Human Core Temperature Prediction for Heat-Injury Prevention

Srinivas Laxminarayan, Mark J. Buller, William J. Tharion, and Jaques Reifman

***Abstract*—Previously, our group developed autoregressive (AR) models to predict human core temperature and help prevent hyperthermia (temperature > 39 °C). However, the models often yielded delayed predictions, limiting their application as a real-time warning system. To mitigate this problem, here we combined AR-model point estimates with statistically derived prediction intervals (PIs) and assessed the performance of three new alert algorithms [AR model plus PI, median filter of AR model plus PI decisions, and an adaptation of the sequential probability ratio test (SPRT)]. Using field-study data from 22 soldiers, including five subjects who experienced hyperthermia, we assessed the alert algorithms for AR-model prediction windows from 15–30 min. Cross-validation simulations showed that, as the prediction windows increased, improvements in the algorithms' effective prediction horizons were offset by deteriorating accuracy, with a 20-min window providing a reasonable compromise. Model plus PI and SPRT yielded the largest effective prediction horizons ($\geq$ 18 min), but these were offset by other performance measures. If high sensitivity and a long effective prediction horizon are desired, model plus PI provides the best choice, assuming decision switches can be tolerated. In contrast, if a small number of decision switches are desired, SPRT provides the best compromise as an early warning system of impending heat illnesses.**

***Index Terms*—Autoregressive (AR) model, core temperature, hyperthermia, prediction interval (PI), sequential probability ratio test (SPRT).**

I. Introduction

HEAT injury is a problem for the United States (U.S.) Armed Forces, especially during deployments to localities with hot and humid climates, and trends show the number of heat injury cases to be on the rise each year [1]. From 2006 through 2010, there were 2887 heat injuries across the services, including 311 cases of heat stroke. The risk of heat injury is modulated by both intrinsic factors (such as genetics, fitness, acclimatization, and sleep quality) and extrinsic factors (such as exercise intensity and duration, clothing and equipment, ambient temperature, relative humidity, and solar radiation). In any case, rising body core temperature is a sign of impending heat injury, starting with heat stress (beyond 37 °C), progressing to hyperthermia and heat exhaustion (beyond 39 °C), and then to heat stroke (beyond 40 °C) [2]. The rising core temperature sets off a cascade of physiological responses to preserve temperature homeostasis (∼37 °C). However, beyond a critical temperature, which depends upon an individual's intrinsic and extrinsic factors, homeostasis cannot be sustained, leading to pathophysiological responses that may culminate in multiorgan dysfunction and death [2], [3].

Early recognition of heat stress combined with changes of activity and cooling strategies—such as cold-water immersion, water spraying, rest in an air-conditioned area, consumption of cold beverages, and ingestion of crushed ice—can reduce the morbidity and mortality associated with heat illnesses [4]. However, certain circumstances make early recognition and intervention challenging. For example, at the height of a military operation or during an athletic competition, soldiers and athletes may not perceive the warning signs of a rising core temperature and impending heat illnesses [5]. This could be achieved by a system capable of recognizing early trends, reliably predicting the onset of core temperature rise, and generating alerts. Sensor technologies, which afford the ability to measure human core temperature via ingestible pills [6], and mathematical predictive models [7]–[9] could be coupled to develop such a hardware/software system and potentially minimize the incidences of heat injuries.

Data-driven models, such as artificial neural networks (ANNs) and autoregressive (AR) models, which use a time series of recent-past measurements to predict their future time courses, have been widely applied for forecasting various physical and biological signals [10]–[12]. (Note that AR models constitute a particular case of ANNs with linear mapping functions.) In our earlier work [8], we used past core temperature measurements to "learn" autocorrelations inherently present in the time-series data and develop AR models to predict future core temperature values for a defined prediction window. We also computed the corresponding prediction intervals (PIs), which provide a measure of reliability of the AR-model predictions [13]. Importantly, we found that because AR models only depend on the frequency of the underlying time-series data, and because the frequency of the core-temperature signal is invariant from individual to individual, AR models could be constituted as "universal" predictors. That is, once an AR model has learned autocorrelations for one individual, it can be directly used to predict the core temperature of other individuals in similar populations without any additional adaption of the model parameters.

Manuscript received January 11, 2014; revised May 12, 2014; accepted June 16, 2014. Date of publication June 20, 2014; date of current version May 7, 2015. This work was supported by the Military Operational Medicine Research Area Directorate of the U.S. Army Medical Research and Materiel Command, Ft. Detrick, MD 21702 USA.

S. Laxminarayan and J. Reifman are with the Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD 21702 USA (e-mail: srinivas@bhsai.org; jaques.reifman.civ@mail.mil).

M. J. Buller and W. J. Tharion are with the U.S. Army Research Institute of Environmental Medicine, Biophysics and Biomedical Modeling Division, Natick, MA 01760 USA (e-mail: mark.j.buller.civ@mail.mil; william.j.tharion.civ@mail.mil).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2014.2332294

However, when we attempted to use AR models for real-time prediction of core temperature, we found a key limitation [9]: the AR models introduced significant prediction time lags (sometimes as long as the prediction window itself) during large excursions of the core temperature signal, which is exactly the condition when we desire the model to be most accurate. Such prediction time lags spring from the observation that AR models are essentially filters applied to an ordered sequence of data samples and, hence, yield delayed responses. This is a well-recognized problem also documented in other applications of predictive models [14]–[16]. To be clinically useful, the predictions must be made with sufficient lead-time (effective prediction horizon: estimated to be ∼20 min in hyperthermia [9]) to allow for proactive interventions that can alter the course of the clinical outcome.

In this study, we aim to address this limitation, so as to allow for the practical use of AR models as an effective tool to accurately predict core-temperature rises and provide clinically useful alerts of impending heat injury. To this end, we developed three new alert algorithms that combine AR model predictions and associated PIs, and assessed their ability to reduce delays in the prediction of hyperthermia. We also compared their performance against that of the AR-model predictions alone. The simplest alert algorithm only uses the current AR-model prediction and PI. The other two algorithms use a series of recent-past predictions and PIs to make the decision. Of these, one alert algorithm provides explicit, adjustable parameters, separate from the AR-model parameters, to tradeoff conflicting measures of performance.

