is essential for the daily practice of medicine. In intensive care, clinical decision-making is often determined by measurements of physiological and other variables to an extent unrivalled by most other medical specialties. Therapeutic success and ultimately outcomes in the critically ill depend on the correct interpretation of such measurements. Therefore, physicians should be aware of metrological concepts and understand the limitations and constraints. In addition, the interactions

between medicine and other scientific disciplines mandate a common language. Since international consensus definitions exist, we must use them and promote them in the medical research and literature. ■■

### DISCLOSURES

**Name:** Pierre Squara, MD.

**Contribution:** This author helped write the manuscript.

**Attestation:** Pierre Squara approved the final manuscript.

**Name:** Michael Imhoff, MD, PhD.

**Contribution:** This author helped write the manuscript.

**Attestation:** Michael Imhoff approved the final manuscript.

**Name:** Maurizio Cecconi, MD, PhD.

**Contribution:** This author helped write the manuscript.

**Attestation:** Maurizio Cecconi approved the final manuscript.

**This manuscript was handled by:** Maxime Cannesson, MD, PhD.

### REFERENCES

1. Bureau International des Poids et Mesures. International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM). Paris: Joint Committee for Guides in Metrology, 2012
2. Wachtel RE, Dexter F. Difficulties and challenges associated with literature searches in operating room management, complete with recommendations. *Anesth Analg* 2013;117:1460–79
3. BIMP. International System of Units, 2013
4. Squara P, Cecconi M, Rhodes A, Singer M, Chiche JD. Tracking changes in cardiac output: methodological considerations for the validation of monitoring devices. *Intensive Care Med* 2009;35:1801–8
5. Cox DJ, Clarke WL, Gonder-Frederick L, Pohl S, Hoover C, Snyder A, Zimelman L, Carter WR, Bobbitt S, Pennebaker J. Accuracy of perceiving blood glucose in IDDM. *Diabetes Care* 1985;8:529–36
6. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–10

7. Squara P, Denjean D, Estagnasie P, Brusset A, Dib JC, Dubois C. Noninvasive cardiac output monitoring (NICOM): a clinical validation. *Intensive Care Med* 2007;33:1191–4
8. Ellmerer M, Haluzik M, Blaha J, Kremen J, Svacina S, Toller W, Mader J, Schaupp L, Plank J, Pieber T. Clinical evaluation of alternative-site glucose measurements in patients after major cardiac surgery. *Diabetes Care* 2006;29:1275–81
9. Plank J, Blaha J, Cordingley J, Wilinska ME, Chassin LJ, Morgan C, Squire S, Haluzik M, Kremen J, Svacina S, Toller W, Plasnik A, Ellmerer M, Hovorka R, Pieber TR. Multicentric, randomized, controlled trial to evaluate blood glucose control by the model predictive control algorithm versus routine glucose management protocols in intensive care unit patients. *Diabetes Care* 2006;29:271–6
10. van den Berghe G, Wouters P, Weekers F, Verwaest C, Bruyninckx F, Schetz M, Vlasselaers D, Ferdinande P, Lauwers P, Bouillon R. Intensive insulin therapy in critically ill patients. *N Engl J Med* 2001;345:1359–67
11. Cecconi M, Rhodes A, Poloniecki J, Della Rocca G, Grounds RM. Bench-to-bedside review: the importance of the precision of the reference technique in method comparison studies—with specific reference to the measurement of cardiac output. *Crit Care* 2009;13:201
12. Critchley LA, Yang XX, Lee A. Assessment of trending ability of cardiac output monitors by polar plot methodology. *J Cardiothorac Vasc Anesth* 2011;25:536–46
13. Kiefer N, Hofer CK, Marx G, Geisen M, Giraud R, Siegenthaler N, Hoeft A, Bendjelid K, Rex S. Clinical validation of a new thermodilution system for the assessment of cardiac output and volumetric parameters. *Crit Care* 2012;16:R98
14. Squara P, Fourquet E, Jacquet L, Broccard A, Uhlig T, Rhodes A, Bakker J, Perret C. A computer program for interpreting pulmonary artery catheterization data: results of the European HEMODYN Resident Study. *Intensive Care Med* 2003;29:735–41
15. Rose L, Presneill JJ, Johnston L, Cade JF. A randomised, controlled trial of conventional versus automated weaning from mechanical ventilation using SmartCare/PS. *Intensive Care Med* 2008;34:1788–95
16. Arnal JM, Wysocki M, Novotni D, Demory D, Lopez R, Donati S, Granier I, Corno G, Durand-Gasselien J. Safety and efficacy of a fully closed-loop control ventilation (IntelliVent-ASV®) in sedated ICU patients with acute respiratory failure: a prospective randomized crossover study. *Intensive Care Med* 2012;38:781–7
17. Le Guen M, Grassin-Delyle S, Cornet C, Genty A, Chazot T, Dardelle D, Liu N, Dreyfus JF, Mazoit JX, Devillier P, Alvarez JC, Sessler DI, Fischler M. Comparison of the potency of different propofol formulations: a randomized, double-blind trial using closed-loop administration. *Anesthesiology* 2014;120:355–64
18. Okabayashi T, Shima Y. Are closed-loop systems for intensive insulin therapy ready for prime time in the ICU? *Curr Opin Clin Nutr Metab Care* 2014;17:190–9
19. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135–60
20. Imhoff M. Glucose monitoring technologies for the intensive care unit. *Care Critically Ill* 2007;23:65–76