Measuring the Repeatability of Simulated Physiology in Simulators

David Cumin, PhD;
Charlotte Chen, MBChB;
Alan F. Merry, MBChB

Background: In simulation, it may be important in some instances that the physiologic responses to given interventions are substantially repeatable. However, there is no agreed approach to evaluating the repeatability of simulators. We therefore aimed to develop such an approach.
Methods: In repeated simulations, we evaluated the physiologic responses to 7 simple clinical interventions generated by a METI (Medical Education Technologies Incorporated, Sarasota, FL) HPS (Human Patient Simulator) simulator in connected and disconnected states and the screen-based Anesoft Anesthesia Simulator. For a selection of variables, we calculated 3 objective measures of similarity (root mean squared error, median performance error, and median absolute performance error). We also calculated divergence over time and compared 3 preprocessing techniques to reduce the effect of clinically irrelevant phase and frequency differences (simple phase shift, complex phase shift, and dynamic time warping).
Results: We collected data from more than 85 hours of simulation time from 255 simulations. The Anesoft physiologic responses were reproduced exactly in each simulation for all variables and interventions. Minor divergence was present between the time series generated with the METI HPS in the connected state but not in the disconnected state. The METI HPS showed some variation between simulations in the raw data. This was most usefully quantified using median absolute performance error as an indicator and was substantially reduced by preprocessing, particularly with dynamic time warping.
Conclusions: The repeatability of the physiologic response of model-controlled simulators to simple standardized interventions can be evaluated by considering divergence over time and the median absolute performance error of individual or pooled variables, but data should be preprocessed to eliminate irrelevant phase and frequency offsets in some variables. Dynamic time warping is an effective method for this purpose.
(*Sim Healthcare* 10:336–344, 2015)

Key Words: Physiology–modeling, Modeling, Training

One of the advantages of simulation is that scenarios can be repeated in a standardized way for practice, assessment, and research.[1–3] Simulators can be used in different ways for different purposes, and physiologic modeling is not always necessary or even relevant.[4] However, in the context of many anesthesia simulations, participants are asked to manage cases by responding to the physiologic data typically monitored during anesthetics. If these data are generated under the direct control of human operators, potential for bias or error is introduced. This bias is acceptable for many purposes, but it would be ideal if clinically realistic data could be generated automatically by modeling, without the possibility of bias. A completely deterministic approach to modeling without changes to parameters might be less realistic than an approach that generated a degree of variation between simulations comparable with that seen between real patients or within an individual patient at different times. Nevertheless, it would be ideal for the same simulated "patient" to show essentially the same physiologic responses to the same interventions, under the same clinical circumstances,[4] particularly in research that uses physiologic data as outcome measures or in the context of evaluating the performance of anesthesiologists.[5] In essence, physiologic responses to particular interventions should be reproducible to the extent that the appropriate subsequent clinical decisions would be the same on each occasion. A relevant standard would be helpful,[6,7] but we could not find one. One requirement for such a standard would be an agreed approach to assessing the reproducibility of data produced by model-based simulators.

There are many techniques for measuring the similarity of sets of time-series data,[8–13] but few have yet been applied to testing the reproducibility of physiologic data from a simulator. In a conference abstract, Mudumbai et al[14] reported poor repeatability of physiologic data generated by a METI (Medical Education Technologies Incorporated, Sarasota, FL) HPS (Human Patient Simulator) simulator in response to blood losses of between 150 and 3000 mL in one simulator and poor stability over time (D. Gaba, personal communication, 2014). However, their work was based on simulations that were short in duration (2 minutes), they evaluated only heart rate and central venous pressure, and they undertook only rudimentary quantitative analysis of the data. In a letter, Garden et al[15] (2004) reported statistically significant differences between physiologic data generated by 5 different METI HPS simulators in response to standardized interventions (ie, poor

From the Department of Anesthesiology (D.C., A.F.M.), School of Medicine, University of Auckland; and Auckland City Hospital (C.C., A.F.M), Auckland, New Zealand.

Reprints: David Cumin, School of Medicine, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand (e-mail: d.cumin@auckland.ac.nz).
Neither D.C. nor C.C. has any potential conflicts of interest. A.F.M. is a director of Safer Sleep, holds approximately 9% of its shares, and advises the company on the design of its products. D.C. was funded by the Ralphe Eve Seelye Scholarship in Anaesthesiology during this work. C.C. was funded by the Faculty of Medical Health Sciences, University of Auckland.

DOI: 10.1097/SIH.0000000000000098

intersimulator repeatability); they also identified a malfunction in a sixth simulator. With respect to intrasimulator repeatability, they commented that, "There were no statistically significant differences within any one simulator during the repeated iterations of the scenario," without further elaboration.[15] In this study Garden et al used a simple repeated-measures analysis of variance on raw data, single repetitions of a simulation for each simulator, a sampling rate of 0.017 Hz (one data point per minute), and short (8 minutes) durations of their simulations. It is unclear whether the approach used in either of these studies was adequate to deal with evaluating the similarity of complex sets of time-series data.

Therefore, we aimed to develop an approach for assessing the reproducibility of model-controlled[4] physiologic responses to a selection of simple, standardized, clinically relevant interventions. We chose 2 simulators for this purpose, on the basis that each has sophisticated physiologic modeling and is commercially available. These were the screen-based Anesthesia Simulator (Anesoft, Issaquah, USA) and the hardware-based METI HPS.[4] Our interest did not lie in comparing these simulators but rather in identifying a suitable method for assessing this aspect of their performance and, by inference, of the performance of model-based simulators in general. As might be expected, the models of these 2 simulators are proprietary, so it is difficult to evaluate them directly. Instead, one must assess their outputs, the time-series of physiologic data that they produce.

METHODS

Patients and Variables

We investigated the hardware-based METI HPS version 6.5 and the screen-based Anesthesia Simulator version 3.0. Both simulators have model-controlled physiology and can respond in a virtual sense to interventions in the form of entries via a computer interface.