To evaluate the performance of these algorithms, we used field-study data from 22 subjects involved in military activities, including five subjects whose core temperature rose beyond 39 °C, the onset of hyperthermia. We assessed the algorithms against four measures of performance, namely, effective prediction horizon, sensitivity, specificity, and decision switches. Two algorithms reliably predicted the onset of hyperthermia ∼20 min in advance with reasonably high sensitivity and specificity, with one of them yielding fewer decision switches. These promising results form the basis for real-time core temperature predictions, which, when coupled with temperature sensors, could provide reliable early warning of impending heat illnesses.

II. Methods

A. AR Model

Given temperature measurements y_{n-i} sampled every *S* min, where *n* is the current discrete time index and $i = 0, 1, \ldots, m-1$, the AR model of order *m* predicts signal $\hat{y}_{n+1}$, at time point $n + 1$, through a linear combination of the antecedent core-temperature samples as follows:

$$\hat{y}_{n+1} = \sum_{i=0}^{m-1} b_i y_{n-i} \tag{1}$$

where *b* denotes the vector of *m* unknown AR coefficients. To make predictions *M* time steps ahead (prediction window $P = M \times S$ min ahead), we iteratively used (1) *M* times, substituting the unobserved signals at $n \geq n+1$ in the summation by their corresponding predicted values. The order *m* of the model specifies the required initial waiting period for which data samples need to be collected before real-time predictions can be made. We chose the sampling period *S* that preserved the important frequencies and rejected the high-frequency noise in the magnitude spectrum of the core-temperature data [17]. Subsequently, we chose *m* to be the number of lags in the data beyond which the partial autocorrelation function was essentially zero [10].

Before applying the AR model, we must first estimate the AR coefficients *b* using some "training" data. To estimate *b*, we used the standard forward–backward least squares procedure (see [17, Ch. 8]) implemented in MATLAB version 7.14 (function *ar*). In this procedure, we first used the entire time series of temperature measurements to form a time-forward convolution matrix, and then a time-backward convolution matrix. Next, we combined the two matrices into one, estimated the corresponding autocorrelation matrix, and inverted it to estimate the coefficients *b*. An advantage of the forward–backward procedure is that it implicitly regularizes the autocorrelation matrix (which can be ill conditioned due to noise in the data) to ensure robust estimates of *b*.

In many safety-critical applications, providing single-point temperature predictions may not be sufficient for decision making and may require information about the uncertainty of the predicted values. In earlier work [18], we developed a technique based on the statistical bootstrap method [19] to estimate prediction uncertainty in the form of PIs. The technique relies on the idea of model resampling [13] rather than data resampling, where a population of models is built based on blocks of data that are randomly drawn from the original time series to form an empirical distribution of models (i.e., a distribution of the model coefficients). Following this procedure [18] to compute the PIs, we estimated the covariance matrix Σ of the AR model from a distribution of models for an *M* time-step-ahead predictor and used the following equation [18]:

$$\mathrm{PI} = Z_{\alpha/2}\sqrt{y^T \Sigma y + \sigma^2} \tag{2}$$

where $Z_{\alpha/2}$ denotes the prediction factor associated with an $\alpha\%$ type I error, y represents a vector of data samples $y = \left[\, y_n \;\; y_{n-1} \;\; \ldots \;\; y_{n-m+1} \,\right]^T$, and σ^2 denotes the variance of the measurement noise. We used the AR model in (1) to make *M* step-ahead predictions (*P*-min-ahead) of core temperature and (2) to estimate the PIs.

As mentioned earlier, we found that AR models introduce significant prediction time lags during large excursions of the core temperature signal [9]. This significantly reduces the effective prediction horizon whenever there is a steep rise in core temperature, which is the case when the temperature nears the hyperthermic threshold of 39 °C. To address this limitation, we investigated three alert algorithms that combine AR model predictions and associated PIs to determine whether the predicted core temperature exceeded a predefined threshold and compared their performance to the AR-model predictions alone.

B. Sequential Probability Ratio Test (SPRT)

Bayesian approaches, which use information from previous data (previous predictions in this case) and current observations (current predictions), offer a natural choice for this decision-making problem. The SPRT [20] is a Bayesian approach that considers increasing evidence from a sequence of observations to make a decision [21], [22]. Briefly, given a sequence of core-temperature samples $X_1, X_2, \ldots$, not necessarily independent, so that $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ is a normal Gaussian process with unknown mean μ_X and a given variance σ_X^2, the SPRT tests the null hypothesis (H_0) that $\mu_X = \mu_0$ against an alternative hypothesis (H_1) that $\mu_X = \mu_1$, where μ_0 and μ_1 denote the mean temperature values below and above the temperature threshold, respectively, with $\mu_0 < \mu_1$. If p_0 and p_1 are the probability density functions governing H_0 and H_1, respectively, then the observed likelihood ratio at decision time *t* (corresponding to the time point *n*) can be represented as $l_n = \prod_{k=0}^{K-1} \frac{p_1(X_{n-k})}{p_0(X_{n-k})}$, where *K* is the length of the sequence of samples being considered.

In order to apply the SPRT algorithm to our problem, we combined three predicted values, namely, the AR-model prediction ($\hat{y}_{n-k}$), the upper PI ($\hat{y}_{n-k} + \mathrm{PI}_{n-k}$), and the lower PI ($\hat{y}_{n-k} - \mathrm{PI}_{n-k}$), using weights θ and ϕ to form:

$$X_{n-k} = \hat{y}_{n-k} - \mathrm{PI}_{n-k}(1 - \theta\phi - \phi) \tag{3}$$

where $k = (0, 1, \ldots, K-1)$ denotes a time index and the weights θ and ϕ are constrained to be between 0 and 1. In (3), when $\phi = 0$, X_{n-k} equals the lower PI; when $\phi = 1$ and $\theta = 0$, X_{n-k} equals the AR-model prediction; and when $\theta = \phi = 1$, X_{n-k} equals the upper PI. Thus, X_n lies between the lower and upper PIs for all values of θ and ϕ. Note that a decision at time *t*, for a prediction window of *P* min, is actually made at time *t–P*, which corresponds to time point *n–M* for *M* steps-ahead AR-model predictions. Then, following Wald's SPRT methodology [20], we

$$\begin{aligned}
&\text{accepted } \mathrm{H}_0 \text{ (below temperature threshold)} \\
&\quad \text{if } \log(l_n) < \log(B); \text{ or} \\
&\text{accepted } H_1 \text{ (above temperature threshold)} \\
&\quad \text{if } \log(l_n) < \log(A); \text{ or} \\
&\text{made no decision and proceeded to time } n+1 \\
&\quad \text{if } \log(B) \le \log(l_n) \le \log(A)
\end{aligned} \tag{4}$$

where *A* and *B* are constants that control the false-positive rate and false-negative rate, respectively, with $0 < B < A < \infty$. The SPRT algorithm required the estimation of seven parameters, the two weights θ and ϕ in (3) and the five parameters μ_0, μ_1, σ_X, *A*, and *B*, from a subject's core-temperature data. A large difference between μ_0 and μ_1, with a small σ_X (which depends on the measurement noise), leads to good separability between temperature values below and above the threshold, respectively. This leads to a significant increase or decrease in $\log(l_n)$, depending on whether the AR model tracks a rise or fall in the core temperature data, respectively. A large difference between $\log(A)$ and $\log(B)$ ensures the reduction of false alerts while maintaining the sensitivity of the algorithm.

