Journal of Occupational and Environmental Medicine, Publish Ahead of Print
DOI: 10.1097/JOM.0000000000003647

OPEN

Ability of Basic Physiological Monitoring to Identify Excessive Occupational Heat Strain

Thomas E. Bernard, PhD [1], 0000-0002-9974-1022

Candi D. Ashley, PhD [2], 0000-0001-5773-2702

Ismail Uysal, PhD[3], 0000-0002-3224-4865

W. Larry Kenney, PhD[4], 0000-0002-6567-9785

[1] Health Policy and Systems Management, College of Public Health, University of South Florida, Tampa FL USA

[2] Exercise Science, College of Education, University of South Florida, Tampa FL USA

[3] Electrical and Computer Engineering, College of Engineering, University of South Florida, Tampa FL USA

[4] Noll Laboratory, Department of Kinesiology, The Pennsylvania State University, University Park PA USA

Corresponding Author

Thomas E. Bernard, University of South Florida College of Public Health, 13201 Bruce B. Downs Blvd, Tampa FL 33612-3805 USA, 1-813-974-6629, tbernar2@usf.edu

Funding

The data were collected previously from five funded studies; none of which had a specific aim to evaluate personal monitoring. These included CDC/NIOSH R01-OH03983 (Bernard, Ashley), CDC/NIOSH T42-OH008438 (Bernard), DuPont (Bernard, Ashley), Electric Power Research Institute (EPRI) under contract RP2705-5 (Bernard, Kenney), and URS through US DoD, Department of the Army (Bernard, Ashley).

Conflict of Interest

None declared by CA, IU and WLK. TEB acknowledges past support from Ogletree, Deakins law firm related to heat stress and current support as a member of the Pocari Sweat® Scientific Advisory Board.

Acknowledgments

We thank the sponsors of our work. We wish to recognize Dr. Margaret Wan, whose dissertation data was part of the study. Her contributions to USF and the profession were notable.

We also wish to thank all the participants and staff that made our studies possible.

TEB, CA and WLK contributed to the study designs and data collection. TEB with IU analyzed the data. TEB, CA, IU and WLK contributed to the writing and review of the paper. All authors (TEB, CA, IU and WLK) agree to the final version.

All of the authors (TEB, CA, IU and WLK) acknowledge that AI was not used in any stage of this study from conception to the final manuscript except as follows. Machine learning, specifically neural networks, was used to develop three of 15 models as described in the methods.

The opinions stated in this paper are those of the authors and are not those of any of the funding organizations.

Reasonable requests for data will be considered and a data usage agreement will be required.

Institution and Ethics approval and informed consent

The University of South Florida IRB approved the TLP 1, TLP 2, TLP 4, PHS 1, PHS 2 and PHS 3 studies. The Penn State IRB approved TLP 3. Participants in each study provided written informed consent after the study was explained to them and they had the opportunity to ask questions.

Clinical Significance

A heat stress exposure becomes unacceptable if core temperature or heart rate becomes excessive or the person wishes to stop. Personal monitoring should be evaluated against all three reasons. The ability to discriminate between cases and non-cases is critical to the evaluation of personal monitoring.

Abstract

Objective: This study assessed the ability of basic physiological metrics to differentiate between acceptable and unacceptable heat strain.

Methods: A database of final core temperature (T_{re}), heart rate (HR), and skin temperature (T_{sk}) for 880 heat stress trials over a wide range of heat stress and clothing was created. Three case definitions based on T_{re}, HR, or fatigue were used to classify the trials as cases (unacceptable heat strain) or non-cases (acceptable heat strain). The area under the Receiver Operating Characteristic curve (AUC) characterized the discrimination ability of 15 individual physiological metrics.

Results: Generally, AUCs ranged from 0.85 allowing acceptable sensitivity with poor specificity to 0.50 with no predictive power.

Conclusions. T_{re}, HR, and fatigue were independent limits of heat strain. A program of personal monitoring needs to recognize independent reasons to classify an exposure as unacceptable.