The METI HPS also has a physical mannequin with sensors to measure inhaled gas composition and an optional drug recognition module, which presumably integrate with the mathematical models underlying the physiology.[4]

To evaluate the METI HPS simulator, we selected as a "patient" a 33-year-old, 70-kg male with no known drug allergies or comorbidities ("Standard Man"). The Anesoft simulator provides no patients exactly like "Standard Man," so we selected 3 relatively similar "patients" for use in evaluating this simulator. These were an 18-year-old, 54-kg female with no comorbidities; a nervous 22-year-old, 78-kg male with hay fever and a penicillin allergy; and a 54-year-old, 68-kg male with no comorbidities. We analyzed variables recorded by the internal logs of the respective simulators across repetitions of simulations in the same simulator (discussed later).

Simulations

With the METI HPS, we evaluated the intrasimulator repeatability of physiologic responses to 7 separate interventions, listed in Table 1. We first repeated each of the interventions 10 times with the simulator in the disconnected state (ie, we only used the computer software for the experiments). We then performed these 7 interventions 5 times each in the "connected" state (ie, with the mannequin connected to the software and anesthesia machine). Fewer repetitions were conducted in the connected state because of the time and cost involved in running the mannequin. The mannequin was not restarted between all simulations in the connected state and the anesthetic machine was not formally checked before running the scenarios. Note that the data were taken from the internal logs of the simulators, not from the monitors. The Anesoft product does not have the capability of simulating a response to blood loss, so with this simulator, we used only interventions 3 to 7. It is a screen-based[4] simulator, so it was not possible to replicate the "connected" interventions, and we simply repeated each of these 5 interventions 10 times using its software. We did this for each of the 3 virtual "patients."

The interventions were initiated via the computer interface, but oxygen and anesthetic vapor were applied directly to the mannequin in the connected state.

For all interventions, the "patients" were administered air (7 L/min) throughout the duration of the simulations. For interventions 3 and 5, 100% oxygen (6 L/min) was administered from 5 minutes before the drug administration throughout the simulation. For intervention 7, 100% oxygen (6 L/min) was administered at 30 seconds, and sevoflurane was administered at 1 minute and discontinued after 10 minutes. With the Anesoft simulator, the oxygen was administered via a virtual facemask. With the METI HPS in the disconnected state, the oxygen was administered by setting the inspired oxygen concentration. With the METI HPS in the connected state, the simulator's trachea was intubated before all simulations began, and all gases were delivered through the endotracheal tube. All interventions began 1 minute after the start of the simulation to allow the models to stabilize and ended after approximately 20 minutes.

Quantifying Similarity of Time Series

The internal log files of both simulators were converted into Matlab (R2006b) structures for analysis. The data were linearly interpolated to 1 Hz and truncated to the duration of the shortest simulation, so all comparisons were of signals with the same length. Other interpolation schemes were tried, but these made little difference to the results. All data of a particular variable (eg, heart rate; Table 2) were compared with all other data of that variable in the 5 or 10 repetitions of a particular intervention (eg, 100-mg propofol with oxygen; Table 1) for the particular simulator and state (disconnected or connected).

Comparisons of the physiologic signals from successive simulations of the same intervention in the same simulator were made using 3 objective measures of similarity: root mean

TABLE 1. Interventions Used to Assess the Repeatability of the Simulators (See Text for Details)

Intervention 1	—	500-mL blood loss at a rate of 33.33 mL/s
Intervention 2	—	2000-mL blood loss at a rate of 33.33 mL/s
Intervention 3	—	100-mg propofol with oxygen
Intervention 4	—	100-mg propofol without oxygen
Intervention 5	—	150-mg propofol with oxygen
Intervention 6	—	150-mg propofol without oxygen
Intervention 7	—	2% sevoflurane with oxygen

TABLE 2. Mean and Range for Each Variable for All Interventions (Raw Data) From the METI HPS in the Disconnected and Connected States

Variable and Units	Mean (Range), Disconnected	Mean (Range), Connected	Anesoft Simulator
Heart rate, beats/min	76.5 (62.0 to 98.0)	75.2 (68.0 to 92.0)	81.8 (0.0–134.0)
Systolic blood pressure, mm Hg	109.1 (87.0 to 119.0)	111.3 (91.0 to 120.0)	109.7 (18.0–149.0)
Diastolic blood pressure, mm Hg	50.6 (39.0 to 55.0)	50.1 (45.0 to 54.0)	74.9 (17.0–105.0)
Mean blood pressure, mm Hg	—	—	65.3 (17.0–96.0)
Central venous pressure, mm Hg	7.5 (−5.0 to 14.0)	5.6 (−7.0 to 11.0)	7.6 (7.0–9.0)
Systolic pulmonary arterial pressure, mm Hg	25.7 (9.0 to 31.0)	24.5 (9.0 to 30.0)	34.3 (18.0–43.0)
Diastolic pulmonary arterial pressure, mm Hg	13.0 (1.0 to 18.0)	11.3 (0.0 to 16.0)	21.6 (14.0–28.0)
Mean pulmonary arterial pressure, mm Hg	—	—	14.5 (10.0–19.0)
Pulmonary capillary wedge pressure, mm Hg	7.0 (−6.0 to 13.0)	5.1 (−7.0 to 10.0)	6.6 (5.0–9.0)
Cardiac output, L/min	5.6 (3.9 to 6.1)	5.8 (5.2 to 6.1)	5.2 (0.1–7.4)
Tidal volume, mL	723.0 (110.0 to 1006.0)	695.1 (0.0 to 1179.0)	251.6 (0.0–801.0)
Respiratory rate, breaths/min	14.5 (5.0 to 21.0)	13.5 (0.0 to 20.0)	13.4 (0.0–24.0)
Oxygen saturation (Spo_2) %	98.5 (89.0 to 100.0)	99.2 (97.0 to 100.0)	83.83 (0.0–99.0)
Alveolar oxygen partial pressure (PAO_2),† mm Hg	300.9 (55.0 to 671.0)	360.8 (94.0 to 769.0)	—
Alveolar carbon dioxide partial pressure ($PACO_2$), mm Hg	40.9 (37.1 to 50.5)	42.2 (0.0 to 70.3)	—
Alveolar nitrogen partial pressure (PAN_2),* mm Hg	377.1 (6.0 to 621.0)	316.2 (−63.0 to 583.0)	—
Alveolar nitrous oxide partial pressure (PAN_2O),† mm Hg	0.0 (0.0 to 0.0)	0.3 (0.0 to 3.0)	—
Inspired oxygen (inO_2), %	—	—	79.0 (20.0–99.0)
Inspired anaesthetic agent (inAA), %	—	—	0.2 (0.0–1.5)
End-tidal anaesthetic agent (etAA), %	—	—	0.2 (0.0–1.5)
Isoflurane partial pressure (PISO),† mm Hg	0.0 (0.0 to 0.0)	0.0 (0.0 to 0.0)	—
Sevoflurane partial pressure (P_{SEVO}), mm Hg	1.0 (0.0 to 15.2)	0.5 (0.0 to 6.9)	—
Halothane partial pressure (PHALO),† mm Hg	0.0 (0.0 to 0.0)	0.0 (0.0 to 0.0)	—
Enflurane partial pressure (PENF),† mm Hg	0.0 (0.0 to 0.0)	0.0 (0.0 to 0.0)	—
Arterial oxygen partial pressure (Pao_2), mm Hg	284.7 (57.0 to 647.0)	342.7 (92.0 to 703.0)	—
Arterial carbon dioxide partial pressure ($Paco_2$), mm Hg	41.0 (37.7 to 50.1)	42.2 (26.9 to 49.5)	—
End-tidal CO_2 ($etCO_2$), mm Hg	—	—	24.8 (0.0–71.0)
Hemoglogin,† g/dL	14.4 (14.4 to 14.4)	14.4 (14.4 to 14.4)	—
Hematocrit,† %	42.3 (42.3 to 42.4)	42.3 (42.3 to 42.3)	—
Blood pH†	7.4 (7.4 to 7.5)	7.4 (7.4 to 7.5)	—
Venous oxygen partial pressure (PvO_2), mm Hg	47.9 (38.8 to 62.4)	51.8 (41.5 to 67.1)	—
Venous carbon dioxide partial pressure ($PvCO_2$), mm Hg	45.5 (43.1 to 52.7)	46.2 (44.0 to 51.9)	—
Temperature, °C	—	—	36.4 (36.3–36.6)
Esophageal temperature (Tesoph),† °C	36.5 (36.5 to 36.5)	36.5 (36.5 to 36.5)	—
Blood temperature,† °C	37.0 (37.0 to 37.0)	37.0 (37.0 to 37.0)	—

