Prospective Validation of *2B-Cool*: Integrating Wearables and Individualized Predictive Analytics to Reduce Heat Injuries

SRINIVAS LAXMINARAYAN[1,2], SAMANTHA HORNBY[1,2], LUKE N. BELVAL[3], GABRIELLE E. W. GIERSCH[3], MARGARET C. MORRISSEY[3], DOUGLAS J. CASA[3], and JAQUES REIFMAN[1]

[1]Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, Fort Detrick, MD; [2]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc., Bethesda, MD; and [3]Korey Stringer Institute, University of Connecticut, Storrs, CT

ABSTRACT

LAXMINARAYAN, S., S. HORNBY, L. N. BELVAL, G. E. W. GIERSCH, M. C. MORRISSEY, D. J. CASA, and J. REIFMAN. Prospective Validation of *2B-Cool*: Integrating Wearables and Individualized Predictive Analytics to Reduce Heat Injuries. *Med. Sci. Sports Exerc.*, Vol. 55, No. 4, pp. 751–764, 2023. **Introduction:** An uncontrollably rising core body temperature (T_C) is an indicator of an impending exertional heat illness. However, measuring T_C invasively in field settings is challenging. By contrast, wearable sensors combined with machine-learning algorithms can continuously monitor T_C nonintrusively. Here, we prospectively validated *2B-Cool*, a hardware/software system that automatically learns how individuals respond to heat stress and provides individualized estimates of T_C, 20-min ahead predictions, and early warning of a rising T_C. **Methods:** We performed a crossover heat stress study in an environmental chamber, involving 11 men and 11 women (mean ± SD age = 20 ± 2 yr) who performed three bouts of varying physical activities on a treadmill over a 7.5-h trial, each under four different clothing and environmental conditions. Subjects wore the *2B-Cool* system, consisting of a smartwatch, which collected vital signs, and a paired smartphone, which housed machine-learning algorithms and used the vital sign data to make individualized real-time forecasts. Subjects also wore a chest strap heart rate sensor and a rectal probe for comparison purposes. **Results:** We observed very good agreement between the *2B-Cool* forecasts and the measured T_C, with a mean bias of 0.16°C for T_C estimates and nearly 75% of measurements falling within the 95% prediction intervals of ±0.62°C for the 20-min predictions. The early-warning system results for a 38.50°C threshold yielded a 98% sensitivity, an 81% specificity, a prediction horizon of 35 min, and a false alarm rate of 0.12 events per hour. We observed no sex differences in the measured or predicted peak T_C. **Conclusion:** *2B-Cool* provides early warning of a rising T_C with a sufficient lead time to enable clinical interventions and to help reduce the risk of exertional heat illness. **Key Words:** CORE BODY TEMPERATURE, HEAT STRESS, INDIVIDUALIZED PREDICTIONS, WEARABLES

Core body temperature (T_C) is recognized as the best single physiological predictor of an impending risk of exertional heat illness (1). However, there is no clear temperature threshold that delineates the transition from a low-risk to a high-risk condition because several important factors, such as hydration level, heat acclimation state, and environmental conditions, modulate this risk (2). For example, Sandell et al. found that marathon and ultramarathon distance runners who collapsed from heat exhaustion during or after a race reported T_C values ranging from 38.00°C to 40.00°C (3,4), whereas Pugh et al. found that the rectal temperature of a marathon winner exceeded 41.00°C (5), where a $T_C > 40.60$°C has been linked to life-threatening heat stroke (6). In fact, Montain et al. (7) found that heat exhaustion occurs over a broad range of T_C values, with no one specific threshold above which exhaustion abruptly increases. Despite such large individual differences in response to exertional heat stress, multiple studies have identified robust population-level relationships between T_C and exhaustion from heat strain during exercise (1,4). For example, in a large meta-analysis involving 747 laboratory studies and 131 field trials of military exercises, Sawka et al. (1) reported that 50% of subjects incurred exhaustion at a T_C of 38.60°C and 39.50°C, respectively.

Address for correspondence: Jaques Reifman, Ph.D., Department of Defense Biotechnology High Performance Computing Software Applications Institute, Telemedicine and Advanced Technology Research Center, U.S. Army Medical Research and Development Command, ATTN: FCMR-TT, 504 Scott Street, Fort Detrick, MD 21702; E-mail: jaques.reifman.civ@health.mil.
Submitted for publication June 2022.
Accepted for publication November 2022.
Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.acsm-msse.org).

0195-9131/23/5504-0751/0
MEDICINE & SCIENCE IN SPORTS & EXERCISE®

DOI: 10.1249/MSS.0000000000003093

The risk of exertional heat illness could be mitigated by continuously monitoring T_C during certain physical heat stress activities, such as sports and military training. However, rectal probes and ingestible temperature pills, the gold standard devices for measuring T_C, are invasive (8) and often not conducive for field monitoring. By contrast, the integration of vital signs collected from nonintrusive wearable sensors embedded in commercial-off-the-shelf (COTS) products routinely used in everyday life (9) with customized algorithms allows for the development of predictive analytical tools to help prevent undesirable outcomes due to exertional and environmental heat stress conditions.

Recently, a handful of approaches have been proposed to integrate vital signs collected from COTS wearable devices with customized algorithms to estimate T_C as an alternative to invasive measurements (10–15). These approaches use a variety of noninvasive physiological variables collected from wearables, ranging from a single heart rate (HR) variable (11) to a few vital signs (HR as well as skin temperature and skin heat flux at multiple body locations) (14) to systems requiring these plus environmental conditions (10,12,15) and anthropometric data (13). In terms of algorithms, some approaches use data-driven, machine-learning (ML) algorithms, such as regression analysis (14,15), others use ML algorithms in the form of an extended Kalman filter, where the relationships between vital signs and T_C are either represented by empirical correlations (11,13) or by physics-based, energy-balance mathematical models with varying complexities (10,12).

Although these approaches show promise, yielding root-mean-square errors (RMSE) between algorithm-estimated T_C and rectal- or gastrointestinal-measured temperatures of <0.50°C, the studies used for their validation have certain limitations: 1) only involved men and used very small sample sizes (10,11); 2) consisted of physical work intensities with short durations (~120 min or less), resulting in nearly monotonic increases in T_C (10–15), which are less challenging to estimate than sharp rises and drops in T_C; 3) the bulk or all measured T_C never exceed 38.50°C (10,11,14), limiting or negating the relevance of the study for heat stress management (15); 4) they assessed the results using a leave-one-out procedure (13,15), which overestimates performance; 5) with one exception (12), they cannot account for between-subject variability; 6) all studies, except for one (11), involved retrospective data analysis, as opposed to a prospective, real-time assessment that more closely resembles actual operational use; and 7) all studies provide estimates of instantaneous T_C ($\hat{T}_C$), lacking the ability to forecast temperature values into the near future to allow for proactive interventions.

Over the years, our U.S. Army group has developed an algorithm that automatically learns how individuals respond to heat stress and provides individualized $\hat{T}_C$ based on vital signs and environmental conditions (12); an autoregressive predictive algorithm that forecasts future values of T_C based on a time series of previous T_C values (16); and a probabilistic algorithm that provides early warnings of an impending rise in T_C beyond undesirable threshold levels (17). To date, we developed and retrospectively validated these three algorithms separately, using previously collected laboratory-grade vital sign sensor data (12,16,17), constant work intensities of short duration and monotonic rises in T_C (12), and a handful of trials where T_C exceeded 39.00°C (17). For example, because we developed the $\hat{T}_C$ estimation algorithm (12) after the prediction and early-warning algorithms (16,17), we assessed the performance of the latter algorithms using a time series of measured T_C, as opposed to $\hat{T}_C$. Hence, we do not know how inaccuracies in the estimation of T_C propagate and affect the ability to provide early warnings.

We have now merged these three algorithms into one software system and integrated it with COTS wearable devices to form the *2B-Cool* system, consisting of a smartwatch that collects vital signs and wirelessly transmits them via Bluetooth to a smartphone, which houses the software and continuously provides individualized 1) values of $\hat{T}_C$, 2) 20-min ahead predictions of T_C, and 3) early warnings. The main objective of this study is to assess the performance of *2B-Cool* to provide an early warning of an impending rise in T_C beyond a threshold of 38.50°C, which is associated with exertional heat illnesses (1,7,12). In particular, we desire to assess the performance of *2B-Cool* during relatively long periods (~8 h) of intermittent rest and work cycles of moderate- to high-intensity levels to challenge the algorithm and to assess whether the use of low-cost commercial wearable devices affects its performance. Other objectives are to determine whether *2B-Cool* yields a similar performance under different environmental and clothing conditions and whether there is a difference in performance between men and women. To address these objectives, we performed a prospective, controlled heat stress laboratory study involving 11 men and 11 women who performed three bouts of frequently fluctuating activity levels on a treadmill over a 7.5-h trial, each under two different environmental conditions and two different clothing types.

METHODS

Study Data

We conducted an exertional heat stress study in an environmentally controlled chamber at the University of Connecticut, where we recruited 22 healthy, young men and women (11 each): age = 20.31 ± 2.01 (18–25) yr, height = 1.70 ± 0.11 (1.55–1.94) m, weight = 67.25 ± 13.56 (47.84–93.30) kg, body mass index = 22.97 ± 2.60 (19.16–28.65) $kg \cdot m^{-2}$, and $\dot{V}O_{2max}$ = 46.81 ± 5.28 (36.34–56.54) $mL \cdot kg^{-1} \cdot min^{-1}$; mean ± SD (range, min–max). We screened the subjects for health problems, prior heat illness, medications, and pregnancy, which could affect the study results. The participating women were normally menstruating (no birth control contraceptives) and were tested in the follicular phase of their menstrual cycle based on self-reporting. Each subject provided a written consent before their participation in the study, which was approved by the Institutional Review Board of the University of Connecticut (Storrs, CT) and by the U.S. Army Human Research Protection Office (Fort Detrick, MD).