Keywords

Heat stress; physiological monitoring; discrimination; performance; ROC; wearables

Learning Outcomes

- Explain the importance of evaluating a personal monitor performance against a case definition.
- Describe the trade-off between sensitivity and specificity, and provide the rationale for the selection of a cut-point in terms of sensitivity and specificity

Introduction

The American Conference of Governmental Industrial Hygienists (ACGIH) [1], International Organization for Standardization (ISO) [2] and the US National Institute for Occupational Safety and Health (NIOSH) [3] have issued professional practice guidelines to assess occupational heat stress exposures. In addition, the US Occupational Safety and Health Administration has started a new rulemaking process [4]. The common exposure assessment is based on wet bulb globe temperature (WBGT) and on the upper limit of the prescriptive zone [5,6]. This method has high sensitivity (0.99) (ability to predict True Positive outcomes) but low specificity (0.05) (ability to predict True Negative outcomes) to distinguish between sustainable and unsustainable heat stress among healthy, hydrated individuals [7]. The need to have high sensitivity across a wide range of heat tolerance [7,8] makes low specificity an inherent feature of heat stress exposure assessment. The imbalance between sensitivity and specificity for heat stress assessment drives the interest in personal monitoring, with the expectation that individual factors that affect heat tolerance can be accounted for [1,3,4,9-13].

While important points have been put forward concerning implementation frameworks [14] and ethical use of personal monitoring [15], understanding the requirements and limits of personal monitoring is fundamental. A recent scoping study on wearables for heat strain discussed validity in terms of how well the wearable compared to widely accepted methods of measurement of a parameter [16]. The authors pointed out the need for "thoughtful consideration related to accuracy, validity, and worker interaction with the devices". Notley et al. described a broad set of parameters that might be considered for physiological monitoring and the importance of having a rationale for limiting criteria [12]. In particular, they noted the validity problems associated with core temperature and heart rate monitoring, and the difficulty in predicting heat exhaustion from those parameters [12]. To have actionable information, it is important to develop a classification of health risk and test the ability of the metric to distinguish between (or among) the classification(s) [10].

A previous paper investigated how well common physiological metrics might distinguish between sustainable and unsustainable heat stress [17]. To characterize the ability of a metric to distinguish these conditions, the Receiver Operating Characteristic (ROC) curve was used. The ROC maps out the relationship between a metric's sensitivity (probability of True Positive) and its specificity (probability of True Negative) as 1-specificity (probability of False Positive) over the range of the metric's values. The ROC is characterized as the area under the ROC curve (AUC). AUC = 1.0 means that the metric is a perfect predictor and AUC = 0.5 means that it has no predictive value. For the task of distinguishing between sustainable and unsustainable heat stress, the AUCs for rectal temperature (T_{re}), heart rate (HR), and skin temperature (T_{sk}) were 0.74, 0.80 and 0.85, respectively. For comparison, the WBGT-based occupational exposure limit has an AUC of 0.85 [7]. That is, for the question about exposure to sustainable heat stress, the physiological metrics were not an improvement over WBGT.

In this paper, we used descriptive data on the AUC for different physiological parameters to set expectations for personal monitoring. First, the ROC curve assessed the ability of a continuous metric to differentiate an adverse heat exposure case (unacceptable strain) from a non-case (acceptable strain). There is a general consensus that unacceptable heat strain includes limits on core temperature and heart rate, and to recognize subjective fatigue/exhaustion as a limit as well

[9,18].

Methods

To estimate the ability of basic physiological measures (rectal temperature, T_{re}; heart rate, HR; and mean skin temperature, T_{sk}) to discriminate between acceptable and unacceptable heat strain (i.e., case vs non-case), data from our laboratories were combined to create the database. The observations were pulled from recorded data using one of two different protocols: Time-Limited Protocols and a Progressive Heat Stress protocol.

Time-Limited Protocol

The time-limited protocols (TLP) were designed to have combinations of environment, metabolic rate, and clothing that would cause unsustainable heat stress. The physiological values were those last reported in the trial data.

- TLP 1: This study looked for differences among commercial and prototype chemical, biological, and radiological protective clothing in three environments at a moderate metabolic rate (about 300 W) [19].
- TLP 2: Twelve participants worked at a metabolic rate of 380 W in three clothing ensembles (with different clothing adjustment values, CAV): (1) work clothes (CAV = 0°C), (2) a microporous coverall (CAV = 2.5°C), and (3) vapor-barrier coveralls (CAV = 6.5°C) at five levels of heat stress (approximately at the clothing adjusted TLV plus 7.0, 8.0, 9.5, 11.5 and 15.0°C) [20].
- TLP 3: 36 participants completed one exposure with a mean metabolic rate of 310 W (20-

min cycle of 5-min periods of sitting, lifting, cranking and walking) in one of two environments (WBGT = 36°C or 41°C) wearing one of three clothing ensembles (cotton work clothes, single-layer cotton coveralls with hood, or a vapor-barrier suit over cotton coveralls with hood) [21].

- TLP 4: This study included work clothes and vapor-barrier coveralls at two levels of heat stress (low and high) at a moderate metabolic rate (about 300 W) [22].