*These variables had clinically significant differences between the connected and disconnected states in the HPS.
†These variables were not used in further analysis.

squared error (RMSE),[16] median performance error (MDPE), and median absolute performance error (MDAPE).[17] The latter 2 measures require a "criterion standard" to be specified as a denominator, but it was not possible to decide this a priori, so the denominator of the equation was modified to be the mean value of the 2 data sets being compared (Eq. (1)). In addition, divergence was calculated to give an indication of any changes in difference over time.[17] The mean and SD of the similarity measures are reported in Tables 3, 4, and 5.

$$\mathrm{PE}(a,b)_i = \frac{a_i - b_i}{{}^1\!/_2(a_i + b_i)} \times 100 \qquad (1)$$

Equation 1: Performance error (PE) for 2 signals, *a* and *b*, over all points in time, *i*. PE is calculated as a percentage.

Differences in Time – Phase and Frequency

Periodic data may have slight offsets in different simulations because of slightly different starting times. For example, respiratory sinus arrhythmia may result in heart rate values that seem different because of differences in the timing of the first breath. There may also be slight differences in frequency. Differences in phase and slight differences in frequency are of no clinical importance when considering the repeatability of simulators. To account for apparent differences attributable to phase offsets, 3 methods to preprocess the data before using the methods mentioned earlier were chosen: simple phase shifting (SPS), complex phase shifting (CPS), and dynamic time warping (DTW). The first 2 methods are concerned primarily with differences in phase. For SPS, the 2 time series are aligned to their points of maximum cross-correlation,[18] calculated using the whole time-series. A refinement of this approach, CPS, recognizes that the timing of interventions after the start of data collection may affect the subsequent phase alignment. Therefore, with CPS, the data are divided into segments at the time of each intervention. Each segment is then aligned in the same manner as in SPS. The segments are subsequently rejoined for analysis. DTW is used in a number of fields, from gesture recognition to robotics.[19] The method stretches

TABLE 3. Mean (SD) of Indicators of the Similarity of the Repeated Series (RMSE [in the Units of the Variable], MDPE [%], MDAPE [%]) for Each Variable Under Difference Phase/Frequency Shifts (RAW, No Phase Shift) Pooled for All Interventions (HPS, Disconnected State)