C. Proposed Alert Algorithms

As mentioned earlier, AR model predictions are significantly delayed during steep rises in core temperature. To address this problem and provide an alert before the core temperature reaches a specified threshold, we compared the performance of three algorithms, which use AR-model predictions and PIs, against the AR model. Of these algorithms, the SPRT is the only one that required estimation of additional parameters. All algorithms output either 0 or 1 (no alert or alert, respectively). The four algorithms are as follows:

1) *Model:* Uses the AR-model prediction at the current time instant to make a decision.
2) *Model+PI:* Uses the upper PI (AR-model predictions plus PI) at the current time instant to make a decision.
3) *Median Filter:* Outputs the median of a finite sequence of decisions made by Model+PI. The length of the sequence of decisions was fixed to an odd number (five) to ensure that the output was either 0 or 1.
4) *SPRT:* Outputs a decision based on (4) after collecting evidence from a series of AR-model predictions and PIs [see (3)]. For a given temperature threshold, we obtained the SPRT parameters by minimizing a cost function [see Appendix A, (A1)] formed as a composite of the four measures of performance described in Section II-D.

D. Measures for Evaluating the Performance of the Proposed Algorithms

To evaluate the proposed alert algorithms, we defined an "event" as an episode where the core temperature measurement rises and remains above a specified temperature threshold for ≥15 min [see Fig. 1(a)]. The event ends when the measured temperature decreases below the threshold and remains below the threshold for ≥15 min. Thus, an event was mapped into a 1 (true response) when the measured temperature was above the threshold and 0 otherwise. Similarly, a model-predicted event was defined as an episode where the algorithm's output (the model-predicted response) was 1 [see Fig. 1(a)].

We evaluated the algorithms' performance using four measures:

1) *Sensitivity*: the fraction of time points during which both the true and the model-predicted responses were 1 [Fig. 1(c); involving time spans A and B], including the time points up to at most 30 min prior to the true event onset when the model-predicted responses were 1 and the true responses were 0 [Fig. 1(c); scenario ii]. This allowed us to reward model-predicted events that arrived up to 30 min before the true event onset, as it takes >30 min for the core temperature to rise by 1 °C during exercise [23]. Sensitivity was expressed as a percentage of time points of the entire time series.
2) *Specificity*: the fraction of time points during which both the true and the model-predicted responses were 0 [Fig. 1(c) and (d); involving time spans C and D]. For computing specificity, we did not consider time points up to at most 30 min prior to the true event onset and time points immediately following the true event, when the

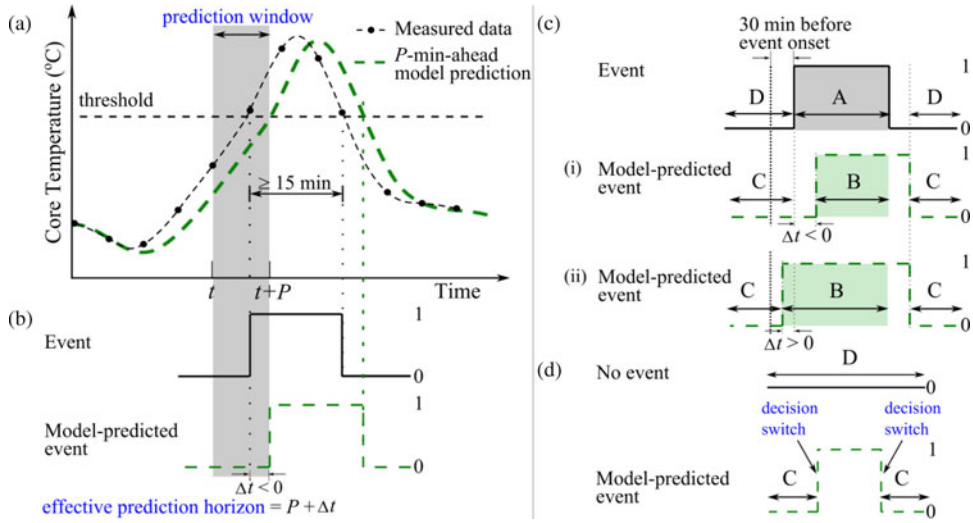

Fig. 1. Definition of a true event and the four measures of algorithm performance. These measures compare a true event, i.e., the episode when the core temperature rose above a certain temperature threshold, with the model-predicted event. (a) Measurement data (dotted line) and P-min-ahead AR-model predictions (dashed line). In this example, the AR-model used data until time point t to make a prediction at time point $t + P$. (b) Binary responses of the true event (solid line) and the model-predicted event (dashed line). The effective prediction horizon was defined as the prediction window P plus the time difference Δt between the onsets of the true and model-predicted events (see text for complete definition). (c) In scenario i, the prediction was delayed ($\Delta t < 0$) and in scenario ii, it anticipated the true event ($\Delta t > 0$). We computed sensitivity and specificity using time markers A, B, C, and D. Sensitivity was defined as the fraction of B included in max(A, B) ($<100\%$ in scenario i and 100% in scenario ii). Specificity was defined as the fraction of C included in max(C, D–Δt) if $\Delta t > 0$ and the fraction of C included in max(C, D) otherwise. (d) Scenario where there was no true event but the model predicted an event. In this case, we could only compute specificity and the number of decision switches; the other two measures were undefined.

model-predicted responses were 1 and the true responses were 0 [Fig. 1(c); scenario ii]. Specificity was expressed as a percentage of time points of the entire time series.

