Journal of Critical Care (2013) **28**, 62–68

ELSEVIER

Journal of Critical Care

Interobserver agreement rate of the spontaneous breathing trial☆

Karina Reis Kappáz Cappati RT[a], **Rodrigo Marques Tonella RT**[b], **Aline Santos Damascena**[c], **Carlos Alberto de Bragança Pereira**[d], **Pedro Caruso MD**[a,e,*]

[a]*UTI - Hospital A C Camargo, São Paulo, Brazil*
[b]*Unidade de Terapia Intensiva, Hospital de Clínicas da Universidade de Campinas, São Paulo, Brazil*
[c]*Bioinformatics Department, Hospital AC Camargo*
[d]*Instituto de Matemática e Estatística, Universidade de São Paulo*
[e]*Respiratory ICU, Pulmonary Department, Heart Institute (InCor), University of São Paulo Medical School, São Paulo, Brazil*

Keywords:
Mechanical ventilation;
Ventilator weaning;
Diagnostic tests;
Interobserver variation;
Bayesian analysis

Abstract

Purpose: During the mechanical ventilation weaning process, the spontaneous breathing trial (SBT) is the confirmatory test of patients' capability to breathe unassisted. However, the SBT interobserver agreement rate (its reliability) is unknown, and our objective was to evaluate it.

Materials and Methods: This is a prospective, multicentric and observational study. Patients were included when the SBT criteria were fulfilled. Two physicians and 2 respiratory therapists (RTs) rated each SBT. The SBT interobserver agreement was measured using κ statistic and also the percentage of agreement with its 95% credible interval (CrI) calculated by a Bayesian inference.

Results: Ninety-three distinct physicians and 91 distinct RTs rated 130 SBTs. The κ coefficient was 0.46 for physicians and 0.57 for RT, indicating a moderate interobserver agreement rate. The percentage of agreement was 87.7% between physicians (95% CrI, 81.0%-92.3%) and 86.2% between RT (95% CrI, 79.2%-91.1%). The physicians' and RT' percentage of agreement were not statistically different ($P = .71$).

Conclusions: The SBT interobserver agreement rate is only moderate for physicians and RT. The percentage of agreement between 2 different SBT observers is 79.2% to 92.3%. Therefore, a relevant percentage of patients will have different extubation decisions depending on the SBT observer.

☆ Conflict of interest: Karina Kappaz received a personal grant from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), a federal unprofitable funding agency, to participate in the study. Karina Reis Kappáz does not have any conflict of interest to disclosure. Rodrigo Marques Tonella does not have any conflict of interest to disclosure. Aline Santos Damascena does not have any conflict of interest to disclosure. Carlos Alberto de Bragança Pereira does not have any conflict of interest to disclosure. Pedro Caruso does not have any conflict of interest to disclosure.

* Corresponding author. Tel.: +55 11 2661 5695; fax: +55 11 2661 5695.
E-mail address: pedro.caruso@hc.fm.usp.br (P. Caruso).

0883-9441/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jcrc.2012.06.013

1. Introduction

Mechanical ventilation is indispensable to many critically ill patients, but due to its complications as ventilator-induced lung injury, ventilator-associated diaphragmatic dysfunction, and ventilator-associated pneumonia and due to its discomfort [1,2], it should be removed without delay. The mechanical ventilation weaning begins with the recognition that is possible to decrease the patient's dependency on the mechanical ventilation. If the weaning runs successfully, the final step will be a confirmatory test [3]. This confirmatory test is a diagnostic test because it intends to diagnose if the patient is able or not to reassume the unassisted breathing; in other words, it intends to determine the likelihood of a successful extubation. The spontaneous breathing trial (SBT) is the confirmatory test of the weaning process and has a large and increasing use worldwide [4].

As for any diagnostic test, SBT should have a high accuracy and reliability or at least these characteristics should be known [5,6]. Unfortunately, this is not the case of the SBT because, although some information about its accuracy is known [3,7], its reliability is unknown. The reliability of an observational test, as the SBT, is determined by the agreement between different observers. A low interobserver agreement rate means a low test reliability and vice versa.