The study followed a randomized crossover design consisting of four trials, where for each ~7.5-h trial, subjects exercised at varying intensities on a treadmill inside the chamber while wearing one of two different clothing types (T-shirt and shorts [TSS] or active combat uniform [ACU]) under one of two different environmental conditions: hot and humid (ambient temperature [T_A] = 30°C and relative humidity [H] = 60%) or very hot and dry (T_A = 36°C and H = 30%). Hence, each subject participated in four trials (two clothing types × two environmental conditions). We selected these environmental conditions to balance the likelihood that subjects would complete each of the four 7.5-h trials and reach a T_C of at least 38.50°C during each trial. For each trial, we measured the environmental conditions using a Kestrel 5400 environmental monitoring device (Nielsen-Kellerman, Boothwyn, PA) and manually entered the data into the smartphone app. We selected the two clothing types (TSS [clo = ~0.11] and ACU [clo = ~0.86]) to simulate military training (TSS) and combat (ACU) scenarios, respectively. Women wore a sports bra in addition to the other clothing (clo = ~0.044) (18). Subjects started each trial at approximately 8:30 AM with a 60-min rest period in the environmental chamber, after which they performed three identical exercise bouts of frequently fluctuating moderate- to high-intensity levels, each lasting for 80 min followed by a 50-min rest (Fig. 1). We selected this relatively long 7.5-h trial with multiple changes in work-intensity levels to challenge *2B-Cool* and assess its ability to capture sharp rises and drops in the temporal dynamics of T_C, which would more closely resemble military training activities (17) than the 2.0-h constant work–intensity trials typically used to assess $\hat{T}_C$ algorithms (10–15).

Upon subject arrival to the laboratory, we used a handheld refractometer to collect a urine sample to confirm euhydration and allowed the subject to continue to the trial if the urine-specific gravity (Model TS400; Reichert Inc., Depew, NY) was ≤1.025. We asked subjects to refrain from strenuous physical activity 24 h before the start of the trial, and to abstain from alcohol and caffeine for 24 and 12 h, respectively. Before data collection, subjects performed a maximal oxygen consumption ($\dot{V}O_{2max}$) test to assess cardiorespiratory fitness and prescribe exercise intensity during each trial (moderate intensity = 30%–40% $\dot{V}O_{2max}$; high intensity = 70%–80% $\dot{V}O_{2max}$). To compute activity intensity, we mapped $\dot{V}O_2$ values into metabolic equivalent units (MET; 1 MET = $\dot{V}O_2$/3.5). The $\dot{V}O_{2max}$ test consisted of incremental stages, where the treadmill speed was increased every 3 min (0.5–1.0 mph) until the subject reached volitional exhaustion. We continuously collected $\dot{V}O_2$ and respiratory exchange ratio using a PARVO metabolic cart (TrueOne 2400; Parvo Medics, Salt Lake City, UT). We required subjects to obtain a $\dot{V}O_{2max}$ of 45 mL·kg^{-1}·min^{-1} (men) or 40 mL·kg^{-1}·min^{-1} (women) to participate in the study.

In all trials, a fan installed in front of the treadmill provided wind at a speed of 2.5 m·s^{-1}, and subjects consumed water *ad libitum* provided from a standard water fountain (temperature of approximately 17.00°C). Subjects ate a standardized breakfast (bagel, 250 kcals; two tablespoons peanut butter, 190 kcals; and a large banana, 120 kcals) and lunch (two slices of white bread, 160 kcals; two tablespoons of peanut butter, 190 kcals; one tablespoon of grape jelly, 50 kcals; and one CLIF BAR, 240 kcals) during each trial. Subjects arrived fasted to control for nutritional intake. For each of the four trials, subjects wore a Samsung Gear S3 smartwatch (Samsung Electronics America, Ridgefield Park, NJ) on the wrist of their nondominant arm, which measured physical activity (A_C) via a three-axis accelerometer at a sampling rate of 25 Hz (19), HR at 1 Hz, and skin temperature (T_S) at 1 Hz. The smartwatch was paired with a Samsung Note 4 smartphone, which housed the software and used the vital sign data to make real-time forecasts. For comparison, subjects also wore a Polar H7 chest strap HR sensor with a 1-Hz sampling rate (Polar Electro Oy, Kempele, Finland); a thermistor (Biopac Systems Incorporated, Santa Barbara, CA) to continuously measure T_S on the neck, shoulder, back, abdomen, chest, thigh, and calf; and a self-inserted, calibrated rectal probe (YSI 400 series probe; Measurement Specialties, Hampton, VA) at 10–15 cm past the anal sphincter to measure T_C at a sampling rate of 250 Hz, utilizing a continuous physiological monitoring system (Biopac Systems Incorporated).

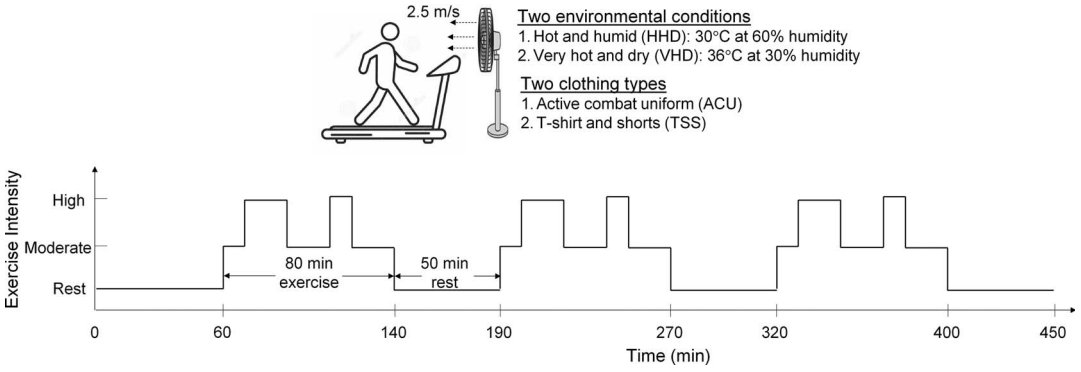

FIGURE 1—Schematic of the exercise protocol for the study conducted at the University of Connecticut. Subjects started each trial at approximately 8:30 AM, with a 60-min rest period. They then performed frequently fluctuating moderate- to high-intensity physical activities (walked or ran on a treadmill with a fan in front providing wind at 2.5 m·s^{-1}) for 80 min, followed by 50 min of rest, which was repeated two more times, for a total trial time of 450 min. Each subject repeated this protocol under two clothing types in two different environmental conditions, for a total of four trials.

Description of the *2B-Cool* System

The *2B-Cool* system is composed of COTS hardware and customized software components. The hardware consisted of a Samsung Gear S3 smartwatch paired with a Samsung Note 4 smartphone, both of which were specially configured for this study. We selected the Samsung Gear S3 because it provides direct access to the sensor data, without requiring third-party involvement or upload to a cloud service, and because we previously carried out an in-house assessment of its sensors' accuracy (12,20). In this assessment, healthy volunteers simultaneously wore both the Samsung Gear S3 and the gold standard devices under everyday ambulatory conditions, where we found that the median absolute difference was ≤5 bpm against the Polar H7 HR sensor, ≤1.30°C against the iButton skin temperature sensor (iButtonLink LLC, Whitewater, WI), and ≤0.032*g* against the ActiGraph wGT3X-BT accelerometer device (ActiWatch LLC, Pensacola, FL) (20).

We uploaded an in-house–developed software onto the two paired devices to control for the continuous wireless transmission of vital signs (A_C, HR, and T_S) from the watch to the phone via a Bluetooth protocol. Once transmitted to the smartphone, we preprocessed and averaged the vital signs to form a time series of 15-s values (see Appendix A, Supplemental Digital Content, Preprocessing of activity, heart rate, and skin temperature data, http://links.lww.com/MSS/C749). The smartphone also housed the *2B-Cool* software, which ran the three individualized algorithms in real time to 1) compute $\hat{T}_C$, 2) predict T_C 20 min into the future, and 3) provide early warning of an impending rise in T_C beyond undesirable threshold levels. During the study, we stored all raw sensor data and algorithm-produced results in the smartphone, as they were generated, and later retrieved them for analysis, as reported herein. Subjects were blinded to the *2B-Cool* results to ensure that they did not influence their performance.

Individualized *2B-Cool* Algorithms

Briefly, the $\hat{T}_C$ estimation algorithm (12) consists of a simplified thermoregulatory model coupled to an ML Kalman filter (21). The thermoregulatory model relates an individual's measured vital signs (A_C, HR, and T_S) and two environmental variables (T_A and H) to T_C, and the Kalman filter adapts the model parameters to customize the T_C estimates so as to reflect the individual's measurements (Fig. 2A). The model is composed of a phenomenological component, which relates A_C to HR via equation 1 in Table B1 (see Supplemental Digital Content, Individualized estimation and prediction algorithms, http://links.lww.com/MSS/C749) and a first-principles, macroscopic energy-balance component, which regulates the heat transfer from the core body to the skin and from the skin to the environment via equations 2 and 3, respectively. The model consists of six adjustable parameters and one fixed heat transfer parameter (α_2). The ML Kalman filter algorithm uses the two environmental variables and the three vital signs to continuously adjust these six parameters so as to customize the T_C estimates to the individual.

2B-Cool generates individualized $\hat{T}_C$ through three major steps. In step 1, the thermoregulatory model takes as inputs T_A, H, and A_C at the current time t_i to estimate HR and T_S (marked by an "o" in Fig. 2B) at time t_i. The algorithm then computes the errors between the estimated and the measured HR and T_S values (errors 1 and 2, respectively, in Fig. 2B) and uses them as inputs to the Kalman filter in step 2. In turn, the filter uses these errors to adapt the six adjustable model parameters, which are then used by the updated model in step 3 to estimate an individualized T_C at t_i (Fig. 2C). Through this process, which repeats itself every 15 s after each new set of vital sign measurements, T_C is continuously customized to the individual to capture the individual's response to exertional heat stress and environmental conditions, as represented by the measurements. We refer the reader to Laxminarayan et al. (12) for additional details.