3) *Effective prediction horizon*: For each true event, we computed the effective prediction horizon by adding the prediction window P to the time difference Δt between the onset of the true event and the onset of the model-predicted event. When $\Delta t < 0$ [Fig. 1(b) and (c); scenario i], the onset of the true event preceded the prediction and the lower limit of Δt was set to $-P$ min. When $\Delta t > 0$ [Fig. 1(c); scenario ii], the onset of the model-predicted event preceded the true event and the upper limit of Δt was set to 30 min (as explained earlier). The reported effective prediction horizon was averaged over the number of events.
4) *Number of decision switches*: the cumulative number of times the model-predicted output transitioned from one state to another (0 to 1 or 1 to 0) that was incongruent with the true state (0 or 1) of the measured temperature [Fig. 1(d) shows two decision switches: the true state was 0 and the model-predicted output changed first from 0 to 1 and then from 1 to 0].

We computed specificity and the number of decision switches regardless of whether or not an event had occurred; however, we only computed sensitivity and effective prediction horizon when a true event occurred. Note that while specificity measured the time period for which the algorithm incorrectly predicted the occurrence of an event, the number of decision switches provided the number of times the algorithm predictions incorrectly switched from one state to another.

E. Study Data

To demonstrate the performance of the proposed algorithms, we used data from a field study involving 22 U.S. Army soldiers [age: 23.1 year (SD 4.1); height: 178 cm (SD 7); weight 81.3 kg (SD 11.1), mean and standard deviation (SD)] who performed regularly scheduled infantry training. The training included a 6-mile foot march while wearing a backpack and carrying equipment weighing on average 14.0 kg (SD 1.4) and exercises, such as digging of ditches, setting up concertina wire, marksmanship drills, running, rolling, and jumping as part of approach to a target. The training duration was 8–14 h in one day. During the training, soldiers wore the advanced combat uniform with a thermal insulation of 1.08 clo and an evaporative potential of 0.41 im/clo. The ambient temperature was 31.5 °C (SD 2.9) with a relative humidity of 66% (SD 16) and a wind speed of 9.9 km/h (SD 3.7). The Institutional Review Board of the U.S. Army Research Institute of Environmental Medicine (Natick, MA) approved the study. Subjects were briefed on the purpose, risks, and benefits of the study and each gave their written informed consent prior to study participation. All training was at the direction of the military unit, i.e., the research team did not interfere with or ask for any alteration to training events; they only monitored physiological parameters of the subjects. The core temperature data were measured using radio-thermometer pills (MiniMitter, Inc., Bend, OR) that transmitted the data to the Hidalgo Equivital EQ-02 (Hidalgo, Ltd., Cambridge, UK) physiological status monitoring (PSM) system [24]. Data were retrieved from the PSM system at the end of the exercise and subsequently analyzed. Pills were ingested at least 12 h before

data collection and had the following technical characteristics: size: 21.9 mm length and 8.5 mm diameter; weight: 1.75 g; sampling period: 15 s; temperature range: 25–50 °C, with accuracy of ±0.25 °C; transmission method: near-field magnetic link.

F. Cross Validation of the Algorithms

We performed a cross validation of the four alert algorithms using the study data as follows:

1) *Step 1:* Identified subjects that had core temperature values above a 38 °C temperature threshold for ≥ 15 min. Eighteen subjects met this criterion.
2) *Step 2:* Trained an AR model and estimated the SPRT algorithm parameters using data from one of the subjects who met the criterion in *Step 1*. Tested the algorithms on the remaining 21 subjects by computing the four measures of performance (sensitivity, specificity, effective prediction horizon, and the number of decision switches) for each algorithm. Repeated this step for each of the subjects who met the criterion in *Step 1*.
3) *Step 3:* Increased the temperature threshold by 0.1 °C and repeated *Steps 1* and *2*.

This procedure was repeated for temperature thresholds from 38.0 to 39.5 °C. There were five subjects with core temperature values beyond 39 °C (hyperthermic threshold) and three subjects with core temperature values beyond 39.5 °C (see Table III in Appendix B).

III. Results and Discussion

To apply the AR model to the core temperature data, we estimated the sampling period S and the model order m as follows. Analysis of the magnitude spectrum of the core temperature data led us to fix S to 5 min to preserve the important frequencies of the signal while rejecting noise. Accordingly, the temperature measurements were downsampled and reported every 5 min. Subsequently, we fixed m to 5 based on analysis of the partial autocorrelation function. For computing the PI in (2), we set $Z_{\alpha/2}$ to 2.78 based on the t-distribution for $\alpha = 5\%$ and $m = 5$ [18]. To further test the universal nature of our AR models [8], for each subject, we compared the root mean squared error (RMSE) between the AR-model fit and the measured data with the average RMSE between each of the other 21 AR-model predictions and the measured data for that subject. We found that the differences between the 22 pairs of RMSEs were not significantly different ($p = 0.55$, paired Wilcoxon signed-rank test) and the mean RMSE difference was 0.05 °C (SD 0.05 °C). For the SPRT algorithm, the memory length K was fixed to 3 (corresponding to a memory of 15 min for $S = 5$ min).

A. Performance of the Algorithms for a Single Temperature Threshold

Fig. 2 shows the performance of the four alert algorithms on a single subject based on a temperature threshold of 39 °C using a model trained on data from another subject (see Section II-F). The top panel shows the measured core temperature data (dotted line), 20-min-ahead AR-model predictions (dashed line), and 95% PIs (shaded area). The next four panels show the depiction of a true event (solid line) and algorithm decisions (dashed lines), for Model, Model+PI, Median Filter, and SPRT algorithms, respectively. The vertical thin dotted line in these four panels marks the 30-min point that precedes the true event onset.