	RMSE				MDPE, %				MDAPE, %			
Variables	**RAW**	**SPS**	**CPS**	**DTW**	**RAW**	**SPS**	**CPS**	**DTW**	**RAW**	**SPS**	**CPS**	**DTW**
Heart rate	0.6 (0.2)	0.6 (0.2)	0.6 (0.2)	0.1 (0.0)	−0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	0.0 (0.0)	0.3 (0.3)	0.3 (0.3)	0.3 (0.3)	0.0 (0.0)
Systolic arterial blood pressure	1.2 (0.5)	1.2 (0.5)	1.2 (0.5)	0.3 (0.1)	−0.0 (0.1)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	0.6 (0.4)	0.6 (0.4)	0.6 (0.4)	0.0 (0.0)
Diastolic arterial blood pressure	0.8 (0.2)	0.8 (0.2)	0.8 (0.2)	0.2 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.8 (0.5)	0.8 (0.5)	0.8 (0.5)	0.0 (0.0)
Central venous pressure	1.9 (0.6)	1.4 (0.3)	1.4 (0.3)	0.4 (0.1)	0.0 (0.9)	0.0 (0.1)	0.0 (0.2)	0.0 (0.0)	35.7 (69.6)	15.5 (17.6)	14.0 (13.8)	0.8 (1.1)
Systolic pulmonary arterial pressure	2.1 (0.6)	2.0 (0.5)	2.0 (0.5)	0.4 (0.1)	−0.0 (0.3)	−0.0 (0.2)	0.0 (0.2)	0.0 (0.0)	4.9 (4.0)	4.4 (3.1)	4.3 (3.0)	0.2 (0.3)
Diastolic pulmonary arterial pressure	2.3 (0.6)	1.8 (0.3)	1.8 (0.3)	0.4 (0.1)	−0.0 (0.2)	−0.0 (0.2)	−0.0 (0.1)	0.0 (0.0)	12.2 (11.8)	7.8 (5.4)	7.6 (5.2)	0.5 (0.7)
Pulmonary capillary wedge pressure	2.1 (0.5)	1.6 (0.3)	1.6 (0.3)	0.4 (0.1)	0.0 (0.0)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	35.9 (68.5)	16.8 (19.0)	15.4 (16.7)	0.8 (1.2)
Cardiac output	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.5 (0.4)	0.5 (0.4)	0.5 (0.4)	0.0 (0.0)
Tidal volume	6.8 (2.3)	6.8 (2.3)	6.8 (2.3)	1.8 (0.6)	−0.0 (0.1)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	0.3 (0.1)	0.3 (0.1)	0.3 (0.1)	0.0 (0.0)
Respiratory rate	0.2 (0.1)	0.2 (0.1)	0.2 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
PAO_2	1.4 (1.1)	1.4 (1.1)	1.4 (1.1)	0.4 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.3 (0.2)	0.3 (0.2)	0.3 (0.2)	0.0 (0.0)
$PACO_2$	0.4 (0.1)	0.4 (0.1)	0.4 (0.1)	0.1 (0.0)	−0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	0.0 (0.0)	0.5 (0.2)	0.5 (0.2)	0.5 (0.2)	0.1 (0.0)
PAN_2	1.2 (1.1)	1.2 (1.1)	1.2 (1.1)	0.3 (0.3)	−0.0 (0.2)	−0.0 (0.2)	−0.0 (0.2)	0.0 (0.0)	0.4 (0.6)	0.4 (0.6)	0.4 (0.6)	0.0 (0.0)
P_{SEVO}	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Pa_{O_2}	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Pa_{CO_2}	0.7 (0.8)	0.7 (0.8)	0.7 (0.8)	0.2 (0.3)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)
Sp_{O_2}	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	0.0 (0.0)	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)
PvO_2	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
PvCO2	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)

time-series data nonuniformly, thereby obviating the potential influence of phase differences. This approach also reduces any differences associated with slight frequency disparities.[20] The 3 methods described earlier (SPS, CPS, and DTW) are shown graphically in Figure 1 using an example pair of physiologic signals (eg, heart rate) from 2 hypothetical simulations.

RESULTS

A total of 255 simulations, representing 85 hours of simulation time, were performed. There were some variables in the METI HPS log files that varied insignificantly over the simulations and would not be expected to vary with the given interventions (starred in Table 2; see Discussion section). These variables were removed from further analysis. All interventions were performed at exactly the same time in each repetition of the simulation according to the logs, which have a resolution of 1 second.

The mean and SD of the divergence for all variables in all simulations with the Anesoft simulator and with the METI HPS in the disconnected state were zero. In the connected state, the mean and SD of the divergence for most variables in each simulation were zero: the maximum mean (SD) divergence was 0.2% (0.3%) per minute and was for PvO_2 during the simulation of intervention 1 under SPS. The RMSE, MDPE, and MDAPE for each variable under each transformation are summarized in Tables 3 to 5 for the METI HPS. Note that MDPE and MDAPE are expressed as a percentage difference to the reference signal (Eq. (1)), whereas the dimensions of RMSE are the units of the variable in question. These measures were all zero for all variables in all interventions and under all preprocessing steps for the Anesoft simulator (ie, repeatability was perfect).

Table 3 shows the values of the indicators of similarity for each method using data from the HPS in the disconnected state, whereas Table 4 shows the same values in the connected state.