Failure of the SBT is defined by objective indices as tachypnea, tachycardia, and hypertension and also by subjective indices, such as agitation, distress, depressed mental status, and evidence of increasing effort [3]. If the SBT judgment was based only on the objective indices, it would be expected a total interobserver agreement. However, the SBT judgment is also based on subjective indices that probably decrease the SBT interobserver agreement (reliability) to an unknown amount.

The objective of the present study is to measure the SBT interobserver agreement of physicians and respiratory therapists (RTs).

2. Methods

We designed an observational, prospective, and multicentric study that was conducted in 6 intensive care units (ICUs) of 3 teaching hospitals from August 2009 to May 2010 (Appendix E1). The institutional review board of the 3 hospitals approved the study. All participants gave written informed consent.

2.1. Weaning protocol

The 3 centers have protocols to guide the weaning process. Patients were daily screened, and a SBT was indicated when the disease for which the patient was intubated had improved, the oxygenation was adequate (PaO_2/fraction of inspired oxygen ≥ 150 with fraction of inspired oxygen $\leq 40\%$ to 50% and positive end-expiratory pressure [PEEP] ≤ 5 to 8 cm H_2O), the cardiovascular status was stable (vasoactive drugs absent; heart rate < 120 beats per minute without acute arrhythmia), the patient was awake with coughing during endotracheal suctioning, core temperature was less than 38.0°C and hemoglobin level greater than 7.0 g/dL or at a level considered acceptable by the primary physician. The SBT was performed using a T-tube with supplemental oxygen or using pressure support ventilation with a level 5 cm H_2O or less and PEEP 5 cm H_2O or less, both lasting up 30 minutes. If the SBT was tolerated, patients were extubated after 30 minutes. Failure of the SBT was decreed if one of the objective or subjective criteria were persistently present: respiratory rate 35 breaths per minute or higher, oxygen saturation as measured by pulse oximetry 88% or less, heart rate 140 beats per minute or higher or acute arrhythmia, systolic blood pressure 90 mm Hg or less or 180 mm Hg or higher, facial signs of distress, increased accessory muscle activity, intense agitation, or depressed level of consciousness.

2.2. Observers and patients

Two physicians and 2 RTs of the regular ICU staff rated each SBT. We divided physicians and RT in 2 categories. The primary physician and primary RT (primary professionals) were those caring for the patient on the day of the study, and the nonprimary physician and nonprimary RT (nonprimary professionals) were those of the regular ICU staff but not caring for the patient.

Patients older than 18 years and mechanically ventilated for more than 48 hours were consecutively included when the primary physician ordered a SBT. Only 1 SBT was observed per patient that was included only once in the same hospital admission. Patients did not receive any intervention due to the study.

2.3. Study design

After the primary physician had ordered the SBT for the primary RT, a researcher who supervised the study but remained uninvolved in management decisions invited a nonprimary ICU physician and a nonprimary RT to observe and rate the SBT as failure or success. The observers were instructed to consider the success of the SBT only when they considered that the patients were able to be extubated. The researcher summarized the patients' clinical conditions and examinations to the nonprimary professionals. The same researcher observed the SBT and requested that the 4 observers did not externalize their impressions during the SBT. At the moment that the primary physician declared the SBT ending, the verdicts of the 4 observers were covertly pointed in a chart. Immediately before and at the last minute of the SBT, we recorded the ventilatory and hemodynamic parameters of the patients (Fig. 1). The clinical decisions and management of the SBT were at the discretion of primary physician.