The prediction algorithm uses a time series of $\hat{T}_C$ values to predict T_C 20 min ahead (17). Briefly, the prediction algorithm uses five values of $\hat{T}_C$ spanning 20 min (i.e., the value at the current time plus four previous values 5 min apart) to make a 20-min ahead prediction and to compute the corresponding 95% prediction interval (PI). For this purpose, the algorithm uses a fifth-order autoregressive model to iteratively predict T_C using equation 4 in Table B1 (see Supplemental Digital Content, Individualized estimation and prediction algorithms, http://links.lww.com/MSS/C749) and compute the associated PI through equation 5 (17). We repeated this process every 15 s to generate a time series of predicted T_C 20 min into the future. We refer the reader to Laxminarayan et al. (17) for additional details.

Finally, to provide early warnings of an impending T_C rise, we used a probabilistic algorithm to accumulate evidence and determine whether the predicted T_C crossed prespecified thresholds associated with an increased risk of exertional heat illnesses. Briefly, as previously reported (17), we combined the predicted T_C and its corresponding 95% PI as inputs to the sequential probability ratio test (22), which assesses this evidence over a 10-min window, to determine whether the accumulated evidence was sufficient to ascertain that the predicted T_C had crossed any of two prespecified thresholds, 38.50°C and 39.20°C. These thresholds correspond to 25% and 75% of individuals, respectively, who reached exhaustion while performing physical activities (7), and indicate low (amber, for >38.50°C) and high risk (red, for >39.20°C). This algorithm has the net effect of increasing the value of the predicted T_C to account for inherent time lags in the predictions (17).

Statistical Analyses

In our prospective laboratory study, *2B-Cool* continuously received vital sign data from the subject, and every 15 s computed $\hat{T}_C$ for the current time, predicted T_C 20 min into the future, and provided amber or red alerts when the predicted T_C was expected to exceed a threshold. We assessed the *2B-Cool*-generated results through the following metrics.

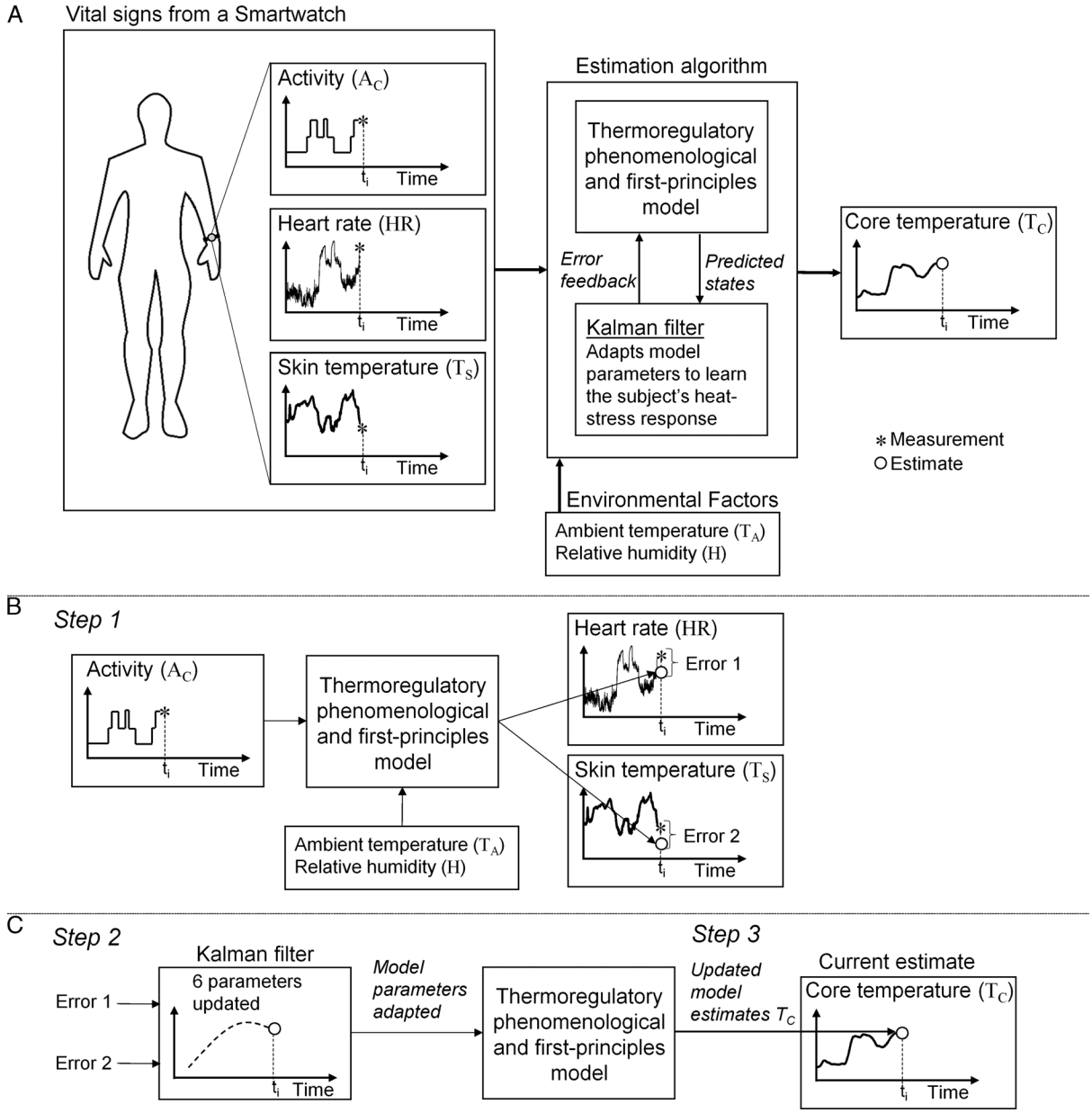

FIGURE 2—Visual representation of the core temperature estimation algorithm, consisting of a thermoregulatory model coupled to an ML Kalman filter. A, The algorithm takes physical activity (A_C), HR, and skin temperature (T_S) measurements (marked by an "∗") collected from an individual and two environmental variables (ambient temperature [T_A] and relative humidity [H]) at the current time point t_i to estimate T_C at t_i. B, In step 1, the thermoregulatory model takes A_C, T_A, and H to estimate HR and T_S (marked by an "o"), from which the algorithm computes errors between the estimates and the corresponding measurements. C, In step 2, the Kalman filter uses these errors to adjust six parameters of the thermoregulatory phenomenological and first-principles model to reflect the individual's response to heat stress and, by doing so, individualizes the model. In step 3, the updated model provides instantaneous T_C estimates ($\hat{T}_C$) at the current time t_i.

Metrics for $\hat{T}_C$. We computed mean bias, mean absolute error (MAE), mean absolute percentage error (MAPE), and RMSE. For each of these metrics, for a given subject and experimental condition (i.e., trial), we compared the measured rectal temperature T_C and $\hat{T}_C$ every 15 s over the entire trial and averaged the results over the number of comparisons in the trial. We defined mean bias as the mean difference between the measured T_C and $\hat{T}_C$ ($\hat{T}_C$ minus measured T_C); MAE as the mean absolute difference between the measured T_C and $\hat{T}_C$; MAPE as the mean of the normalized absolute difference between the measured T_C and $\hat{T}_C$, where we divided each difference by the measured T_C and multiplied the mean by 100; and RMSE as the square root of the mean of the sum of the squared differences between T_C and $\hat{T}_C$. We also computed Pearson's correlation coefficient r for each trial and averaged the results over the number of trials. To assess *2B-Cool*'s accuracy in estimating the timing and magnitude of the peak rectal temperature measurement (T_{Cmax}) in each of the three exercise bouts of a trial, we computed the following statistics: 1) the absolute difference between T_{Cmax} and $\hat{T}_C$ at the time of T_{Cmax}, 2)

the absolute difference between T_{Cmax} and $\hat{T}_{Cmax}$, 3) the absolute time difference between T_{Cmax} and $\hat{T}_{Cmax}$, and 4) the 95% limits of agreement (LOA) of the difference between T_{Cmax} and $\hat{T}_C$ at the time of T_{Cmax}, using the analysis suggested by Bland and Altman (23,24), which accounts for the relationship between difference and magnitude as well as for repeated measurements.

Metrics for T_C predictions. To assess *2B-Cool*'s 20-min ahead predictions, we computed the RMSE between the predicted and the measured T_C as described above and the fraction of measured T_C values that fell within the 95% PI. We computed the fraction over an entire trial and averaged the results over the number of comparisons in the trial.

Metrics for early-warning alerts. To evaluate the ability of the system to provide early-warning alerts, we followed the approach developed in our previous study (17). Briefly, we labeled the entire 7.5-h duration of each trial with a "true" binary time series, with "1" denoting the time points when the measured T_C rose and stayed above a temperature threshold (38.50°C or 39.20°C) for at least 5 min, indicating a "true event," and "0" for the remaining time points. Similarly, we labeled the entire trial with a "predicted" binary time series based on the decisions of the probabilistic algorithm and then assessed *2B-Cool*'s ability to predict true events by computing four performance metrics (17):

1. Sensitivity: the fraction of time during which the true and predicted responses were 1.
2. Specificity: the fraction of time during which the true and predicted responses were 0, when a true event did not occur.
3. Effective prediction horizon for true events: 20 min plus the time difference between the onset of a true event and the onset of a predicted event. When the predicted event lagged behind the true event, the effective prediction horizon was reduced by that time lag.
4. False alarm rate: the number of incorrect algorithm-predicted transitions from 0 to 1 per hour.