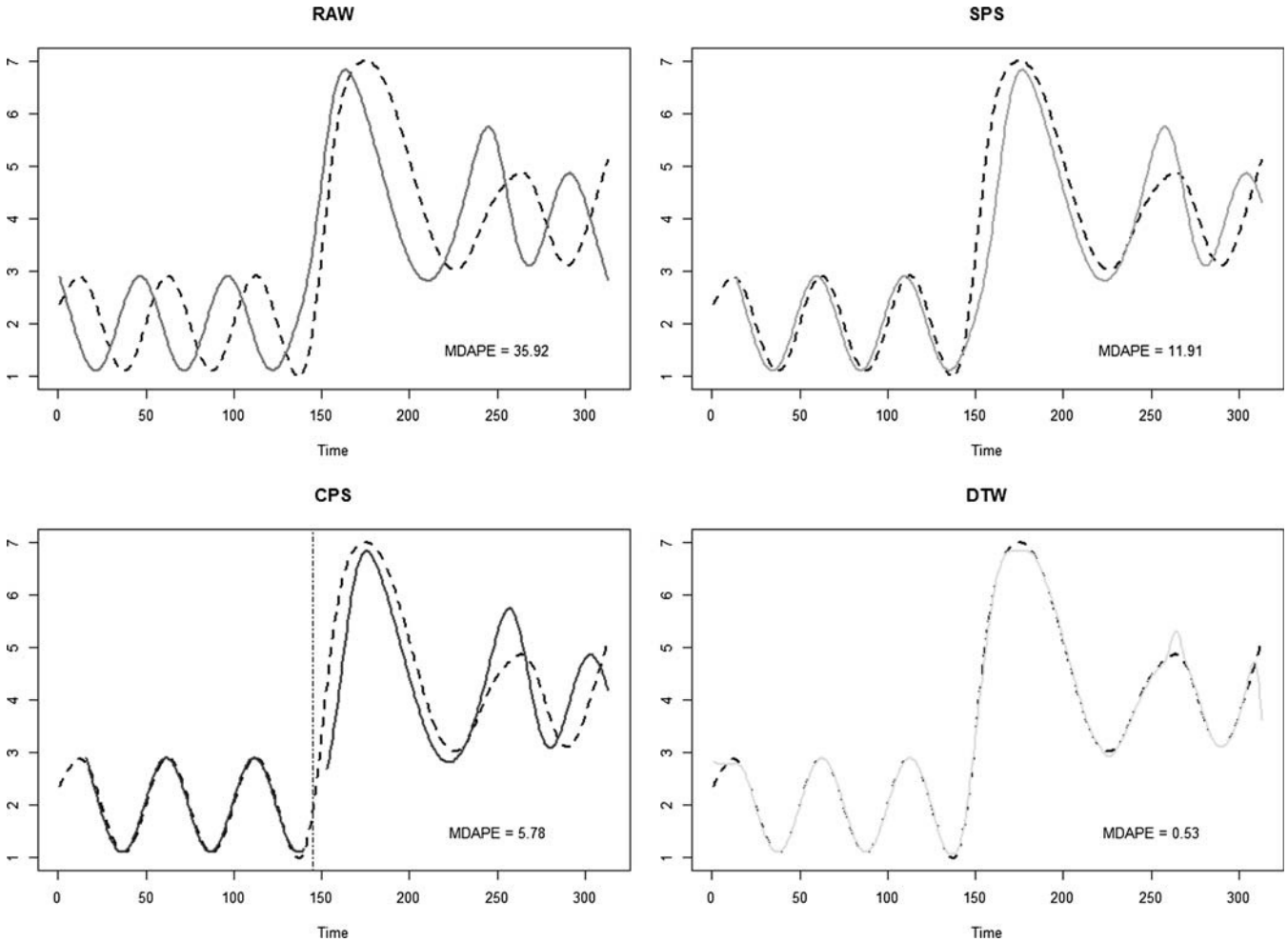

FIGURE 1. An illustrative example of 4 methods used to preprocess time series. The 2 lines represent 2 signals of, for example, heart rate, in repeated simulations (a and b in Eq. (1)). In this example, the dashed line represents the reference signal. RAW shows the data without preprocessing. With SPS, the nonreference time series is shifted until maximum cross-correlation is achieved. With CPS, the time series is divided into segments at the time of each intervention (there is 1 intervention at 145 seconds in this example, as indicated by the vertical line, and so there are 2 segments). Each segment is shifted until maximum cross-correlation is achieved, after which the segments are joined together and the similarity of the new signal with the reference signal is calculated. With DTW, the nonreference series (coloured line) is essentially divided into segments at each point and stretched or compressed to minimize the difference between the 2. For more details on this method, see Wang and Gasser[21] (1997). It can be seen that the MDAPE is progressively reduced by the application of each of these methods in turn (from 35.92% in RAW to 0.53% in DTW).

TABLE 4. Mean (SD) (RMSE [in the Units of the Variable], MDPE [%], MDAPE [%]) for Each Variable Under Difference Phase/Frequency Shifts (RAW, No Phase Shift) Pooled for All Interventions (HPS, Connected State)

	RMSE				MDPE, %				MDAPE, %			
Variables	**RAW**	**SPS**	**CPS**	**DTW**	**RAW**	**SPS**	**CPS**	**DTW**	**RAW**	**SPS**	**CPS**	**DTW**
Heart rate	0.7 (0.2)	0.7 (0.2)	0.7 (0.2)	0.2 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.5 (0.4)	0.5 (0.4)	0.5 (0.4)	0.0 (0.0)
SABP	1.3 (0.3)	1.3 (0.3)	1.3 (0.3)	0.3 (0.1)	−0.0 (0.1)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	0.8 (0.3)	0.8 (0.3)	0.8 (0.3)	0.0 (0.1)
DABP	0.9 (0.2)	0.9 (0.2)	0.9 (0.2)	0.2 (0.0)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	1.2 (0.5)	1.2 (0.5)	1.2 (0.5)	0.0 (0.0)
Central venous pressure	2.2 (0.4)	2.0 (0.4)	2.0 (0.4)	0.5 (0.3)	−0.0 (2.1)	0.0 (2.1)	−0.0 (2.1)	0.0 (0.0)	21.4 (9.0)	20.2 (9.3)	19.6 (9.0)	1.6 (1.9)
Systolic pulmonary arterial pressure	2.2 (0.3)	2.2 (0.3)	2.2 (0.3)	0.5 (0.2)	0.1 (0.6)	0.1 (0.6)	0.1 (0.6)	0.0 (0.0)	4.8 (2.3)	4.8 (2.3)	4.8 (2.3)	0.3 (0.4)
Diastolic pulmonary arterial pressure	2.4 (0.4)	2.3 (0.4)	2.3 (0.4)	0.5 (0.2)	0.4 (2.0)	0.4 (2.0)	0.4 (2.0)	0.0 (0.0)	14.8 (9.4)	14.3 (9.5)	14.4 (9.5)	0.9 (0.9)
Pulmonary capillary wedge pressure	2.2 (0.4)	2.1 (0.3)	2.1 (0.4)	0.5 (0.2)	0.3 (2.8)	0.3 (2.8)	0.2 (2.3)	0.0 (0.0)	23.9 (10.6)	22.1 (11.0)	21.2 (10.4)	1.6 (1.5)
Cardiac output	0.1 (0.0)	0.1 (0.0)	0.1 (0.0)	0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	−0.0 (0.0)	0.0 (0.0)	0.5 (0.4)	0.5 (0.4)	0.5 (0.4)	0.0 (0.0)
Tidal volume	49.7 (35.6)	40.3 (34.2)	37.8 (33.3)	17.5 (25.6)	−1.2 (3.4)	−1.2 (3.4)	−1.2 (3.4)	−0.6 (2.7)	1.8 (3.2)	1.8 (3.2)	1.8 (3.2)	0.7 (2.7)
Respiratory rate	1.0 (0.6)	0.8 (0.5)	0.8 (0.5)	0.3 (0.4)	−0.6 (2.1)	−0.6 (2.1)	−0.6 (2.1)	0.0 (0.0)	1.1 (2.3)	1.0 (2.3)	1.1 (2.3)	0.0 (0.4)
PAO_2	25.7 (35.3)	25.3 (34.9)	25.7 (35.3)	6.2 (9.1)	−0.2 (1.8)	−0.2 (1.8)	−0.2 (1.8)	0.0 (0.4)	1.4 (1.3)	1.4 (1.3)	1.4 (1.3)	0.3 (0.4)
$PACO_2$	3.1 (1.2)	2.5 (0.6)	2.5 (0.6)	0.6 (0.3)	−0.4 (1.4)	−0.4 (1.4)	−0.4 (1.4)	−0.0 (0.1)	2.8 (1.5)	2.8 (1.5)	2.8 (1.5)	0.2 (0.2)
PAN_2	25.1 (35.2)	24.4 (34.4)	22.9 (33.6)	6.4 (9.5)	4.1 (10.8)	4.1 (10.8)	4.1 (10.9)	2.3 (6.1)	7.2 (13.2)	7.2 (13.1)	6.9 (12.2)	2.8 (6.4)
P_{SEVO}	0.6 (0.2)	0.5 (0.2)	0.5 (0.1)	0.1 (0.0)	−12.2 (149.9)	−11.4 (138.8)	−10.7 (136.7)	0.0 (0.0)	191.8 (23.1)	190.5 (27.6)	189.2 (28.9)	8.3 (18.8)
Pao_2	0.1 (0.2)	0.1 (0.2)	0.1 (0.2)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
$Paco_2$	22.0 (32.0)	22.0 (32.0)	22.0 (32.0)	3.9 (7.3)	−0.3 (1.7)	−0.3 (1.7)	−0.3 (1.7)	0.1 (0.4)	1.0 (1.4)	1.0 (1.4)	1.0 (1.4)	0.2 (0.4)
Spo_2	0.8 (0.5)	0.8 (0.5)	0.8 (0.5)	0.3 (0.4)	−0.3 (0.8)	−0.3 (0.8)	−0.3 (0.8)	−0.1 (0.5)	0.5 (0.8)	0.5 (0.8)	0.5 (0.8)	0.1 (0.5)
PvO_2	0.3 (0.3)	0.3 (0.3)	0.3 (0.3)	0.1 (0.2)	0.0 (0.3)	0.0 (0.3)	0.0 (0.3)	0.0 (0.0)	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)	0.0 (0.0)
PvCO2	0.7 (0.9)	0.7 (0.9)	0.7 (0.9)	0.1 (0.1)	−0.1 (0.5)	−0.1 (0.5)	−0.1 (0.5)	−0.0 (0.1)	0.4 (0.4)	0.4 (0.4)	0.4 (0.4)	0.0 (0.1)