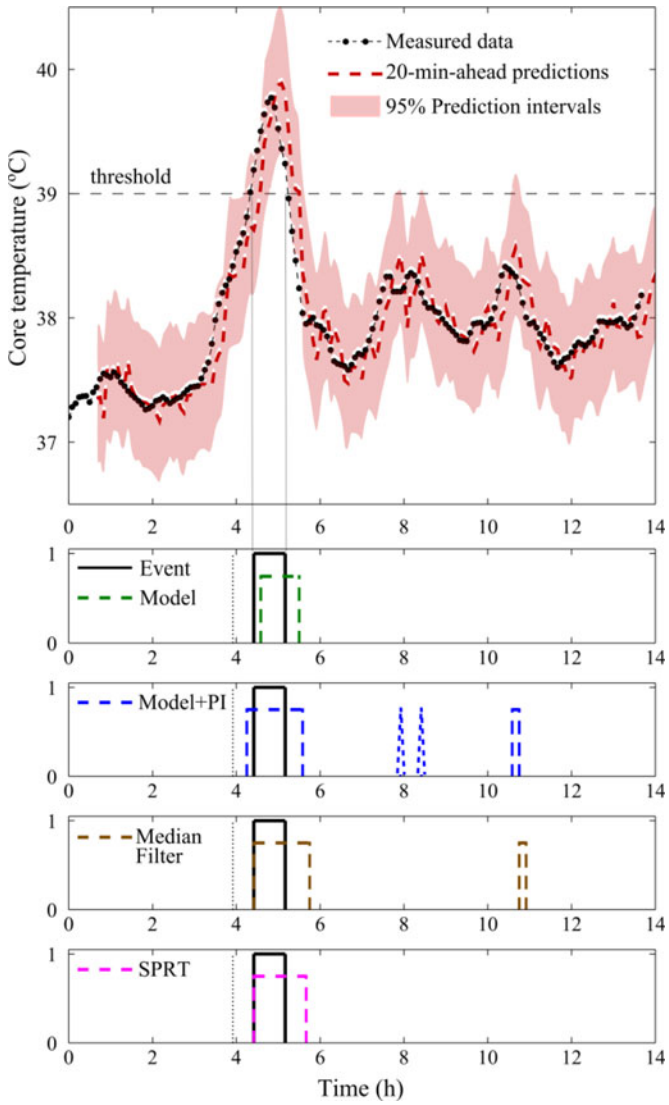

Fig. 2. Top panel shows the measured core temperature data from one subject, 20-min-ahead AR model predictions, and the corresponding PIs. The next four panels show the depiction of a true event (solid line) and algorithm decisions (dashed lines), for Model, Model+PI, Median Filter, and SPRT algorithms, respectively. The vertical thin dotted line in these four panels marks the 30-min point that precedes the true event onset.

As shown in Fig. 2 (top panel), we observed that the upper PI leads the measured data, whereas the model predictions lag the data. Hence, by construction, Model yielded the largest specificity, the smallest sensitivity, and the shortest effective prediction horizon compared with any of the other three algorithms. Model+PI yielded the largest sensitivity (100%) and the longest effective prediction horizon (30 min; the predicted event anticipated the actual event by 10 min beyond the 20-min

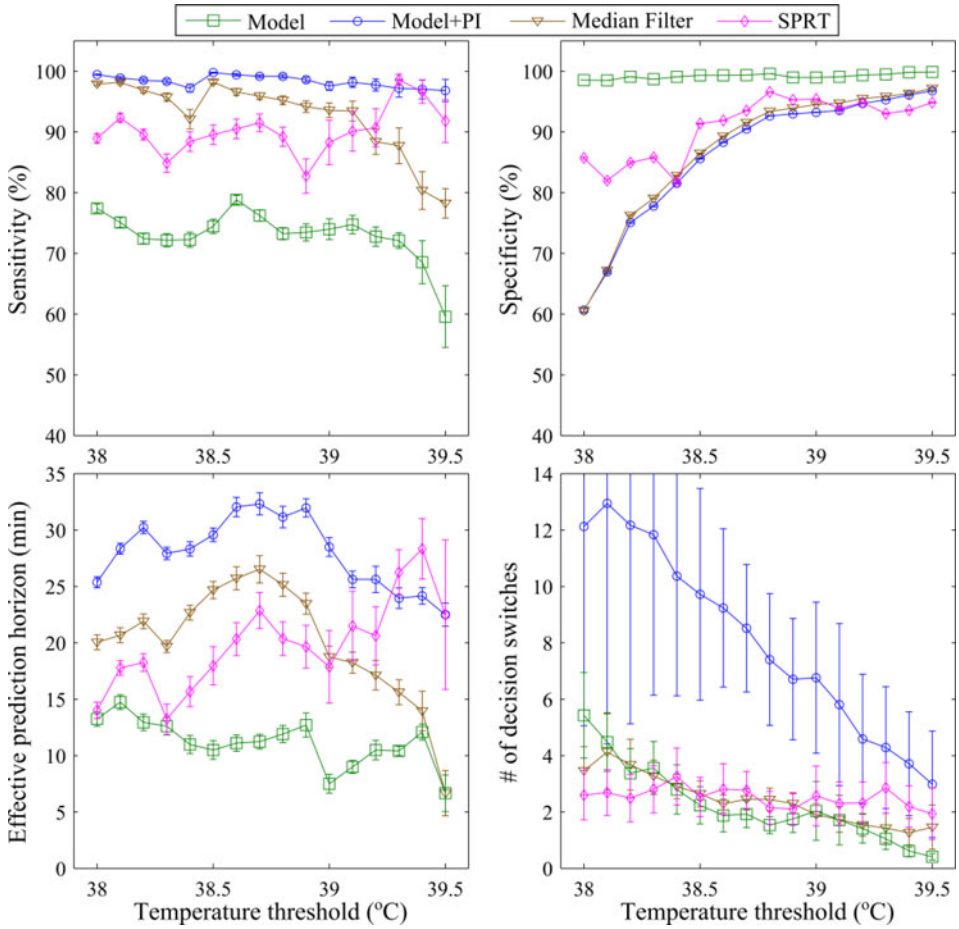

Fig. 3. Cross-validation study of the four algorithms across temperature thresholds from 38.0 to 39.5 °C. Plots show the mean (standard error) across the subjects for each of the four performance measures for 20-min-ahead AR-model predictions. Sensitivity and effective prediction horizon were defined only for subjects with events, i.e., those whose core temperature exceeded the corresponding temperature threshold.

prediction window). However, due to inherent noise in the data, Model+PI also yielded a large number of false alerts, leading to reduced specificity and a large number of decision switches compared to Model (specificity: 97% versus 100%; number of decision switches: 6 versus 0).

When compared to the decisions of Model+PI, Median Filter yielded fewer decision switches at the cost of a reduced effective prediction horizon (number of decision switches: 2 versus 6; effective prediction horizon: 20 min versus 30 min). SPRT had the same effective prediction horizon as Median Filter (20 min) but, importantly, had no decision switches (0 versus 2). A fourfold cross-validation study of five subjects whose temperature exceeded 39 °C in eight events (Table III) confirmed the aforementioned findings to hold true across our dataset (see Table I). In these five sets of computations, the range of the SPRT parameters obtained by optimizing (A1) was as follows: θ (0.80–0.81), ϕ (0.71–0.80), μ_0 (38.0–38.4), μ_1 (40.0–40.2), σ_X (0.20–0.25), $\log(A)$ (-89 to 76), and $\log(B)$ (-95 to 19). The large range of values for $\log(A)$ and $\log(B)$ was driven by one subject [#11; $\log(A) = 76$ and $\log(B) = 19$], who experienced temperature excursions just below the 39 °C threshold, and by the goal of the parameter optimization routine to reduce the number of false alerts, which tended to maximize the differences between these two parameters.