Progressive Heat Stress Protocol

The progressive heat stress protocol (PHS) studies started under conditions of low heat stress, established thermal equilibrium, and then increased the level of heat stress in small increments until thermal equilibrium was lost and core temperature could no longer achieve a steady state [8]. The physiological values of T_{re}, HR, and T_{sk} were those 15 minutes past the critical condition (environmental conditions associated with loss of thermal equilibrium) and were considered to be in the unsustainable range [17].

- PHS 1: This study included five clothing ensembles (cotton work clothes, cotton coveralls, plus three limited-use coveralls of polyolefin, water-barrier and vapor-barrier). The heat stress conditions represented 20, 50, and 70% relative humidity at 300 W [23].
- PHS 2: This study included five clothing ensembles (work clothes, cotton coveralls, plus three limited-use coveralls of polyolefin, water-barrier and vapor-barrier). The heat stress conditions represented metabolic rates from 100 to 500 W at 50% RH [24].
- PHS 3: This study included three ensembles of protective clothing with various configurations used for decommissioning chemical weapons. The metabolic rate was

approximately 300 W [25].

Data

The University of South Florida IRB approved the TLP 1, TLP 2, TLP 4, PHS 1, PHS 2 and PHS 3 studies. The Penn State IRB approved TLP 3. Participants in each study provided written informed consent after the study was explained to them and they had the opportunity to ask questions. The number of participants for each study and participants' sex, average age, height and weight are detailed in Table 1.

Table 1. Participant characteristics by study

Case Definitions

Physiological responses are reflected in metrics like rectal temperature (T_{re}), heart rate (HR) and mean skin temperature (T_{sk}). Volitional fatigue was a subjective decision made by the individual to stop a trial. Limiting T_{re} was an obvious target because the prevention of heat stroke is premised on leaving a margin of safety between a measured value and 40°C [26-28]. For instance, ISO suggested limiting mean predicted T_{re} to 38.0°C which was associated with 95th percentile observed value of 39.2°C, to manage the probability of a heat stroke [26,28].

HR was another obvious physiological measure of heat strain because it represents cardiovascular response to the combination of metabolic and thermoregulatory demands, and it is relatively easy to measure. Setting a limit near maximum heart rate (e.g., 90% of age predicted maximum HR) was a reasonable point to limit cardiovascular involvement [21,29,30].

Volitional fatigue (or exhaustion) was a third endpoint. Data on military training demonstrated that exhaustion was common during heat stress exposures for which core temperature and HR were acceptable [31].

We used three case definitions to distinguish among possible and independent outcomes. The case definitions were:

- T_{re} Case: T_{re} > 38.5 °C. This temperature limit was likely to include unsustainable heat stress [32], proposed to limit heat stress exposures using personal monitoring [21], and likely to limit the risk of heat stroke [26,33].
- HR Case: HR > 90% age-predicted maximum HR ($HR \geq 0.9\,(208 - 0.7\ Age)$). The observed HR should approach the individual's maximum heart rate, which marked the limit of cardiovascular response. The age-predicted HR was based on a combination of studies reporting the relation between measured HR_{max} and age [34] and 90% of the value should include the lower bound of the population.
- Fatigue Case: Volitional fatigue when the participant wished to stop the trial. Fatigue is a subjective limit and may be described as an early stage of heat exhaustion. It represents the point at which an individual should to take a break.

Data Analysis

For the purposes of this paper, T_{re}, HR, and T_{sk} formed Primary (physiological) Metrics. These were examined as standalone personal monitoring metrics. In addition to the Primary Metrics, which were common to all the trials, Table 2 lists other metrics that were available for this paper.

Table 2. Metrics available to predict cases

Some of the datasets included surrogate measures for T_{re}. These included oral temperature T_{oral}) (TLP4), ear canal temperature (T_{ear}) (TLP4), and disk temperature (T_d) (TLP3 and TLP4). Disk temperature was skin temperature under a 4.2 cm diameter by 0.8 cm thick foam disk designed to insulate the skin sensor from the ambient environment [21]. In this paper, the disk was held in place by a chest strap that also included electrodes for HR measurement.

Combinations of the Primary Metrics were labeled Combined Metrics. Four Combined Metrics included each pair of Primary Metrics (T_{re}, HR, T_{sk}) plus all three. Combinations of metrics should improve the ability to discriminate.