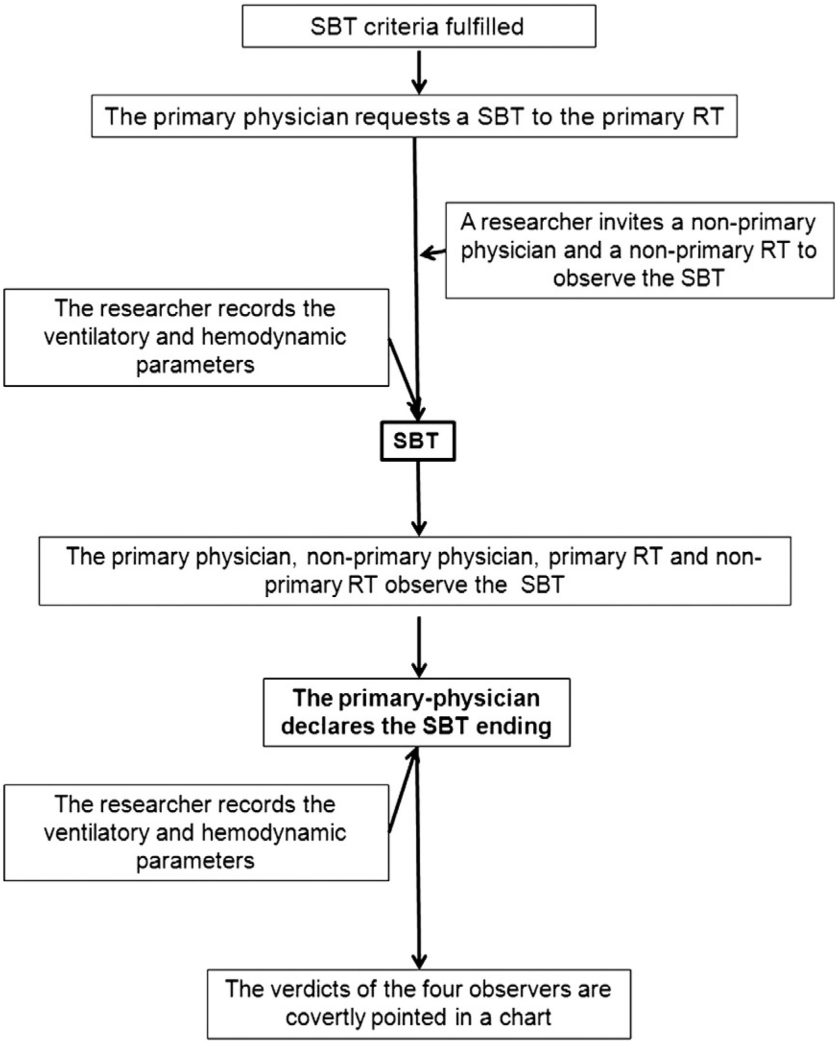
SBT criteria fulfilled
The primary physician requests a SBT to the primary RT
A researcher invites a non-primary physician and a non-primary RT to observe the SBT
The researcher records the ventilatory and hemodynamic parameters
SBT
The primary physician, non-primary physician, primary RT and non-primary RT observe the SBT
The primary-physician declares the SBT ending
The researcher records the ventilatory and hemodynamic parameters
The verdicts of the four observers are covertly pointed in a chart

Fig. 1 Flow chart of the study.

2.4. Statistical analyses

We measured the SBT interobserver agreement of 2 different pairs: physicians and RT.

The interobserver agreement was measured using the percentage of agreement between observers with its 95% credible interval (CrI) calculated using a Bayesian inference. The interobserver agreement was also measured using κ statistic.

2.5. Statistic

We calculated the Cohen κ coefficient [8] and classified the agreement strength as previously proposed [9].

2.6. Bayesian inference

First, we calculated the percentage of agreement (see "Methods" section for further details). However, the percentage of agreement is influenced by chance, so we quantified the uncertainty of the measured percentage of agreement using a Bayesian inference, the normalized likelihood function [10-14].

Likelihood is the expression of a belief in a parameter value given some observed outcome. In frequentist inference, a given value of the parameter returns the density of the different frequencies, and in Bayesian inference, a given frequency returns the density of the different parameters (in the present study, the percentage of agreement).