TABLE I
CROSS-VALIDATION ANALYSIS FOR 20-MIN-AHEAD AR-MODEL PREDICTIONS AT A TEMPERATURE THRESHOLD OF 39 °C

Algorithm	Sensitivity (%)	Specificity (%)	Effective prediction horizon (min)	Number of decision switches
Model	74.0 (1.7)	99.3 (0.3)	8 (1)	2
Model+PI	97.6 (0.7)	94.1 (1.3)	29 (1)	8
Median Filter	93.6 (1.2)	95.1 (1.3)	19 (1)	2
SPRT	88.3 (3.6)	96.3 (1.4)	18 (3)	2

Values are mean (standard error) from a fourfold cross-validation study of five subjects whose core temperatures exceeded 39 °C in eight distinct events (see Table III in Appendix B).

B. Performance of the Algorithms for Temperature Thresholds From 38.0 to 39.5 °C

Fig. 3 shows the overall cross-validation results of each of the four algorithms for 20-min-ahead AR-model predictions for

temperature thresholds between 38.0 and 39.5 °C, for every 0.1 °C increment. Due to varying numbers of subjects with core temperature values beyond a specified temperature threshold (see Appendix B, Table III), the number of subjects from which we computed sensitivity and effective prediction horizon varied. For all temperature thresholds, we computed sensitivity and effective prediction horizon using fewer subjects than we used to compute specificity and the number of decision switches, which was calculated for each of the 22 subjects.

Similar to the example in Fig. 2, Model (see Fig. 3, squares) yielded the smallest sensitivity, the shortest effective prediction horizon, and the largest specificity across all temperature thresholds. Model+PI (see Fig. 3, circles) yielded high sensitivity and a long effective prediction horizon, while producing a very large number of decision switches across all temperature thresholds. Median Filter's performance (see Fig. 3, triangles) was generally between that of Model+PI and Model across all measures. In contrast, the SPRT algorithm (see Fig. 3, diamonds) yielded varying performance across the temperature thresholds, partially because we estimated different SPRT parameters for each temperature threshold (see Section II-C).

In Fig. 3 (right panels), the specificity of Model+PI increased and the number of decision switches decreased with increasing temperature thresholds. This was because the sustained steep rise in the temperature data reduced the effects of measurement noise. This allowed the AR model to better track the data, leading to fewer prediction errors for Model+PI at higher temperature thresholds.

The SPRT algorithm yielded sensitivities around 90% for most temperature thresholds and yielded an overall increasing trend in the effective prediction horizon with increasing thresholds, especially beyond 39 °C. This is because at higher thresholds μ_1 (the mean temperature above the threshold) was significantly higher than μ_0 (the mean temperature below the threshold), leading to a wider separation between core temperature values above and below the temperature threshold and improved estimates of A and B [see (4)], which reduced the false-positive rate and the false-negative rate.

C. Performance of the Algorithms as a Function of the AR Model Prediction Window

To evaluate the performance of the algorithms as a function of prediction window of the AR model, we set P to 15, 20, and 30 min and repeated the cross-validation study of the four algorithms for hyperthermic temperature thresholds (39.0–39.5 °C). For all algorithms, sensitivity and specificity decreased, while the effective prediction horizon increased with increasing P (see Table II). For all algorithms, and in particular Model+PI, the number of decision switches tended to increase with increasing P. SPRT yielded effective prediction horizons comparable to Model+PI for $P \geq 20$ min (see Table II, $P = 20$: 23 versus 25; $P = 30$: 32 versus 30, all units in min) with a significantly lower number of decision switches.

The cross-validation studies also suggested that, in general, the SPRT parameters estimated from one subject's data could be directly applied to other subjects for the same temperature threshold. However, we obtained more robust parameter estimates, in particular for parameters A and B, when the training subject's data both exceeded the temperature threshold and hovered just below it. Using such a procedure to select subjects from which to estimate the SPRT parameters, as opposed to using all subjects whose temperature exceeded the threshold as performed here, would have improved the specificity of the SPRT algorithm in Tables I and II.

TABLE II
CROSS-VALIDATION STUDY OF THE FOUR ALGORITHMS AS A FUNCTION OF AR-MODEL PREDICTION WINDOW FOR HYPERTHERMIC TEMPERATURE THRESHOLDS (39.0–39.5 °C)

Algorithm	Sensitivity† (%)	Specificity* (%)	Effective prediction horizon† (min)	Number of decision switches*
15 min-ahead AR-model predictions				
Model	80.7 (2.6)	99.6 (0.0)	9 (1)	1
Model+PI	98.7 (0.7)	95.5 (0.0)	23 (1)	4
Median Filter	93.1 (1.9)	96.1 (0.0)	13 (1)	2
SPRT	93.8 (2.3)	95.6 (0.1)	18 (3)	2
20 min-ahead AR-model predictions				
Model	70.3 (2.5)	98.9 (0.0)	9 (1)	1
Model+PI	97.4 (1.2)	94.9 (0.0)	25 (1)	5
Median Filter	87.0 (2.2)	95.7 (0.1)	15 (1)	2
SPRT	92.7 (2.7)	94.3 (0.1)	23 (3)	2
30 min-ahead AR-model predictions				
Model	51.5 (2.6)	99.2 (0.0)	11 (1)	2
Model+PI	88.1 (1.8)	93.4 (0.1)	30 (2)	7
Median Filter	69.7 (3.3)	94.4 (0.1)	19 (2)	2
SPRT	87.8 (4.2)	91.6 (0.2)	32 (4)	3

Table shows values averaged over temperature thresholds from 39.0 to 39.5 °C.
*Values are mean (standard error) across all 22 subjects in the cross-validation study (see Table III in Appendix B).
†Values are mean (standard error) across subjects with events in the cross-validation study (see Table III in Appendix B).

IV. CONCLUSION

We found that, as expected, using PIs along with AR-model predictions increased the effective prediction horizon, enabling earlier detection of the onset of core temperature rise than otherwise possible using AR-model predictions alone. We also found that none of the three proposed alert algorithms was consistently superior in each of the four assessed measures of performance. While Model+PI yielded the largest sensitivity and the longest effective prediction horizon, it also yielded the largest number of decision switches. In contrast, delaying alert decisions—through filtering (Median Filter) or through accumulation of evidence (SPRT)—yielded a small number of decision switches at the cost of reduced sensitivity and effective prediction horizon.