For computing specificity, we discounted the times up to 30 min before and after a true event when the true response was 0 and the predicted response was 1 so as not to penalize the algorithm when the event did occur and the algorithm predicted it, but there was not an exact match in the duration of the event (17). This definition has no impact on specificity when the algorithm incorrectly predicted an event that did not occur.

Acceptance criteria. Because *2B-Cool*'s main goal is to provide an early warning of an impending rise in T_C beyond thresholds of clinical relevance, we desire high sensitivity and specificity for such events. However, there are no standard acceptance criteria for these statistics. Hence, we set the acceptance criteria for sensitivity and specificity to >90% because this was the value achieved in our previous work (17). It should be more challenging for *2B-Cool* to reach this threshold because here we used a time series of estimated T_C, as opposed to a time series of measured T_C used in our previous work. $\hat{T}_C$ estimates have errors that propagate through the iterative 5-min predictions and make it harder to correctly predict an impending event. For early warnings, we set the acceptable criterion for the effective prediction horizon to >20 min. This should provide sufficient lead time to intervene and help avoid an exertional heat illness because for exercise-induced increases in body temperature, it takes >30 min for T_C to rise by 1.00°C (25). Similar to Casa et al. (26) and Goodman et al. (27), we used a mean bias smaller than ±0.27°C as the acceptance criterion for $\hat{T}_C$. Goodman et al. argued that this value is larger than the precision of temperature sensors but still small enough to allow for the detection of T_C differences associated with physiological and psychological consequences (28) as well as differences associated with circadian (29) and ovulatory (30) T_C rhythms of ~0.50°C.

Statistical differences. For T_C, we assessed statistical differences in mean bias, MAE, and RMSE between each pair of experimental conditions, and between men and women for each condition, by using the Wilcoxon rank-sum test, a nonparametric statistical test that compares two paired groups without requiring the data in the groups to be normally distributed. This test is recommended over the *t* test when the population characteristics are unknown (31), which is our case. The null hypothesis was that the median difference between the two groups was the same, which we rejected when the *P* value was <0.05. We used the same test to assess statistical differences between T_{Cmax} and $\hat{T}_{Cmax}$ for each of the three exercise bouts for each experimental condition. Similarly, we tested whether T_{Cmax} was different between men and women for each of the three exercise bouts for each experimental condition. For the early-warning system, we used the test to assess differences in each of the four metrics (sensitivity, specificity, effective prediction horizon, and false alarm rate), for each of the two temperature thresholds, between each pair of experimental conditions and between men and women for each condition. Likewise, we used the Wilcoxon rank-sum test to assess statistical differences in RMSE between men and women of HR measured by the Polar H7 versus those from the Samsung smartwatch, for each experimental condition.

Power calculation. To determine the number of subjects for the study, we performed a sample size calculation as part of the Institutional Review Board protocol. We used the equality formula proposed by Chow et al. (32), where the null hypothesis was that the difference between the $\hat{T}_C$ and the measured T_C exceeded a given RMSE and the alternate hypothesis was that the difference was smaller than the RMSE. Considering the range of RMSE for measured T_C between 38.50°C and 39.00°C from our previously work (12), we estimated effect sizes between 0.53 and 0.78 and determined that we would need between 14 and 32 subjects to reject the null hypothesis with an 80% power at a 5% significance level. We recruited 22 subjects, which is the mid-point range for the number of the sample size estimation. We used MATLAB (MathWorks, Natick, MA) version 9.7 R2019b or 9.12 R2022a to perform all statistical analysis calculations.

RESULTS

In the study, we collected data from 88 trials (22 subjects-4 conditions each). However, we could not use the data from six trials because battery problems with the smartwatch prevented us from collecting vital signs in two cases, a building fire alarm interrupted data collection in one case, and we observed problems with the measured rectal temperature in three cases (unreliable values or large measurement gaps). Hence, we assessed *2B-Cool* for a total of 82 trials, which elicited average values of minimum and maximum HR of 72 and 185 bpm, T_S of 31.00°C and 36.50°C, and T_C of 36.70°C and 39.10°C, respectively. We measured HR and Ts with the smartwatch and T_C with the rectal probe.

Performance of the estimation algorithm. Figure 3 shows the measured T_C (dotted blue line), the *2B-Cool*-estimated $\hat{T}_C$ (solid red line), and the corresponding activity levels (solid gray line) for a representative man (subject 4) and woman (subject 16). The RMSE between T_C and $\hat{T}_C$ across the eight trials ranged from 0.23°C to 0.78°C, whereas in some cases *2B-Cool* underestimated T_C (Fig. 3D and E) and in other cases it overestimated T_C (Fig. 3A and G). We observed similar results for the other subjects.

Table 1 summarizes the results of the *2B-Cool*-estimated $\hat{T}_C$. Over the 82 trials, the system overestimated the measured T_C with a mean bias of 0.16°C (SD = 0.34°C), RMSE of 0.45°C (SD = 0.17°C), MAE of 0.39°C (SD = 0.17°C), MAPE of 1.03% (SD = 0.44%), and Pearson's *r* of 0.89 (SD = 0.12). To assess *2B-Cool*'s accuracy in estimating the timing and magnitude of the peak rectal temperature measurement T_{Cmax} in each of the three exercise bouts of a trial, we compared them with the corresponding $\hat{T}_C$ at the times of T_{Cmax}. Across the 82 trials, the average absolute error between T_{Cmax} and $\hat{T}_C$ was 0.37°C (SD = 0.29°C) (Table 2), the mean bias ranged from −0.14°C to −0.24°C, and the LOA ranged from 0.57°C to 0.62°C (Fig. 4). In computing the LOA, we accounted for the linear relation between the difference ($\hat{T}_C - T_{Cmax}$) and the magnitude of the measurement (T_{Cmax}) as well as the repeated measurements across the four experimental conditions for a subject (23,24). The average absolute difference between T_{Cmax} and $\hat{T}_{Cmax}$ was 0.36°C (SD = 0.26°C), and the average time difference was 13 min (SD = 13 min) (Table 2). The largest difference occurred when subjects wore ACU, where on average *2B-Cool* underestimated the peak by as much as 0.50°C (Table 2).

TABLE 1. Performance of the *2B-Cool* instantaneous estimates $\hat{T}_C$ as compared with the measured rectal temperature (T_C) in terms of mean bias ($\hat{T}_C$ minus T_C), MAE, RMSE, MAPE, and Pearson's correlation coefficient *r*.

	Bias (°C)	MAE (°C)	RMSE (°C)	MAPE (%)	Pearson's *r*
Men	0.16 (0.35)	0.39 (0.16)	0.46 (0.17)	1.03 (0.44)	0.91 (0.05)
Women	0.16 (0.32)	0.39 (0.17)	0.45 (0.18)	1.03 (0.45)	0.86 (0.16)
Overall	0.16 (0.34)	0.39 (0.17)	0.45 (0.17)	1.03 (0.44)	0.89 (0.12)

Entries indicate average values and 1 SD within parentheses over the four experimental conditions for men and women separately, and combined, over the 82 trials.

Performance of T_C predictions and early-warning alerts. The 20-min ahead predictions yielded an overall average RMSE over all 82 trials of 0.56°C (SD = 0.16°C) (see Table C1, Supplemental Digital Content, Comparisons of measured vs estimated and predicted core temperature, http://links.lww.com/MSS/C749). The average width of the 95% PI was ±0.62°C (SD = 0.02°C), and nearly 75% of the measured T_C fell within these limits.

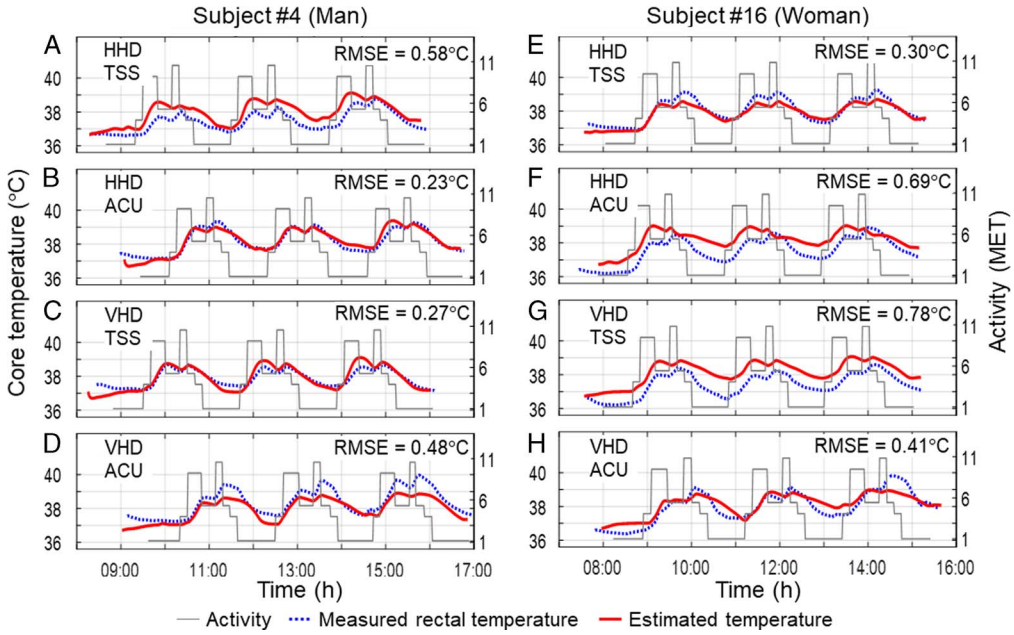

FIGURE 3—Comparison between measured rectal temperature (*dotted blue line*) and *2B-Cool* instantaneous estimates (*solid red line*) for two subjects, as they performed three bouts of time-varying, moderate- to high-intensity physical activity (*solid gray line*) under four conditions. Panels A through D show the results for subject 4 (man, 18 yr of age), whereas panels E through H show the results for subject 16 (woman, 21 yr of age). Panels A and E show the results for the hot and humid condition (HHD; ambient temperature of 30°C and 60% relative humidity), where subjects wore a TSS. Panels B and F show the results for HHD, where subjects wore an ACU. Panels C and G show the results for the very hot and dry condition (VHD; ambient temperature of 36°C and 30% relative humidity), where subjects wore a TSS. Panels D and H show the results for VHD, where subjects wore an ACU.