Briefly, we recorded the frequencies of agreements and disagreements of the SBT verdict in our sample. With these frequencies, we calculated the posterior probability distribution as follow:

$$f(\theta|x) = \frac{\theta^{x+a-1}(1-\theta)^{y+b-1}}{B(x+a;y+b)}$$

where, θ is the unknown parameter of interest, x is the observed frequency (known) and $y = n - x$, for n denoting the sample size. In the present study, θ is the unknown percentage of agreement; x (y) is the frequency of agreement

Table 1 Ventilatory and hemodynamic patients' characteristics immediately before and at the end of the SBT

	Before	End
Pressure support in cm H_2O	10 (7-10)	
PEEP in cm H_2O	6 (5-8)	
Fraction of O_2 in %	30 (30-40)	
Tidal volume in mL	496 (165)	481 (172)
Respiratory rate per minute	20 (6)	23 (7)
Rapid shallow breathing index	46 (24)	58 (44)
Systolic arterial pressure in mm Hg	138 (24)	141 (23)
Diastolic arterial pressure in mm Hg	74 (14)	76 (15)
Heart rate per minute	93 (17)	98 (20)
Spo_2 in %	97 (3)	96 (3)
pH	7.44 (0.06)	7.40 (0.07)
Pao_2 in mm Hg	98 (30)	83 (25)
$Paco_2$ in mm Hg	40 (9)	43 (12)

Data are expressed as mean (and SD in parentheses) or median (and 25%-75% interquartile range in parentheses). Spo_2, pulse oximetry (peripheral saturation of O_2).

(disagreement) observed; *a* and *b* are real numbers that represent our prior information about the population parameters (prior sample). If there is no prior information, we consider the normalized likelihood as the posterior density, so $a = b = 1$.

Results were expressed as mean and SD or median and 25% to 75% interquartile range. The percentages of agreement were compared using the χ^2 test. The statistical software SPSS 19.0 (IBM Corporation, Somers, NY) and R (R Development Core Team, Vienna, Austria) were used.

3. Results

3.1. Patients' and observers' characteristics

One hundred thirty SBTs were observed and rated. The numbers of SBT per center were 60 (46.2%), 45 (34.6%), and 25 (19.2%). The mechanical ventilation length of stay before the SBT was 5.6 days (SD, 3.4). The diagnoses at ICU admission and intubation causes are depicted in Table E1. One hundred twelve SBTs (86.2%) were the first SBT attempt; 85 SBTs were performed with low level of pressure support; and 45, through a t-tube. The SBT duration median was 38 minutes (interquartile range, 30-120). The reintubation rate was 11.5%. The ventilatory and hemodynamic parameters of the patients are depicted in Table 1.

Table 2 Spontaneous breathing trial failures according to the category of the observer

	n = 130 (%)
Physicians	
Primary physician	15 (11.5%) [a,b]
Nonprimary physician	19 (14.6%) [c,d]
RT	
Primary RT	25 (19.2%) [e]
Nonprimary RT	27 (20.8) [f]

(%) in the percentage of SBT observed.
[a] $P < .01$ primary physician × nonprimary physician.
[b] $P < .01$ primary physician × primary RT.
[c] $P < .01$ nonprimary physician × nonprimary RT.
[d] $P < .01$ nonprimary physician × primary RT.
[e] $P < .01$ primary RT × nonprimary RT.
[f] $P < .01$ primary physician × nonprimary RT.

Ninety-three distinct physicians and 91 distinct RTs participated in the study (Table E2).

3.2. Differences in the SBT verdicts

The percentage of SBT failures varied according to group that rated it. The RT considered failure of the SBT more frequently than their physician peers. The nonprimary professionals considered failure of the SBT more frequently than their primary peers (Table 2).