Increasing the prediction window of the AR model increased the effective prediction horizon for all algorithms at the expense of a decreased sensitivity, a decreased specificity, and an increased number of decision switches.

For practical applications, we suggest the use of a 20-min AR-model prediction window and a temperature alert threshold set to 39 °C. This should provide a sufficiently high threshold to reduce spurious alerts, while allowing for useful lead time for proactive cooling interventions that could alter the course

TABLE III
NUMBER OF EVENTS FOR EACH SUBJECT AT DIFFERENT TEMPERATURE THRESHOLDS

Subjects	Temperature threshold (°C)															
	38.0	38.1	38.2	38.3	38.4	38.5	38.6	38.7	38.8	38.9	39.0	39.1	39.2	39.3	39.4	39.5
1	1	1	1	1	1	1	1	1	1	1						
2	3	3	3	3	3	3	3	3	2	1						
3	3	3	3	4	3	3	2	1	1	1						
4	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
5	1	1	1	1	1											
6	2	2	2	2	2	1										
7	3	3	3	3	3	3	3	3	3	3	3	2	2	2	1	1
8																
9	1															
10	2	1	1	1	1	1	1									
11	3	4	4	3	3	3	3	2	1	1	1	1	1			
12	1	1														
13																
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
15	3	3	2	1	1											
16																
17																
18	1	1	1													
19	3	3	2	1	1	1	1	1	1	1						
20	2	2	2	1	1	1										
21	4	4	3	2	1	1	1	1	1	1	1	1	1	1	1	1
22	1	1	1	1												
Total	36	35	32	27	24	21	18	15	13	12	8	7	7	6	5	4

of the clinical outcome [4]. We recommend two possible alert algorithms: When the intent is to prevent the occurrence (or reoccurrence) of heat injury at any cost, we recommend the use of Model+PI, which, in addition to its simplicity, provides extremely high sensitivity (∼98%) and a long effective prediction horizon (29 min). When spurious alerts are the main concern, we recommend the use of SPRT, which provides the smallest number of decision switches with reasonable overall performance.

A practical limitation of the proposed concept is the use of an invasive sensor to measure core temperature. We are currently developing an alternative approach that combines environmental and activity measurements with phenomenological and first-principle models to estimate core temperature and obviate the need for ingestion of a temperature pill. We are also in the process of integrating the Model+PI and SPRT algorithms into the Equivital EQ-02 sensor electronics module [24] (Hidalgo, Ltd.) and assessing the performance of the integrated system during military training exercises. Together with the work described here, these efforts shall lead to a hardware/software system for real-time alerting of an increasing core temperature and a reduced incidence of heat injuries.

APPENDIX A
OPTIMIZING THE SPRT ALGORITHM

The seven parameters of the SPRT algorithm (see Section II-C; the two weights θ and ϕ in (3) and the five parameters μ_0, μ_1, σ_X, A, and B) were estimated for a subject by minimizing the following cost function:

$$J = 10\,(1 - \text{Sensitivity}) + (1 - \text{Specificity}) + \left|\frac{t_d - 30 - t_m}{t_d - 30}\right| + \frac{\#\text{of decision switches}}{2\times(\#\text{ of events})} \quad \text{(A1)}$$

where t_d and t_m denote the time of event onset and the model-predicted event onset, respectively, both in minutes. We set a 30-min window before the event onset to force the optimization algorithm to provide the best-possible effective prediction horizon. The four measures, namely, sensitivity, specificity, effective prediction horizon, and number of decision switches, are described in Section II-D. The number of decision switches was normalized by twice the number of true events because each model-predicted event that is not a true event causes two decision switches. In other words, the fourth term in (A1) penalizes model-predicted events that were not true events. We used the Nelder–Mead simplex method implemented in MATLAB version 7.14 (function *fminsearch*) to estimate the SPRT parameters that minimize (A1). Our initial efforts at optimizing (A1) without weighting the cost for sensitivity led to poor effective prediction horizons. Hence, we performed several optimizations by weighting the sensitivity by 5, 10, 20, and 30 and found that for weights beyond 10 the algorithm performance did not improve substantially.

APPENDIX B
CROSS-VALIDATION STUDY TABLE

Table III shows the number of subjects whose core temperature crossed a temperature threshold (from 38.0 to 39.5 °C, with 0.1 °C increments) and triggered an event. It also shows the total number of distinct events for each temperature threshold.

DISCLAIMER

The opinions and assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U.S. Army or of the

U.S. Department of Defense. This paper has been approved for public release with unlimited distribution.

REFERENCES

[1] *Medical Surveillance Monthly Report: Heat Injury Issue*, Armed Forces Health Surveillance Center, Silver Spring, MD, USA, vol. 18, no. 3, pp. 6–8, Mar. 2011 [Online]. Available: http://www.afhsc.mil/viewMSMR?file = 2011/v18_n03.pdf, accessed on 11/21/2013

[2] T. P. Yeo, "Heat stroke: A comprehensive review," *AACN Clin. Issues*, vol. 15, no. 2, pp. 280–293, Jun. 2004.

[3] Y. Epstein and W. O. Roberts, "The pathopysiology of heat stroke: An integrative view of the final common pathway," *Scand. J. Med. Sci. Sports*, vol. 21, pp. 742–748, Dec. 2011.

[4] M. Brearley, "Crushed ice ingestion—A practical strategy for lowering core body temperature," *J. Mil. Veterans' Health*, vol. 20, no. 2, pp. 25–30, 2012.

[5] D. S. Moran, Y. Heled, L. Still, A. Laor, and Y. Shapiro, "Assessment of heat tolerance for post exertional heat stroke individuals," *Med. Sci. Monit.*, vol. 10, no. 6, pp. CR252–CR257, Jun. 2004.

[6] C. Byrne and C. L. Lim, "The ingestible telemetric body core temperature sensor: A review of validity and exercise applications," *Brit. J. Sports Med.*, vol. 41, no. 3, pp. 126–133, Mar. 2007.

[7] M. J. Buller, W. J. Tharion, S. N. Cheuvront, S. J. Montain, R. W. Kenefick, J. Castellani, W. A. Latzka, W. S. Roberts, M. Richter, O. C. Jenkins, and R. W. Hoyt, "Estimation of human core temperature from sequential heart rate observations," *Physiol. Meas.*, vol. 34, no. 7, pp. 781–798, Jul. 2013.