Derived Metrics used the Primary Metrics to determine an established or proposed metric. The Physiological Strain Index (PSI) was proposed as a method to combine rectal temperature and heart rate into a single index [35]. PSI has a nominal range from 0 to 10 with 0 representing no physiological strain, and 10 representing maximum physiological strain. PSI gained interest in the literature for basic concept and is widely used [12]. For this paper, baseline T_{re} and HR were taken as 36.5°C and 70 bpm, respectively. PSI was treated as a continuous variable.

$$PSI = 5*(T_{re} – 36.5)/(39.5-36.5) + 5*(HR – 70)/(180-70) \qquad 1$$

ACGIH suggested a sustained HR limit based on age [1]; see Equation 2. This was a dichotomous variable where $HR \geq HR_{acgih}$ is Unacceptable and $HR < HR_{acgih}$ is Acceptable.

$$HR_{acgih} = 180 - Age \tag{2}$$

Recovery heart rate was first proposed by Brouha [36] and further studied by Fuller and Smith [37]. It was to be used with oral temperature to make a decision about the level of heat stress. A version of recovery heart rate was suggested by NIOSH [3] and the ACGIH [1]. Only HR recovery at 1 min (HRR1) was used in this paper and was treated as dichotomous with decision points at either 110 or 120 bpm.

Another Derived Metric was based on moving time averages (MTA) of heart rate to help filter out peak HR in favor of a trend [21]. Four averaging periods were considered: 5 min (MTA5), 10 min (MTA10), 20 min (MTA20) and 30 min (MTA30).

One suggestion to limit exposures with protective clothing was to look at how skin temperature converges on rectal temperature [38]. A narrow skin temperature to core temperature gradient minimizes heat loss, and impairs cardiovascular responses and work in the heat [39]. While intriguing, the convergence concept is not settled [40].

Area under ROC Curves

To characterize the ability of a metric to distinguish between cases and non-cases, the Receiver Operating Characteristic (ROC) curve was constructed for continuous metrics. The strength of this approach was its wide use in biomedical research to validate the ability of a diagnostic test to classify a disease state as absent vs. present. Specifically, an ROC curve is a plot of the metric's sensitivity (probability of True Positive among cases) versus 1-specificity (probability of False Positive among non-cases). The curve is constructed by systematically changing the cut-point that defines a positive result over the observed range of the metric. The ROC curve is characterized as the area under the ROC curve (AUC). The AUC summarizes the metric's overall diagnostic ability and can be used as a global measure of the accuracy of the metric [41]. By considering all possible cut-points, the need to specify any given cut-point is avoided. Further, the ROC analysis does not assume a linear relationship between the metric and the outcome; only that it be monotonic.

For single metrics, specifically the primary and surrogate metrics and the continuous derived metrics, the ROC curve was plotted from observed metric values and case status over all observations. The metric values were ordered from lowest to highest. The non-case and case counts for each metric were noted at each value, then the cumulative total of non-cases and cases was determined for increasing metric values. The next step was to determine the probability of a non-case and a case at each metric value by dividing the counts by the overall total [42]. The trapezoidal method of integration under the curve was used to compute the AUC [43].

The preceding method cannot be used for multivariate prediction using the Combined Metrics. For this reason, logistic regression and a neural network were used to initially model the data. The logistic regression was a simple model as described in Equation 3. For the models with two metrics, one of the terms was dropped. The resulting model was then used to determine the AUC of the ROC.

$$\ln(\text{odds}) = a + b_1T_{re} + b_2HR + b_3T_{sk} + e \quad (3)$$

A neural network does not start with a preconceived model (e.g., logistic equation) where its parameters are learned through gradient descent over its cost function (e.g., MSE) with backpropagation of prediction error through the model. For this effort, the neural network had the following characteristics: an input layer explicitly defined according to the number of features in the dataset (two or three, depending on which specific prediction scenario was implemented), two hidden layers containing 16 and 8 neurons respectively, each using the Rectified Linear Unit (ReLU) activation function, and an output layer with a single neuron employing the sigmoid activation function, suitable for binary classification tasks such as case definitions in this application. Two hidden layers were selected to effectively model the complexity inherent in the dataset while preventing excessive computational complexity and mitigating the risk of overfitting. The first hidden layer, with 16 neurons, provides sufficient capacity to capture essential nonlinear relationships present in the input features. The second hidden layer, consisting of 8 neurons, acts as an additional transformation stage, refining the learned representations and enhancing the network's ability to generalize. The model was developed using Python 3 (*ipykernel*), leveraging several essential libraries including Pandas and NumPy for data handling and numerical computations; TensorFlow (compatible with version 2.0 or higher) and Keras (compatible with version 2.3 or higher) for constructing and training the neural network; scikit-learn for data preprocessing, standardization, and evaluation metrics (ROC curves and AUC scores); and Matplotlib for visualization purposes. All software and associated libraries are open source published under the Python Software Foundation License V2. Training employed the Adam optimizer, binary cross-entropy loss function, and included early stopping with a patience parameter of 5 epochs to prevent overfitting. Additionally, training was performed using a 5-fold cross-validation strategy, with each fold trained for up to 25 epochs and with a batch size of 16 using early stopping with a patience parameter of 5 epochs to prevent overfitting. An accuracy logging callback was incorporated to monitor model performance every 25 epochs. Finally, ROC curves were plotted to evaluate and visually represent the performance and robustness of the neural network model across multiple datasets. To ensure the validity and reliability of the results, the ROC curves were averaged across the 5 test folds for each specific prediction scenario and average AUC values were calculated for each plot.