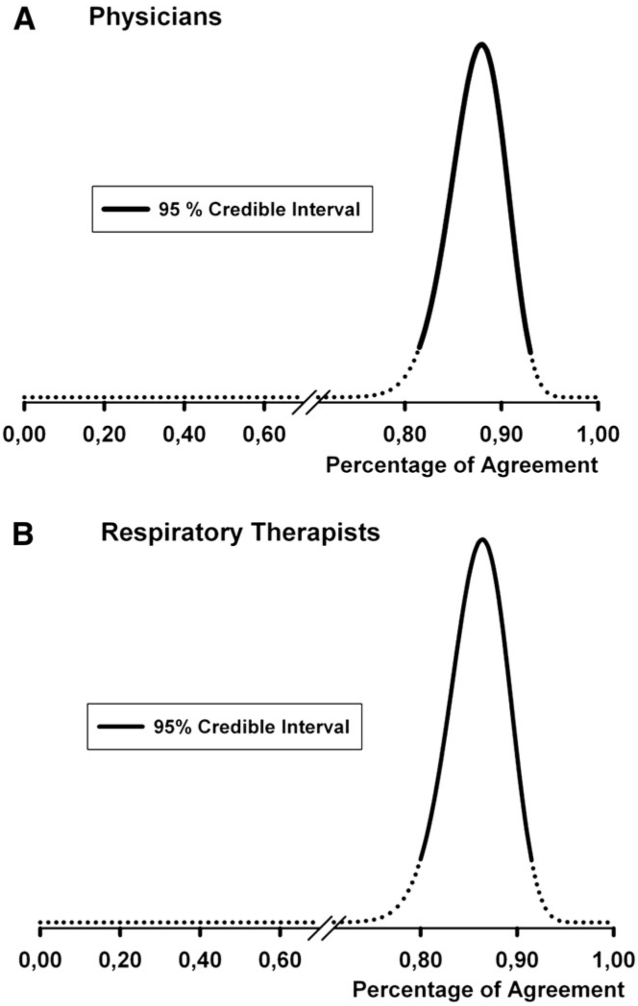

Fig. 2 Percentage of agreement and the 95% CrI of the SBT between physicians (A) and respiratory therapists (B). The values in the y-axis are likelihood.

3.3. Spontaneous breathing trial interobserver agreement

The κ coefficient was 0.46 for physicians and 0.57 for RT, which is a moderate interobserver agreement for both categories of observers.

The physicians agreed in 114 SBTs and disagreed in 16 resulting in a percentage of agreement of 87.7% with a 95% CrI of 81.0% to 92.3%. The RT agreed in 112 SBT and disagreed in 18 resulting in a percentage of agreement of 86.2% with a 95% CrI of 79.2% to 91.1% (Fig. 2). The percentages of agreement of physicians and RT were not statistically different (87.7% × 86.2%; $P = .71$).

Considering the 16 SBTs, which the physicians disagreed, in 13 (81%), the objective indices that defined a SBT failure were absent. In the other 3 SBTs, the objective index of SBT failure was tachypnea. Considering the 18 SBTs, which the RT disagreed, in 16 (89%), the objective indices of SBT failure were absent. In the other 2, the objective indices of SBT were systolic hypertension or tachycardia.

4. Discussion

The SBT is a pivotal diagnostic test in the intensive care medicine, but its interobserver agreement (reliability) was unknown hitherto. We showed that the interobserver agreement of a SBT is moderate and that the percentage of agreement between 2 different observers is 79.2% to 92.3%.

4.1. Hypotheses for the imperfect SBT interobserver agreement

We hypothesized that there are 2 reasons for the imperfect agreement between 2 observers of the same SBT. The first reason is the high amount of subjective indices (agitation, distress, depressed mental status, diaphoresis, and evidence of increasing effort) presented in the judgment of a SBT [3]. The second reason is that the judgment of a SBT is a decision making, and it is subject to the factors that influence any decision making, medical [15] or not [16].

We did not ask to the observers the reason why they considered that the SBT failed, but at the beginning and ending of the SBT, we recorded the objective indices involved in the SBT judgment. In more than 80% of the SBTs that the observers disagreed, the objective indices that defined a SBT failure were absent, allowing us to infer that the disagreement was predominantly caused by the presence of the subjective indices.

The subjective indices may be signaling a marked increase in respiratory load or may be neuropsychologic signals in a patient with anxiety or delirium but without a prohibitive increase in the respiratory load, therefore, able to be extubated. In a study about the pathophysiologic basis of acute respiratory distress in patients who failed a SBT [17], the authors found that 20% of patients that failed a SBT had the same respiratory mechanics as those that succeed the trial. They hypothesized that a misjudgment of the SBT could be the cause of the failure and that the psychiatry alterations could have contributed to the misjudgment. In fact, the incidence of psychiatry alterations in weaning patients as delirium [18], anxiety, and depression [19] is high. In our population, the 3 patients with objectives indices of SBT failure and that the physicians disagreed about the SBT had delirium in the study day, suggesting that 1 physician considered the objective index of failure as a marker of increased respiratory load and the other considered it as a neuropsychologic signal derived from anxiety or delirium.