[8] A. Gribok, M. Buller, and J. Reifman, "Individualized short-term core-temperature prediction in humans using biomathematical models," *IEEE Trans. Biomed. Eng.*, vol. 55, no. 5, pp. 1477–1487, May 2008.

[9] A. Gribok, M. Buller, R. Hoyt, and J. Reifman, "A real-time algorithm for predicting core-temperature in humans," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 4, pp. 1039–1045, Jul. 2010.

[10] C. Chatfield, *Time-Series Forecasting*. London, U.K.: Chapman & Hall, 2002.

[11] Y. Lu, A. V. Gribok, W. K. Ward, and J. Reifman, "The importance of different frequency bands in predicting subcutaneous glucose concentration in type 1 diabetic patients," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 8, pp. 1839–1846, Aug. 2010.

[12] M. Khashei and M. Bijari, "An artificial neural network (p, d, q) model for timeseries forecasting," *Expert Syst. Appl.*, vol. 37, no. 1, pp. 479–489, 2010.

[13] N. Oleng, A. Gribok, and J. Reifman, "Error bounds for data-driven models of dynamical systems," *Comput. Biol. Med.*, vol. 37, no. 5, pp. 670–679, May 2007.

[14] B. W. Bequette, "Continuous glucose monitoring: Real-time algorithms for calibration, filtering, and alarms," *J. Diabetes Sci. Technol.*, vol. 4, no. 2, pp. 404–418, Mar. 2010.

[15] J. Reifman, S. Rajaraman, A. Gribok, and W. K. Ward, "Predictive monitoring for improved management of glucose levels," *J. Diabetes Sci. Technol.*, vol. 1, no. 4, pp. 478–486, Jul. 2007.

[16] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series," *IEEE Trans. Biomed. Eng.*, vol. 54, no. 5, pp. 931–937, May 2007.

[17] L. Marple, *Digital Spectral Analysis With Applications*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1987.

[18] A. V. Gribok, M. J. Buller, R. W. Hoyt, and J. Reifman, "Providing statistical measures of reliability for body core temperature predictions," in *Proc. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2007, pp. 545–548.

[19] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. London, U.K.: Chapman & Hall, 1993.

[20] A. Wald, "Sequential tests of statistical hypotheses," *Ann. Math. Statist.*, vol. 16, no. 2, pp. 117–186, 1945.

[21] L. Chen, A. T. Reisner, X. Chen, A. V. Gribok, and J. Reifman, "Are standard diagnostic test characteristics sufficient for the assessment of continual patient monitoring?" *Med. Decis. Making*, vol. 33, no. 2, pp. 225–234, Feb. 2013.

[22] M. Kulldorff, R. L. Davis, M. Kolczak, E. Lewis, T. Lieu, and R. Platt, "A maximized sequential probability ratio test for drug and vaccine safety surveillance," *Sequential Anal.: Des. Methods Appl.*, vol. 30, no. 1, pp. 58–78, Jan. 2011.

[23] C. A. Horswill, J. R. Stofan, S. C. Lovett, and C. Hannasch, "Core temperature and metabolic responses after carbohydrate intake during exercise at 30 degrees C," *J. Athletic Train.*, vol. 43, no. 6, pp. 585–591, Oct.–Dec. 2008.

[24] Y. Liu, S. H. Zhu, G. H. Wang, F. Ye, and P. Z. Li, "Validity and reliability of multiparameter physiological measurements recorded by the Equivital LifeMonitor during activities of various intensities," *J. Occupat. Environ. Hygiene*, vol. 10, no. 2, pp. 78–85, 2012.

Srinivas Laxminarayan received the B.E. degree in electronics engineering from the University of Pune, Pune, India, in 2000, the M.S. degree in electrical engineering from the Illinois Institute of Technology, Chicago, IL, USA, in 2003, and the Ph.D. degree in electrical engineering from Northeastern University, Boston, MA, USA, in 2011.

He is currently a Research Scientist in the Department of Defense Biotechnology High Performance Computing Software Applications Institute, Frederick, MD, USA. His current research interests include development of biomathematical models of physiology, time series analysis, computational neuroscience, signal processing, control theory, Bayesian schemes for parameter optimization, and machine learning.

Mark J. Buller received the B.Sc. (Hons.) degree in applied psychology from the University of Wales, College of Cardiff, Cardiff, U.K., in 1991, and the M.Sc. degree in computer science from Brown University, Providence, RI, USA, in 2008, where he is currently working toward the Ph.D. degree in computer science.

He is with the U.S. Army Research Institute of Environmental Medicine, Natick, MA. He has been involved as a Lead in the development of wearable physiological monitoring systems for more than ten years, and is a member of the North Atlantic Treaty Organization Research Technology Group "Real-Time Physiological and Psycho-Physiological Status Monitoring for Human Protection and Operational Health Applications." His research interests include understanding human health state in harsh environments using machine-learning techniques.

William J. Tharion received the B.S. and M.S. degrees in exercise science from the University of Massachusetts, Amherst, MA, USA, in 1980 and 1984, respectively, and the M.B.A. degree with a concentration in marketing and management from Northeastern University, Boston, MA, in 1994.

He has been a Human Factors Research Psychologist in the U.S. Army Research Institute of Environmental Medicine, Natick, MA, for 29 years. He is an author of more than 150 journal manuscripts and technical reports. His research interests include assessment of human factors issues associated with equipment developed for soldiers.

Mr. Tharion is a member of the Human Factors and Ergonomics Society and the American Physiological Society.

Jaques Reifman received the B.S. degree in civil engineering from Rio de Janeiro State University, Rio de Janeiro, Brazil, and the B.B.A. degree in business administration from Rio de Janeiro Federal University, Rio de Janeiro, in 1980 and 1985, respectively, and the M.S.E. and Ph.D. degrees in nuclear engineering from the University of Michigan, Ann Arbor, MI, USA, in 1985 and 1989, respectively.

He is currently a Senior Research Scientist in the Department of the Army, U.S. Army Medical Research and Materiel Command, Fort Detrick, MD, USA, where he is also the Founder of the Department of Defense Biotechnology High Performance Computing Software Applications Institute, Frederick, MD. His current research interests include physiological signal processing, statistical pattern recognition, artificial intelligence, data mining, biomathematical modeling, systems biology, bioinformatics, genomics, and proteomics.