TABLE 5. Mean (SD) of the Similarity for Each Variable Under Difference Phase/Frequency Shifts (RAW, No Phase Shift) for All Interventions (HPSTM, Unconnected and Connected States)

		RMSE				MDPE, %				MDAPE, %			
Intervention		RAW	SPS	CPS	DTW	RAW	SPS	CPS	DTW	RAW	SPS	CPS	DTW
Unconnected	1	1.1 (1.6)	1.0 (1.5)	1.0 (1.5)	0.2 (0.4)	0.0 (0.2)	−0.0 (0.0)	−0.0 (0.1)	0.0 (0.0)	4.5 (10.0)	2.9 (5.6)	2.7 (5.2)	0.3 (0.7)
	2	1.2 (1.7)	1.1 (1.6)	1.1 (1.6)	0.2 (0.4)	−0.0 (0.6)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	22.1 (64.3)	8.2 (18.8)	7.3 (16.1)	0.1 (0.3)
	3	1.0 (1.3)	0.9 (1.2)	0.9 (1.2)	0.2 (0.4)	−0.0 (0.0)	−0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	1.8 (3.8)	1.4 (2.5)	1.3 (2.4)	0.0 (0.0)
	4	0.9 (1.4)	0.8 (1.4)	0.8 (1.4)	0.2 (0.4)	0.0 (0.0)	0.0 (0.1)	0.0 (0.0)	0.0 (0.0)	1.7 (3.4)	1.7 (3.3)	1.7 (3.3)	0.3 (0.7)
	5	1.2 (1.9)	1.1 (1.9)	1.1 (1.9)	0.2 (0.4)	0.0 (0.1)	0.0 (0.1)	0.0 (0.1)	0.0 (0.0)	1.8 (3.7)	1.4 (2.6)	1.4 (2.5)	0.0 (0.0)
	6	1.1 (1.7)	1.0 (1.6)	1.0 (1.6)	0.3 (0.5)	−0.0 (0.2)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	2.7 (5.4)	2.0 (3.6)	2.0 (3.5)	0.3 (0.7)
	7	1.4 (2.0)	1.3 (2.0)	1.3 (2.0)	0.4 (0.5)	−0.0 (0.1)	−0.0 (0.1)	−0.0 (0.1)	0.0 (0.0)	1.3 (2.1)	1.2 (1.9)	1.1 (1.8)	0.0 (0.0)
Connected	1	4.8 (17.8)	3.9 (15.0)	3.9 (15.0)	2.0 (10.8)	−0.5 (41.3)	0.0 (40.0)	0.7 (38.7)	−0.3 (1.7)	14.5 (43.2)	14.3 (43.0)	14.2 (43.0)	1.5 (6.0)
	2	3.2 (8.7)	2.4 (3.8)	2.4 (4.2)	0.7 (1.3)	5.8 (31.5)	5.3 (30.4)	4.9 (29.7)	−0.0 (0.2)	17.6 (45.3)	17.6 (45.3)	17.5 (45.2)	1.2 (7.9)
	3	9.1 (20.9)	8.8 (20.3)	8.3 (19.3)	2.2 (4.9)	−1.1 (29.8)	−0.6 (23.3)	−1.0 (25.7)	0.6 (3.2)	14.3 (37.8)	13.3 (35.8)	13.0 (35.1)	1.1 (3.5)
	4	9.2 (18.3)	8.5 (16.7)	7.5 (14.8)	3.5 (8.7)	−5.4 (32.2)	−5.5 (32.0)	−5.3 (30.6)	0.4 (2.2)	14.3 (42.9)	13.7 (42.5)	13.7 (42.5)	0.8 (2.4)
	5	2.8 (8.5)	2.1 (4.8)	2.2 (5.0)	0.8 (2.4)	4.6 (28.8)	4.1 (24.4)	4.1 (24.4)	−0.0 (0.0)	13.7 (44.4)	13.7 (44.4)	13.5 (44.4)	0.1 (0.3)
	6	4.6 (17.7)	4.5 (17.7)	4.5 (17.7)	2.8 (13.3)	−7.0 (34.6)	−6.8 (34.1)	−6.7 (32.7)	−0.0 (0.0)	13.5 (43.1)	13.9 (44.3)	13.5 (43.1)	0.5 (3.0)
	7	15.8 (33.6)	15.2 (33.0)	15.2 (33.0)	1.5 (3.8)	−0.2 (38.3)	−0.0 (33.4)	−0.0 (33.4)	−0.0 (0.0)	13.3 (43.3)	13.1 (43.3)	13.1 (43.3)	1.2 (6.7)

The values of the indicators of similarity for all variables listed in Tables 3 and 4 were then pooled for each intervention for the HPS in the disconnected and connected states, respectively (Table 5), to give an overall measure of similarity.