4.2. Use of 2 methods to measure the interobserver agreement

The present study evaluated the interobserver agreement with 2 different statistical methods, the κ statistic, and a Bayesian inference based on the normalized likelihood function. Since 1960 [8], κ statistic is the method used in medical studies dedicated to evaluate the interobserver agreement, but it has several limitations [20-24] that impelled the researchers to an alternative [24,25]. The percentage of agreement between the different observers complemented with the estimation of its uncertainty is a more direct and intuitive evaluation than κ and allows the comparison of interobserver agreement among different studies and populations.

The likelihood function demands massive calculations that probably avoided its use in the first studies devoted to measure interobserver agreement, but this is not a limitation nowadays due to high computational capacity. Considering the aforementioned, we propose that the likelihood function be offered as an alternative or complementary method to measure the interobserver agreement.

4.3. Differences in the SBT verdicts

We noticed that the primary professionals were more liberal to consider an extubation than their nonprimary peers. We hypothesized that the higher primary professional's knowledge about the patient's clinical conditions may turn them more confident to decide on extubation, especially to differentiate a marked increase in respiratory load from the neuropsychologic signals of patient with anxiety or delirium. In addition, a deeper knowledge about the cough and muscular strengths may turn the primary professionals more confident to consider an extubation.

We also noticed that the physicians are more liberal to consider an extubation than the RT. We hypothesized that the RTs are more conservative because they may take in account other variables that are less evident to physicians, such as cough intensity and tracheal secretions characteristics. The RT conservativeness about the extubation may be relevant in the ICU where the RTs execute the SBT and declare the result to the physician. In that organizational structure, the RT would report more SBT failures than the physicians would consider, probably leading to an increase in the mechanical ventilation length of stay. This is important to note that our study cannot conclude the reasons for the differences in SBT verdicts between physicians and RT or primary and nonprimary professionals.

We decided to evaluate the interobserver agreement rate of physicians and RT because both are usually involved in the initiation, monitoring, and judgment of a SBT [26]. However, in other hospitals, the role of the RT is performed by nurses.

4.4. Implications of the SBT interobserver agreement

Because the SBT is a diagnostic test that brings about import prognosis and treatment decisions [27,28], it ideally should have a perfect interobserver agreement that would turn it independent from the observer. However, we demonstrated that the interobserver agreement is only moderate and around 10% to 20% of the decisions to extubate the patient would be different with different observers. We believe that the present study reinforces the need for a better confirmatory test of mechanical ventilation weaning process because the SBT reliability is not adequate and the sources of this unreliability that are the subjective indices are not remediable. On the other hand, the disagreement among the ICU team might be positive if it prompts a team discussion regarding the causes of SBT failure, its approach, and treatment.

4.5. Limitations

One limitation of the present study is that the observers could be influenced by the primary physician verdict. A fundamental assumption underlying the measure of the interobserver agreement is that the observers are independent [29]. We tried to preserve the independence of the observations, but due to the nature of the SBT, it is impossible to avoid any hint from the primary physician. However, if some observations were not independent, the κ coefficient would be overestimated that would only reinforce the inadequate interobserver agreement because a true κ and percentage of agreement would be even lower. Another limitation already mentioned is that we did not ask to the observers the reason or reasons why they considered that the SBT failed. However, we recorded the objective indices involved in the SBT judgment at the beginning and ending of the SBT that allows us to notice that in most of the SBT disagreement, the objective indices of SBT failure were absent. This observation leads us to infer that the subjective indices were the cause of the disagreement.

5. Conclusions

The SBT interobserver agreement rate is only moderate for physicians and RT. The percentage of agreement between 2 different SBT observers is 79.2% to 92.3% without difference between physicians' and RT' percentage of agreement. Therefore, a relevant percentage of patients will have different extubation decisions depending on the SBT observer that reinforces the need for a better confirmatory test on the patients' capability to breathe unassisted. The present study also noted that respiratory therapists consider failure of the SBT more frequently than physicians. Finally, we offer an alternative method to measure interobserver agreement that is more intuitive than κ statistic and also allows the comparison of interobserver agreement among different studies and populations.