TABLE 2. Average performance of the *2B-Cool* system in estimating the magnitude and time of the measured peak core temperature (T_{Cmax}) for each of the three exercise bouts in the 7.5-h trials, under the four experimental conditions.

		Hot and Humid (30°C with 60% Humidity)		Very Hot and Dry (36°C with 30% Humidity)	
	Exercise Bout	TSS	ACU	TSS	ACU
$\lvert T_{Cmax} - \hat{T}_C \rvert$ (°C)	1	0.33 (0.26)	0.39 (0.36)	0.32 (0.22)	0.39 (0.31)
	2	0.30 (0.24)	0.30 (0.26)	0.37 (0.26)	0.43 (0.28)
	3	0.36 (0.23)	0.40 (0.33)	0.36 (0.32)	0.54 (0.41)
$\lvert T_{Cmax} - \hat{T}_{Cmax} \rvert$ (°C)	1	0.37 (0.25)	0.37 (0.32)	0.33 (0.21)	0.34 (0.31)
	2	0.34 (0.22)	0.29 (0.25)	0.41 (0.23)	0.42 (0.26)
	3	0.30 (0.22)	0.40 (0.34)	0.35 (0.29)	0.41 (0.30)
$\lvert$Time T_{Cmax} – Time $\hat{T}_{Cmax} \rvert$ (min)	1	7 (10)	8 (11)	6 (8)	9 (11)
	2	15 (13)	13 (15)	14 (14)	15 (12)
	3	20 (15)	11 (13)	18 (15)	17 (13)
T_{Cmax} (°C)					
Men	1	38.4 (0.4)	39.0 (0.4)	38.7 (0.4)	39.0 (0.4)
	2	38.7 (0.4)	38.8 (0.4)	39.1 (0.4)	39.2 (0.4)
	3	38.8 (0.4)	39.0 (0.4)	39.0 (0.6)	39.1 (0.4)
Women	1	38.7 (0.4)	38.7 (0.5)	38.7 (0.4)	38.9 (0.5)
	2	38.7 (0.4)	38.6 (0.5)	38.7 (0.6)	38.9 (0.6)
	3	38.9 (0.3)	38.8 (0.6)	38.8 (0.5)	39.3 (0.5)
Overall average	1	38.5 (0.4)	38.9 (0.5)	38.7 (0.3)	39.0 (0.4)
	2	38.7 (0.4)	38.7 (0.4)	38.9 (0.5)	39.0 (0.5)
	3	38.8 (0.3)	38.9 (0.5)	38.9 (0.6)	39.2 (0.5)
$\hat{T}_{Cmax}$ (°C)					
Overall average	1	38.5 (0.3)	38.5 (0.3)	38.7 (0.3)	38.7 (0.2)
	2	38.7 (0.3)	38.5 (0.3)	38.7 (0.4)	38.8 (0.3)
	3	38.7 (0.3)	38.6 (0.3)	38.8 (0.4)	38.7 (0.4)

The table also shows the average magnitude of T_{Cmax} for men, women, and overall as well as the *2B-Cool*-estimated peak temperature $\hat{T}_{Cmax}$. Values within parentheses indicate 1 SD. All pairwise Wilcoxon rank-sum test comparisons were larger than 0.05.
$\hat{T}_C$, *2B-Cool*-estimated core body temperature at the time of T_{Cmax}.

To evaluate the performance of the early-warning alerts, we identified episodes of true events where the measured T_C exceeded one or both of the two alert-triggering thresholds (38.50°C or 39.20°C) and assessed the ability of *2B-Cool* to detect these events by computing four metrics (see Methods). Across the 82 trials, we recorded a total of 227 true events where $T_C > 38.50$°C (average duration, 39 min), of which 58 also exceeded 39.20°C (average duration, 25 min). For the 38.50°C threshold, the early-warning algorithm yielded an overall sensitivity of 98%, specificity of 81%, effective prediction horizon of 36 min, and false alarm rate of 0.12 events per hour (Fig. 5). For the 39.20°C threshold, *2B-Cool* yielded a sensitivity of 87%, specificity of 77%, effective prediction horizon of 35 min, and false alarm rate of 0.35 events per hour.

Figure 6 illustrates the results for a subject under one experimental condition (no. 10, very hot and dry condition wearing TSS). Figure 6A shows the measured (dotted blue line), estimated (solid red line), and 20-min ahead predicted T_C (dashed green line) along with the 95% PI (shaded green area), whereas Figure 6B and C shows the corresponding episodes of true events (dotted blue line) and algorithm-predicted, alert-triggering events (solid green line). For this subject, *2B-Cool* predicted the five true events with an overall sensitivity and specificity ≥0.90 and an effective prediction horizon of 23 min for the 38.50°C threshold and 35 min for the 39.20°C threshold. For the 39.20°C threshold, the algorithm yielded two false alerts, resulting in a false alarm rate of 0.27 events per hour over the 7.5-h-long trial.

The effect of smartwatch HR and T_S measurements on T_C. The *2B-Cool* algorithms are agnostic of the vital sign–measuring device. Nevertheless, because vital signs measured by wearable devices may be contaminated by noise artifacts (20), we assessed the robustness of *2B-Cool*'s estimation algorithm with HR measured by different types of devices and with the unavailability of T_S data. To characterize the effect of error measurements in HR collected with the Samsung Gear S3, we compared them against those of the widely used Polar H7 chest strap device. Over the 82 trials for the 22 subjects, we observed an average RMSE between the Polar H7 and the smartwatch of 12 bpm (SD = 7 bpm). Interestingly, the RMSE for women (average of 15 bpm) was 6 bpm larger than the RMSE for men (not statistically significant; $P = 0.58$), perhaps because the watchband was not as tight on women with small wrists. To characterize the effect of these HR differences in the T_C estimates, we performed simulations where we substituted the watch's HR measurements with those of the Polar H7 and reestimated T_C. Surprisingly, we observed no significant differences in $\hat{T}_C$, with the Polar H7 yielding an RMSE of 0.48°C versus 0.45°C for Samsung Gear S3. To assess the robustness of the estimation algorithm to a lack of T_S data, we performed simulations where we estimated T_C without using T_S and obtained very similar results (0.46°C without vs 0.45°C with).

DISCUSSION

2B-Cool provides an early warning of an impending rise in T_C to thresholds associated with heat illnesses. Although the exact threshold for the onset of heat illnesses varies between individuals and depends on environmental conditions, within the context that the increase in temperature is due to exertional heat stress, there is considerable evidence to link specific ranges in T_C with the risk of heat exhaustion at the population level (1,7). Inevitably, any threshold selection would lead to false positives and false negatives. However, the thresholds

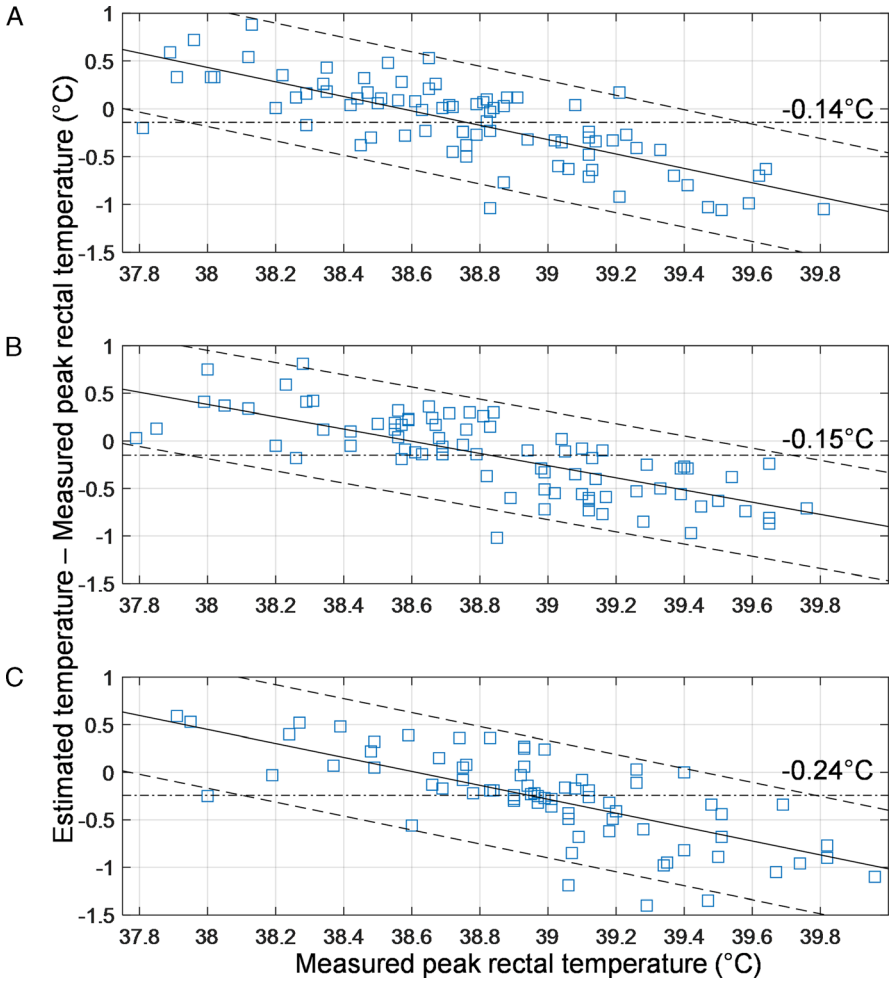

FIGURE 4—Bland–Altman plot of the difference between the estimated core temperature ($\hat{T}_C$) and the measured peak rectal temperature (T_{Cmax}) at the time of the peak, for each of the three exercise bouts of the 7.5-h trials across the four experimental conditions (82 trials). First bout (A), second bout (B), and third bout (C). In each plot, the *dash-dot line* represents the bias, and the *two dashed lines* represent the 95% LOA, which were ±0.62°C, ±0.57°C, and ±0.62°C for the first, second, and third exercise bouts, respectively. The biases are presented by the following equations (23): $y = 29.06 - 0.75x$ (A); $y = 24.76 - 0.64x$ (B); and $y = 28.30 - 0.73x$ (C).

used in *2B-Cool* can be adjusted to the specific application. For example, in the U.S. Armed Forces, greater than 95% of the 2000 annual cases of heat exhaustion and heat stroke occur during field training, as opposed to deployment (33). This gives us the flexibility to modulate the threshold to increase sensitivity to heat illness and minimize casualties at the expense of ending a training exercise early for certain Service members.