DISCUSSION

This article has compared approaches for quantifying the similarity of the modeled physiologic responses of simulators to a selection of simple interventions. The Anesoft simulator showed perfect repeatability in its modeled responses to our interventions (ie, RMSE, MDPE, MDAPE in all preprocessing conditions were all zero). A likely explanation would be that the models are deterministic. The METI HPS showed some variation between simulations for most variables in all 7 interventions, but this was primarily a function of slight phase and frequency differences. In this study, our interest was in developing an objective method to quantify this variation rather than in evaluating its clinical significance, which presumably would have required seeking consensus from a suitable sample of clinicians. However, for both simulators, we thought the degree of similarity between simulations in the responses to the interventions investigated was more than adequate for most applications of simulation in education, assessment, and research related to clinical anesthesia, in that we thought it unlikely that the variation seen would have been sufficient to prompt differences in clinical decisions made on the basis of the responses.

The poor reproducibility previously reported by Mudumbai et al[14] could perhaps have been attributable to slight phase and frequency offsets that are not clinically relevant. This interpretation is supported by the marked improvement in our indicators of similarity when our data were processed with the DTW algorithm (Tables 3–5). There was little improvement in these indicators for many of the variables when only phase was accounted for (ie, using the SPS and CPS algorithms). This suggests that phase is not the only cause for the differences and that slight variations in frequency may also contribute. The DTW algorithm therefore seems to be particularly useful when comparing time series of this type.

None of the interventions included the administration of nitrous oxide or anesthetic vapors other than sevoflurane, so it is unsurprising that PSIO, PHALO, and PENF were zero throughout the simulations. There was some minor variation in PAN_2O during the simulations in the connected state with the METI HPS. This could be due to residual N_2O in the anesthetic circuit. However, the variation (0–3 mm Hg) would not be clinically relevant unless there was a medical reason that required the exclusion of N_2O from the circuit. Similarly, the variation in hematocrit (42.3%–42.4%) or pH (7.4–7.5) would not be significant in either the disconnected or connected states. There was no change in hemoglobin, esophageal temperature, or blood temperature values. These variables would not be expected to change much between simulations of the given interventions, and so, removing them from analysis was appropriate.

There are some differences in the overall measures of similarity when comparing interventions (Table 5). The high SD in RMSE and MDAPE for interventions 1, 4, and 6 (Table 6) could be attributable to the fact that some of the simulations

were conducted on different days when atmospheric conditions may have been slightly different and so the composition of gases in the environment may have varied. Barometric pressure, temperatures, and humidity were not measured. Interventions 3, 5, and 7 used oxygen as part of the intervention, and so, gas compositions in the environment are less important because oxygen would make up the majority of the administered gas.

Root mean squared error might be expected to be larger for variables such as tidal volume (when measured in milliliters, as is usual) as these values are in the order of 10^3, whereas the values of most other variables are in the order of 10^2 (eg, heart rate) or 10^1 (eg, central venous pressure). Therefore, RMSE may not be the most appropriate measure for a global estimate of overall variability. The MDAPE may be a more appropriate measure than MDPE because it gives a more meaningful estimate of the size of the differences, rather than their direction, which may be less important in this context.[22] The modification made to Eq. 1 (ie, the denominator representing the mean value of the data) will still capture significant deviations without unfairly penalizing a set of simulations if the time series selected as the criterion standard was considerably different from the rest.

Other similarity measures were considered for this study but were not as appropriate for our purpose as those chosen earlier. The Bland-Altman method,[23] for example, is more suitable for comparing different measurement techniques than comparing different measures from the same technique and is less useful for time-series data.[24] Similarity measures based on symbolic representations, such as symbolic aggregate approximation,[21] inherently cause some loss of information and are more suited to fast searching or indexing of data than evaluating repeatability. The data were not normalized before processing—a step often taken for this type of work[19]—because the absolute differences may be important from a clinical perspective.

The choice of linear interpolation to 1 Hz was done to cater for nonuniform time steps in the logs of the simulators and to provide a standard method that could apply to other simulators. There may be subtle differences in the results given a different frequency or interpolation scheme. An analysis of this was out of scope for this work.

There are some clinically significant differences in mean and SD of PAO_2, PAN_2, and Pa_{O_2} between simulations run in the disconnected and connected states, respectively (Tables 4 and 5). These differences are likely to reflect the composition of the physical gases when the mannequin is connected to the circuit. Thus, in the disconnected state, all variables were highly repeatable (Table 4), but in the connected state, the mean and SD of the MDAPE for PAN_2 and P_{SEVO} were larger than the other variables. The connected state is more relevant for most uses of the METI HPS in our facility (where we apply drug interventions via the computer but use real vapors and gases and standard anesthesia monitors), whereas the disconnected state allows a more controlled evaluation of the repeatability of the modeled physiology. The difference in P_{SEVO} could be due to minor differences in the amount of volatile agent manually "dialed up" during the simulations. Similarly, differences in the amount of nitrous oxide in the circuit could account for differences in PAN_2. Some differences may also be due to a small leak in the endotracheal tube as the completeness of the seal was not verified. In assessing the relevance of these minor discrepancies, it is regrettable that there are no benchmarks or standards to inform a more objective conclusion, but our subjective view was that they were of little consequence for most applications of this simulator. Interestingly, there were some variables in some interventions that displayed perfect repeatability once warped. The zero values for divergence suggest that differences did not accumulate over time.