Supplementary materials related to this article can be found online at http://dx.doi.org/10.1016/j.jcrc.2012.06.013.

Acknowledgments

Financial support: Karina Reis Kappáz Capatti received a scholarship from CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), a governmental funding agency, to participate in the study.

References

[1] Pingleton SK. Complications of acute respiratory failure. Am Rev Respir Dis 1988;137:1463-93.

[2] Rotondi AJ, Chelluri L, Sirio C, et al. Patients' recollections of stressful experiences while receiving prolonged mechanical ventilation in an intensive care unit. Crit Care Med 2002;30:746-52.

[3] Boles JM, Bion J, Connors A, et al. Weaning from mechanical ventilation. Eur Respir J 2007;29:1033-56.

[4] Esteban A, Ferguson ND, Meade MO, et al. Evolution of mechanical ventilation in response to clinical research. Am J Respir Crit Care Med 2008;177:170-7.

[5] Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.

[6] Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645-51.

[7] Tobin MJ, Jubran A. Variable performance of weaning-predictor tests: role of Bayes' theorem and spectrum and test-referral bias. Intensive Care Med 2006;32:2002-12.

[8] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.

[9] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

[10] Blume JD. Likelihood methods for measuring statistical evidence. Stat Med 2002;21:2563-99.

[11] Pawitan Y. In all likelihood: statistical modelling and inference using likelihood (ed First Edition). Oxford: Oxford University Press; 2001.

[12] Pereira BB, Pereira CAB. A likelihood approach to diagnostic tests in clinical medicine. REVSTAT – Stat J 2005;3:77-98.

[13] Severini TA. Likelihood methods in statistics (ed First Editon). Oxford University Press: Oxford; 2000.

[14] Sproutt DA. Statistical inference in science (ed First Edition). New York: Spring-Verlag; 2000.

[15] Sox HC, Blatt MA, Higgins MC, et al. Medical decision making (ed Second Edition). Boston: Butterworth-Heinerman; 2006.

[16] Lauwereyns J. The anatomy of bias: how neural circuits weigh the options (ed First Edition). Cambridge: MIT Press; 2010.

[17] Jubran A, Tobin MJ. Pathophysiologic basis of acute respiratory distress in patients who fail a trial of weaning from mechanical ventilation [see comments]. Am J Respir Crit Care Med 1997;155: 906-15.

[18] Girard TD, Kress JP, Fuchs BD, et al. Efficacy and safety of a paired sedation and ventilator weaning protocol for mechanically ventilated patients in intensive care (Awakening and Breathing Controlled trial): a randomised controlled trial. Lancet 2008;371:126-34.

[19] Jubran A, Lawm G, Kelly J, et al. Depressive disorders during weaning from prolonged mechanical ventilation. Intensive Care Med 2010;36: 828-35.

[20] Blackman NJ, Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. Stat Med 2000;19:723-41.

[21] Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46:423-9.

[22] Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.

[23] Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.

[24] Guggenmoos-Holzmann I. How reliable are chance-corrected measures of agreement? Stat Med 1993;12:2191-205.

[25] Agresti A. Modelling patterns of agreement and disagreement. Stat Methods Med Res 1992;1:201-18.

[26] Ely EW, Baker AM, Dunagan DP, et al. Effect on the duration of mechanical ventilation of identifying patients capable of breathing spontaneously. N Engl J Med 1996;335:1864-9.

[27] Funk GC, Anders S, Breyer MK, et al. Incidence and outcome of weaning from mechanical ventilation according to new categories. Eur Respir J 2010;35:88-94.

[28] Sellares J, Ferrer M, Cano E, et al. Predictors of prolonged weaning and survival during ventilator weaning in a respiratory ICU. Intensive Care Med 2011;37:775-84.

[29] Thompson WD, Walter SD. Kappa and the concept of independent errors. J Clin Epidemiol 1988;41:969-70.