In the hardware platform discussed in this report, *2B-Cool* requires a smartwatch to collect vital signs and a smartphone to run the software. This platform supports the Nett Warrior system being developed by the U.S. Army to increase soldier situational awareness, where sensors are distributed and edge computation is performed in the soldier's smartphone (34). A similar concept would be equally useful for civilians because approximately 30% of the U.S. adult population wear fitness, health, or smartwatch devices (35) and nearly 85% of households own a smartphone (36), both of which are routinely used in everyday life for monitoring fitness- and health-related conditions (37–39). Importantly, *2B-Cool* is platform agnostic. For example, to monitor firefighters and other first responders, we recently implemented *2B-Cool* in a smartphone/dashboard platform, where individual smartphones send vital sign data via a commercial cellular network to a remotely located dashboard, which simultaneously monitors dozens of individuals.

We particularly designed the experimental part of the study to challenge *2B-Cool* with long trials lasting for 7.5 h and consisting of three bouts of increasing and decreasing physical intensity levels (from 1 to 11 METs) to allow us to assess its ability to capture sharp rises and drops in the temporal dynamics of T_C. This is in contrast to a previous retrospective validation of the underlying *2B*-Cool estimation algorithm, where the experiments only lasted for 2 h and involved constant physical activity, which resulted in a nearly monotonic rise in T_C (12). In addition, in the original validation, only seven temperature time profiles reached values that exceeded 39.00°C, whereas here we assessed the app in 58 events, in which the measured T_C exceeded 39.20°C for nearly 1,500 min.

In general, we did not observe any significant differences in the peak rectal temperature measurements across the four

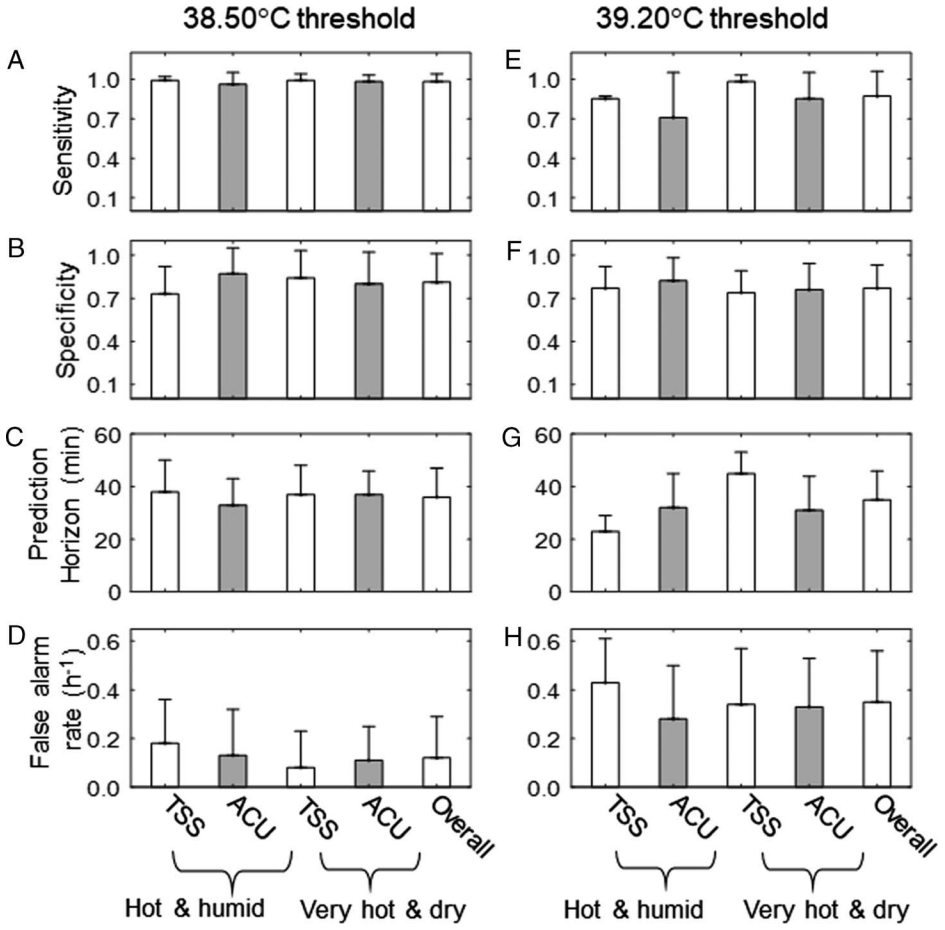

FIGURE 5—Four metrics of performance for *2B-Cool*'s early-warning system across the four experimental conditions (hot and humid, very hot and dry, TSS, and ACU). The bar plots indicate the average performance across the 22 subjects (11 men and 11 women), with the error bars representing 1 SD. Panels A through D show the results for a core temperature threshold of 38.50°C, whereas panels E through H show the results for a 39.20°C threshold. Panels A and E show the algorithm's sensitivity for detecting a core temperature rise beyond the two thresholds, whereas panels B and F show the algorithm's specificity. Panels C and G show the prediction horizon for the detection of a rising core temperature above the thresholds, whereas panels D and H show the false alarm rate per hour. Hot and humid condition, ambient temperature of 30°C and 60% relative humidity; very hot and dry condition, ambient temperature of 36°C and 30% relative humidity. All pairwise Wilcoxon rank-sum test comparisons between conditions and between sexes were larger than 0.05.

experimental conditions, except for nonsignificant, slightly larger values when subjects wore ACU instead of TSS (Table 2). To assess the performance of the estimation algorithm, we used a mean bias smaller than ±0.27°C as the acceptance criterion. This is the same criterion used by Casa et al. (26) to assess the validity of different T_C-measuring devices against gold standard rectal temperature measurements. Over the 82 trials, we obtained a mean bias of 0.16°C (SD = 0.34°C) (Table 1). When we compared the mean bias across the 22 subjects in the four conditions, we observed small not statistically significant positive and negative biases, where the system's performance varied slightly within a subject across conditions and between subjects for a given condition (see Table C2, Supplemental Digital Content, Comparisons of measured vs estimated and predicted core temperature, http://links.lww.com/MSS/C749). In terms of MAE, we obtained an overall average value of 0.39°C (SD = 0.17°C) (Table 1). We believe that such an absolute error is acceptable because it is smaller than the daily temperature variation associated with the circadian rhythm (~0.50°C) (29), and it is similar to the day-to-day variability in resting rectal temperature in young, healthy men after controlling for time of day (SD from 0.22°C to 0.39°C) (40).

For the estimations, we obtained RMSE at the individual level ranging from 0.16°C to 0.94°C (see Table C3, Supplemental Digital Content, Comparisons of measured vs estimated and predicted core temperature, http://links.lww.com/MSS/C749) for an overall RMSE of 0.45°C (SD = 0.17°C) (Table 1), which is larger than the 0.33°C (SD = 0.18°C) RMSE we observed in our previous assessment (12). This larger RMSE is likely more representative of the actual RMSE of the estimates because, as compared with our previous work, the experimental trials here were much more challenging. Reassuringly, the estimation algorithm yielded similar RMSE when we interchanged the Polar H7 chest strap HR measurements for the Samsung Gear S3 measurements, suggesting that the estimation algorithm is robust to inaccuracies in HR measurements (~12 bpm) as well as device-specific measurement artifacts. This supports