A limitation of this study relates to the fact that the DTW algorithm does not uniformly stretch the time scales of each series to be compared over the length of the data: we did not see any obvious aberrations in the signals after DTW, but we did not formally test this. Most of the variables show very small RMSE, MDPE, and MDAPE once transformed. These small values indicate a very good level of repeatability, especially given that MDPE and MDAPE values less than 20% have been considered acceptable in pharmacologic simulation studies.[22] Future work could establish a benchmark for clinical acceptability of difference performance errors. These results are based on 2 simulators, and it would be useful to test the methodology on other available model-driven simulators.[4] We measured intrasimulator repeatability and did not, like Garden et al,[15] measure intersimulator repeatability, but the approach used here would also be applicable to comparing simulators in this way. For example, it might be important to know whether a fully validated scenario developed in one country could be transferred to another and work in substantially the same way, so that validated scenarios could be freely shared and an ever-expanding scenario library could be created and used with confidence.

Our interventions were simple. It is possible that more complex interventions or several simultaneous interventions would be associated with greater variation between simulations. Our work provides an approach that could be used to assess repeatability in more complex scenarios or even the repeatability of "instructor-driven" simulations. It would be useful to develop a standard for acceptable levels of repeatability in this context, keeping in mind that a small amount of variation around a consistent mean may be perceived as more natural than perfect repeatability. This would need to account for possible differences in the timing of interventions also as it may be difficult to exactly replicate complex cases. In these simple interventions, the timing of events was perfect insofar as the resolution of the logs (1 second) is concerned.

Repeatability is only one aspect of how a simulator performs: a model may be perfectly repeatable and yet produce physiologic data that are not clinically realistic. We have not addressed this aspect of the 2 simulators in this study. We also acknowledge that the 2 simulators evaluated in this study are quite different: one is hardware based (users interact with a mannequin), and one is screen based (users interact with a computer). We have focused solely on their physiology, which could be applied to either mode of interaction. We have not considered the physical components of the simulation (such as moulage, facilitators, or airway

anatomy) that might be present. We have also not considered the potential difference between the modeled physiology (from the logs) and that displayed to clinicians through monitors in the case of the connected HPS. There are also differences in the variables modeled by each simulator. Our intention was not to compare these simulators but rather to explore methods for evaluating the repeatability of their physiologic responses to a selection of interventions.

In conclusion, the repeatability of the physiologic response of model-based simulators to simple standardized interventions can be evaluated by considering divergence over time and the MDAPE of individual or pooled variables, but data should be preprocessed to eliminate irrelevant phase and frequency offsets in some variables. The DTW is an effective method for this purpose.

REFERENCES

1. Gaba DM. The future vision of simulation in health care. *Qual Saf Health Care* 2004;13:i2–i10.
2. Good ML. Patient simulation for training basic and advanced clinical skills. *Med Educ* 2003;37:14–21.
3. Maran NJ, Glavin RJ. Low- to high-fidelity simulation—a continuum of medical education? *Med Educ* 2003;37:22–28.
4. Cumin D, Merry AF. Simulators for use in anaesthesia. *Anaesthesia* 2007;62:151–162.
5. Blike G, Cravero J, Andeweg S, Jensen J, Christoffersen K. Standardized simulated events for provocative testing of medical care system rescue capabilities. *Advances in Patient Safety: From Research to Implementations* 2005.
6. Cumin D, Merry AF, Weller JM. Standards for simulation. *Anaesthesia* 2008;63:1281–1284.
7. Cumin D, Weller JM, Henderson K, Merry AF. Standards for simulation in anaesthesia: creating confidence in the tools. *Br J Anaesth* 2010;105:45–51.
8. Das G, Gunopulos D. Time Series Similarity and Indexing. *The Handbook of Data Mining*. 2003. Available at http://tandf.net/books/details/9781410607515/.
9. Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment* 2008;1:1542–1552.
10. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;8: 307–310.
11. Ito S, Mardimae A, Han J, et al. Non-invasive prospective targeting of arterial P(CO2) in subjects at rest. *J Physiol* 2008;586:3675–3682.
12. Knesaurek K, Machac J, Zhang Z. Repeatability of regional myocardial blood flow calculation in ^{82}Rb PET imaging. *BMC Med Phys* 2009;9:2.
13. Strouthidis NG, White ET, Owen VM, Ho TA, Garway-Heath DF. Improving the repeatability of Heidelberg retina tomograph and Heidelberg retina tomograph II rim area measurements. *Br J Ophthalmol* 2005;89:1433–1437.
14. Mudumbai S, Lighthall G, Sun C, Harrison K, Davies F, Howard S. Reproducibility of a model driven simulator and sensitivity to initial conditions: research abstract: 23. *Simul Healthc* 2007;2(1):58.
15. Garden AL, Robinson BJ, Arancibia CU, et al. Unrecognized malfunction in computerized patient simulators. *Br J Anaesth* 2004;93:873–875.
16. Li XR, Zhao Z. Measures of performance for evaluation of estimators and filters. *Proceedings of SPIE Conference on Signal and Data Processing of Small Targets* 2001;4473–61.
17. Varvel JR, Donoho DL, Shafer SL. Measuring the predictive performance of computer-controlled infusion pumps. *J Pharmacokinet Biopharm* 1992;20:63–94.
18. Orfanidis SJ, ed. *Optimum Signal Processing: An Introduction*. London, England: Collier Macmillan; 1988.
19. Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery* 2003;7:349–371.
20. Wang K, Gasser T. Alignment of curves by dynamic time warping. *The Annals of Statistics* 1997;25:1251–1276.
21. Lin J, Keogh E, Lonardi S, Chiu B. A symbolic representation of time series, with implications for streaming algorithms. 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery; 2003.
22. Hendrickx JF, Lemmens HJ, Shafer SL. Do distribution volumes and clearances relate to tissue volumes and blood flows? A computer simulation. *BMC Anesthesiol* 2006;6:7.
23. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet* 1995;346:1085–1087.
24. Myles PS, Cui J. Using the Bland-Altman method to measure agreement with repeated measures. *Br J Anaesth* 2007;99: 309–311.