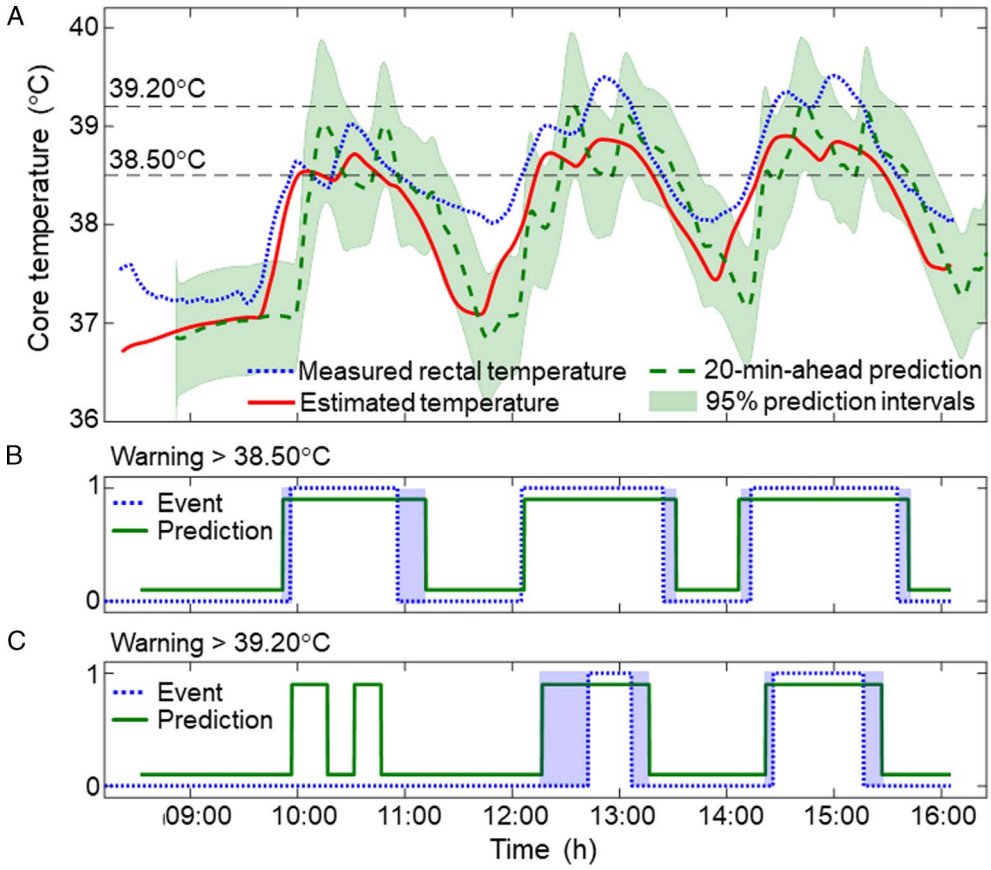

FIGURE 6—Example of *2B-Cool*'s individualized instantaneous estimates of T_C, 20-min ahead predictions, and early warnings for one male subject (no. 10) under a very hot and dry environmental condition (ambient temperature of 36°C and a 30% relative humidity), while wearing TSS. A, Measured rectal temperature (*dotted blue line*) and *2B-Cool*'s individualized estimate (*solid red line*; estimation RMSE = 0.45°C), 20-min ahead prediction (*dashed green line*; prediction RMSE = 0.52°C), and the corresponding 95% PI (*shaded light-green region*; 79% of measured data fell within the PI). B, Ground-truth events (a total of three, with an average duration of 72 min), where the measured rectal temperature exceeded 38.50°C (*dotted blue line*) and the predicted events (*solid green line*). In this case, *2B-Cool* yielded a sensitivity of 0.99, a specificity of 1.00, an effective prediction horizon of 23 min, and a false alarm rate of zero events per hour. The *blue shaded regions* denote the times before and after a true event when we discounted incorrect predicted responses in the calculations of sensitivity and specificity (see Methods). C, Same data as in panel B for the 39.20°C threshold. There were a total of two true events with an average duration of 37 min, where *2B-Cool* yielded a sensitivity of 1.00, a specificity of 0.90, an effective prediction horizon of 35 min, and a false alarm rate of 0.26 events per hour.

the notion that the required accuracy in vital sign measurements is highly dependent on the effect that their inaccuracies may have on the downstream application. Because obtaining accurate measurements of T_S is often challenging and some wearable devices lack the ability to measure it, we performed simulations where we estimated T_C without using T_S. Similar to our previous finding (12), the lack of T_S measurements did not impact *2B-Cool*'s ability to estimate T_C (0.46°C without vs 0.45°C with), allowing its integration with wearables that do not measure T_S.

In terms of the Bland–Altman LOA, our average results (±0.60°C) were in line with those observed in other estimation studies (11,13,41,42) and smaller than that of the gastrointestinal temperature sensor (±0.99°C) deemed by Casa et al. (26) as the only alternative to rectal temperature probes. However, while we computed the Bland–Altman LOA at discrete T_{Cmax} time points, as the metric was originally designed (43), these studies computed the LOA across the entire time series data with high and low temperature values, which could have averaged out the inaccuracies in detecting peak T_C values. Nevertheless, we did observe trends in bias at T_{Cmax} in each of the three bouts of the 82 trials, where *2B-Cool* overestimated T_{Cmax} for T_C smaller than ~39.10°C and underestimated it above this value (Fig. 4). The underestimation is primarily due to the time lag of the estimates (average of 13 min; Table 2) caused by the smoothing of the raw HR data (17). To account for this delay, in the early-warning algorithm we combined the predicted T_C with the 95% PI to create a larger, more effective T_C, which had the net effect of minimizing or eliminating this time lag (see Methods and Laxminarayan et al. [17]). We also observed that the MAE between T_{Cmax} and its estimated value $\hat{T}_{Cmax}$ was 0.36°C (Table 2), which is smaller than the range of daily variations in T_C between 0.39°C and 0.50°C discussed above.

As expected, the average RMSE of the 20-min ahead predictions of 0.56°C was slightly larger than the average RMSE of 0.45°C for $\hat{T}_C$ (see Table C1, Supplemental Digital Content, Comparisons of measured vs estimated and predicted core temperature, http://links.lww.com/MSS/C749) because the autoregressive model propagates errors in $\hat{T}_C$ during its iterations to predict T_C. This RMSE was also larger than the value

we observed in our previous assessment (0.33°C) (16). However, our previous assessment was based on the prediction of eight subjects where part of the subject's data were used to "train" the model and the remaining data were used for assessing the predictions in the same subjects. Importantly, 73% (SD = 19%) of the measured T_C fell within *2B-Cool*'s 95% PI (average of ±0.62°C, SD = 0.02°C), suggesting that we could not distinguish between the study data and the algorithm predictions in nearly three out of four T_C measurements.

The objective of this study is to assess the performance of the main functionality of *2B-Cool*: to provide an early warning of an impending rise in T_C associated with exertional heat illnesses. Thus, the steps of estimating T_C and making 20-min ahead predictions are only the means to achieve the desired functionality. To this end, we investigated four performance metrics for early warning (sensitivity, specificity, prediction horizon, and false alarm rate) and defined the acceptance criteria for sensitivity and specificity at >90% and for prediction horizon at >20 min (see Methods). For the 227 events where $T_C > 38.50$°C, we observed a 98% sensitivity and an 81% specificity, and for the 58 events where $T_C > 39.20$°C, we observed an 87% sensitivity and a 77% specificity (Fig. 5). Compared with our previous work (17), where we observed an 88% sensitivity and a 96% specificity for only seven events when T_C exceeded 39.00°C, we noted a reduction in specificity for an increase in sensitivity, which is an acceptable tradeoff to help reduce the risk of heat illnesses in the U.S. military during field training. For both thresholds, we observed an effective prediction horizon of ~35 min, exceeding the 20-min acceptance criterion.

The ability to provide an early warning for the risk of an impending heat illness is a distinguishing feature of *2B-Cool*, which is not available in any other approach available on the market today (11,13,41,42). An early warning is desirable because it can provide sufficient lead time to enable clinical interventions, reduction of work intensity levels, or relocation to cooler environments, all of which could help reduce the risk of an impending exertional heat illness. This early-warning projection necessarily assumes that the individual will continue the same level of exercise intensity over the next few minutes, which may not hold true. However, at the height of military training or during an athletic competition, soldiers and athletes may not perceive the warning signs of a rising T_C and having a system with a high sensitivity (>90%) to recognize such trends early could help reduce the risk of undesirable outcomes. In addition, the *2B-Cool* smartphone interface provides plots of $\hat{T}_C$ and 20-min ahead T_C as a function of time, allowing users to visualize trends in the data.

A head-to-head comparison between *2B-Cool*'s T_C estimates and those provided by other approaches on the market today would require the assessments be made against a common heat stress study, which is not possible here. Based on reported results, when compared with rectal probes or ingestible temperature pills, these approaches provide mean biases ranging from 0.01°C to 0.29°C, MAE around 0.30°C, RMSE around 0.32°C, and LOA on the entire time series data from ±0.35°C to ±0.64°C (11,13,41,42). Some of these results are arguably more promising than those obtained here; however, we also obtained better statistics in our previous studies when we did not challenge the algorithms as we did in this work. Assessment studies of short duration (60 to 120 min) with constant physical activity, where T_C increases nearly monotonically and never exceeds 38.00°C (41) or seldom exceeds 38.50°C (11,42), and that are assessed using data from the same study in which the algorithm was developed based on a leave-one-out validation procedure (13) are not sufficiently challenging to generate reproducible performance statistics. The experimental study reported here provides a good benchmark by which to assess future T_C estimation algorithms as the technology evolves.

2B-Cool performance on men and women. We observed no statistical differences in the measured T_{Cmax} between men and women in any condition (Table 2). Similarly, we observed no differences in *2B-Cool*'s ability to compute $\hat{T}_C$ between men and women. To investigate the presence of systematic errors, we separately computed the estimation bias for men and women and observed a consistent, however nonsignificant, positive bias in each of the four conditions, with an average overprediction of T_C in both men and women of ~0.15°C (see Table C2, Supplemental Digital Content, Comparisons of measured vs estimated and predicted core temperature, http://links.lww.com/MSS/C749). With regard to *2B-Cool*'s 20-min ahead predictions, we again did not observe sex differences, with average RMSE of 0.57°C for men and 0.52°C for women. In addition, we observed no sex differences in the four metrics used to assess the performance of *2B-Cool*'s early-warning algorithm. These results suggest that *2B-Cool*'s performance is indistinguishable between men and women, consistent with the experimental data.

Limitations of the _2B-Cool_ system. *2B-Cool* cannot be used to monitor changes in T_C under any environmental and physiological conditions. Specifically, *2B-Cool* cannot capture nonmonotonic variations in T_C that occur on a time scale of about 10 min or less because we used a filtering (smoothing) algorithm to remove noise in the raw HR and T_S measurements, which inherently induced a lag in the T_C estimates. Hence, it is not clear how the system would perform in stop-and-go activities of varying lengths and varying intensities, such as those in American Football. Nevertheless, we designed *2B-Cool* to capture the onset of heat illnesses in quasi-steady physical activities, which necessarily require T_C to rise and remain elevated for considerably longer periods of time. *2B-Cool* cannot be used to capture a rise in T_C because of a disease condition or in the absence of physical activity. In addition, the app cannot be used to identify a risk of hypothermia ($T_C < 35$°C) because the underlying models do not represent cold-related mechanisms. It is also not known whether *2B-Cool*'s performance would deteriorate when applied to an older, more heterogeneous population. Finally, before deployment, it is critical to validate *2B-Cool* in field settings under a diverse set of clothing and environmental conditions.

CONCLUSIONS

This study suggests that *2B-Cool* is a promising tool to monitor and forecast rises in T_C because of steady physical activities in hot and humid environments irrespective of sex, when subjects wore a military uniform or minimal clothing. We observed a mean bias of 0.16°C for core temperature estimates, which is within the daily temperature changes associated with individual variability. Importantly, *2B-Cool* provides early warning of a rising temperature beyond a clinically meaningful threshold of 38.50°C with a 98% sensitivity and 80% specificity and a sufficient lead time (~35 min) to enable clinical interventions to help reduce the risk of exertional heat illnesses. The natural next step to mature the technology is to perform a field study to assess whether the promising results reported here are reproducible in an uncontrolled environment.

This study was supported by the Military Operational Medicine Research Program Area Directorate of the U. S. Army Medical Research and Development Command, Fort Detrick, MD, and the Defense Medical Research and Development Program. In addition, the HJF was supported under Contract Numbers W81XWH-17-2-0024 and W81XWH-20-C-0031.

The authors thank Maxim Khitrov, Kamal Kumar, Adam Kapela, and Nikolaos Tountas for their help in designing the study and developing the *2B-Cool* app. They also thank Scott Montain for valuable discussions.

The results of the study are presented clearly, honestly, and without fabrication, falsification, or inappropriate data manipulation. The results of the present study do not constitute endorsement by the American College of Sports Medicine.

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, the U. S. Department of Defense, or The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc. (HJF). This paper has been approved for public release with unlimited distribution.

S. L. and J. R. received royalties for the licensing of *2B-Cool* to Seaclaid LLC.

S. L., S. H., and J. R. designed the study. L. N. B., G. E. W. G., M. C. M., and D. J. C. performed the experimental study. S. L. and J. R. analyzed the data and drafted the manuscript. All authors have reviewed the manuscript and approved the submitted version.

REFERENCES

1. Sawka MN, Latzka WA, Montain SJ, et al. Physiologic tolerance to uncompensable heat: intermittent exercise, field vs laboratory. *Med Sci Sports Exerc*. 2001;33(3):422–30.
2. Sawka MN, Young AJ, Latzka WA, et al. Human tolerance to heat strain during exercise: influence of hydration. *J Appl Physiol (1985)*. 1992;73(1):368–75.
3. Sandell RC, Pascoe MD, Noakes TD. Factors associated with collapse during and after ultramarathon footraces: a preliminary study. *Phys Sportsmed*. 1988;16(9):86–94.
4. Kenefick RW, Sawka MN. Heat exhaustion and dehydration as causes of marathon collapse. *Sports Med*. 2007;37(4–5):378–81.
5. Pugh LG, Corbett JL, Johnson RH. Rectal temperatures, weight losses, and sweat rates in marathon running. *J Appl Physiol*. 1967;23(3):347–52.
6. Yeo TP. Heat stroke: a comprehensive review. *AACN Clin Issues*. 2004;15(2):280–93.
7. Montain SJ, Sawka MN, Cadarette BS, Quigley MD, McKay JM. Physiological tolerance to uncompensable heat stress: effects of exercise intensity, protective clothing, and climate. *J Appl Physiol (1985)*. 1994;77(1):216–22.
8. Ganio MS, Brown CM, Casa DJ, et al. Validity and reliability of devices that assess body temperature during indoor exercise in the heat. *J Athl Train*. 2009;44(2):124–35.
9. King CE, Sarrafzadeh M. A survey of smartwatches in remote health monitoring. *J Healthc Inform Res*. 2018;2(1–2):1–24.
10. Hirata A, Miyazawa T, Uematsu R, et al. Body core temperature estimation using new compartment model with vital data from wearable devices. *IEEE Access*. 2021;9:124452–62.
11. Hunt AP, Buller MJ, Maley MJ, Costello JT, Stewart IB. Validity of a noninvasive estimation of deep body temperature when wearing personal protective equipment during exercise and recovery. *Mil Med Res*. 2019;6(1):20.
12. Laxminarayan S, Rakesh V, Oyama T, et al. Individualized estimation of human core body temperature using noninvasive measurements. *J Appl Physiol (1985)*. 2018;124(6):1387–402.
13. Moyen NE, Bapat RC, Tan B, et al. Accuracy of algorithm to non-invasively predict core body temperature using the Kenzen wearable device. *Int J Environ Res Public Health*. 2021;18(24):13126.
14. Niedermann R, Wyss E, Annaheim S, et al. Prediction of human core body temperature using non-invasive measurement methods. *Int J Biometeorol*. 2014;58(1):7–15.
15. Richmond VL, Davey S, Griggs K, Havenith G. Prediction of core body temperature from multiple variables. *Ann Occup Hyg*. 2015; 59(9):1168–78.
16. Gribok AV, Buller MJ, Hoyt RW, Reifman J. A real-time algorithm for predicting core temperature in humans. *IEEE Trans Inf Technol Biomed*. 2010;14(4):1039–45.
17. Laxminarayan S, Buller MJ, Tharion WJ, Reifman J. Human core temperature prediction for heat-injury prevention. *IEEE J Biomed Health Inform*. 2015;19(3):883–91.
18. McCullough EA, Jones BW, Huck J. A comprehensive database for estimating clothing insulation. *ASHRAE Trans*. 1985;91(2A):29–47.
19. Sasaki JE, John D, Freedson PS. Validation and comparison of ActiGraph activity monitors. *J Sci Med Sport*. 2011;14(5):411–6.
20. Chaudhury S, Yu C, Liu R, et al. Wearables detect malaria early in a controlled human-infection study. *IEEE Trans Biomed Eng*. 2021; 69(6):2119–29.
21. Kalman RE. A new approach to linear filtering and prediction problems. *J Basic Eng*. 1960;82(1):35–45.
22. Wald A. Sequential tests of statistical hypotheses. *Ann Math Statist*. 1945;16(2):117–86.
23. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8(2):135–60.
24. Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. *J Biopharm Stat*. 2007;17(4):571–82.
25. Horswill CA, Stofan JR, Lovett SC, Hannasch C. Core temperature and metabolic responses after carbohydrate intake during exercise at 30 degrees C. *J Athl Train*. 2008;43(6):585–91.
26. Casa DJ, Becker SM, Ganio MS, et al. Validity of devices that assess body temperature during outdoor exercise in the heat. *J Athl Train*. 2007;42(3):333–42.
27. Goodman DA, Kenefick RW, Cadarette BS, Cheuvront SN. Influence of sensor ingestion timing on consistency of temperature measures. *Med Sci Sports Exerc*. 2009;41(3):597–602.
28. Goldman RF. Introduction to heat-related problems in military operations. In: Pandolf KB, Burr RE, Wenger CB, editors. *Medical Aspects of Harsh Environments*. Washington (DC): Office of the Surgeon General at TMM Publications; 2001. pp. 3–49.
29. Kräuchi K, Wirz-Justice A. Circadian rhythm of heat production, heart rate, and skin and core temperature under unmasking conditions in men. *Am J Physiol Regul Integr Comp Physiol*. 1994;267(3): R819–29.
30. Coyne MD, Kesick CM, Doherty TJ, Kolka MA, Stephenson LA. Circadian rhythm changes in core temperature over the menstrual cycle: method for noninvasive monitoring. *Am J Physiol Regul Integr Comp Physiol*. 2000;279(4):R1316–20.

31. Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes: comparative power of the *t*-test and and Wilcoxon Rank-Sum test in small samples applied research. *J Clin Epidemiol*. 1999;52(3):229–35.
32. Chow S-C, Shao J, Wang H. A note on sample size calculation for mean comparisons based on noncentral t-statistics. *J Biopharm Stat*. 2002;12(4):441–56.
33. Armed Forces Health Surveillance Branch. Update: heat injuries, active component, U.S. Armed Forces, 2020. *MSMR*. 2021;28(4):10–5.
34. Nett Warrior. Available at https://apps.dtic.mil/sti/pdfs/ADA626617.pdf, accessed on August 23, 2022.
35. Chandrasekaran R, Katthula V, Moustakas E. Patterns of use and key predictors for the use of wearable health care devices by US adults: insights from a national survey. *J Med Internet Res*. 2020;22(10):e22443.
36. Martin M. Computer and Internet use in the United States: 2018. *US Department of Commerce*. Available at https://www.census.gov/newsroom/press-releases/2021/computer-internet-use.html, accessed on August 23, 2022.
37. Witt D, Kellogg R, Snyder M, Dunn J. Window into human health through wearables data analytics. *Curr Opin Biomed Eng*. 2019;9:28–46.
38. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJM. Complex human activity recognition using smartphone and wrist-worn motion sensors. *Sensors (Basel)*. 2016;16(4):426.
39. Haghayegh S, Khoshnevis S, Smolensky MH, Diller KR, Castriotta RJ. Accuracy of wristband fitbit models in assessing sleep: systematic review and meta-analysis. *J Med Internet Res*. 2019;21(11):e16273.
40. Consolazio CF, Johnson RE, Pecora LJ. Physiological variability in young men. In: *Physiological Measurements of Metabolic Functions in Man*. New York: McGraw-Hill; 1963. pp. 453–80.
41. Verdel N, Podlogar T, Ciuha U, et al. Reliability and validity of the CORE sensor to assess core body temperature during cycling exercise. *Sensors (Basel)*. 2021;21(17):5932.
42. Mazgaoker S, Ketko I, Yanovich R, Heled Y, Epstein Y. Measuring core body temperature with a non-invasive sensor. *J Therm Biol*. 2017;66:17–20.
43. Myles PS, Cui J. Using the Bland–Altman method to measure agreement with repeated measures. *Br J Anaesth*. 2007;99(3):309–11.