For the two metrics with dichotomous outcomes (HR recovery and the ACGIH age-limited HR), 2x2 tables were created for each of the case definitions. Sensitivity and specificity were computed.

Results

Descriptive Statistics

The endpoint observations for all of the metrics are described in Table 3 as the mean, standard deviation, and number of observations by case status for each case definition. The shaded areas did not meet the criterion for independence of the case definition from the metric. Where the metric was independent of the case definition, there was considerable overlap between the case and the non-case values.

Table 3. Distribution of metrics (mean, standard deviation and count) by case definition for single observations. The cells highlighted in gray mark those combinations of case

definition and metric that are not independent and those cells that have too few observations (<5) in one cell to compute an ROC.

Table 4 is a crosstab of Fatigue by T_{re} and HR. There were 131 cases defined by T_{re} > 38.5 °C, 32 cases defined by HR > 0.9 (209 – 0.7 Age), and 149 cases of Fatigue. An endpoint observation may meet multiple case definitions. For instance, 41 of the 131 T_{re} cases and 11 of the HR cases were also marked by Fatigue.

Table 4. Crosstab of case definitions (T_{re} and HR by Fatigue)

Area under the Receiver Operating Characteristic (ROC) Curve (AUC)

The AUC for the single metrics was computed as described in the methods and reported in Table 5. The AUCs for the Combined Metrics were reported by the software package used to compute the logistic regressions and the neural networks. As expected, the AUCs were very high when the case definition used the metric.

Table 5. AUCs for each metric and case definition. Shaded AUCs represent a lack of independence between the metric and case definition.

When the Derived Metrics had prescribed cut-points, 2x2 tables were constructed, and the sensitivity and specificity were computed. These are shown in Table 6.

Table 6. 2x2 tables for the dichotomous metrics (HR_{acgih}, HRR_{110}, and HRR_{120})

Discussion

As expected for each physiological metric in Table 3, the case average was higher than the non-case average indicating greater physiological strain for a case. The difference also suggests that there is some possibility of distinguishing cases from non-cases. Because the distributions overlapped, perfect discrimination between cases and non-cases was not expected.

An important finding was the distribution of cases across the case definitions (see table 4). There were 90 cases of just T_{re}; 21 cases of just HR; and 108 cases of just Fatigue. This points to the importance of considering all three end points to assess unacceptable heat strain. For this sample of trials, a limiting HR represented an important but clear minority of cases compared to Fatigue and T_{re}. Looking from a somewhat different perspective, 30% (41/131) of T_{re} cases were also associated with Fatigue cases; as were 35% of HR cases. This is consistent with the 40% of heat exhaustions during US Army laboratory studies reported when T_{re} > 38.5°C [31].

Core Temperature

From Table 5 for the T_{re} Cases, the AUCs where T_{re} is a metric alone or with others are very near 1.0. This is a trivial insight considering the requirement to know T_{re} perfectly. The implications when a surrogate for T_{re} was used provide more insight. T_{ear} did quite well at 0.96 followed by T_{oral} at 0.89. T_d had a weaker performance at 0.78. A suggestion for using surrogate measures of T_{re} are discussed below.

PSI at 0.87 did well considering that it is a pre-determined blend of T_{re} and HR. Again, this finding is based on perfect knowledge of T_{re}. PSI had a similar AUC of 0.81 (95% CI:0.77-0.85) for predicting the upper limit of the prescriptive zone [17].

Looking at the three dichotomous variables based on HR, HRR_{100} (HR recovery at 1 minute with a threshold of 100 bpm) showed a sensitivity of 1.0 based on 6 T_{re} Cases, all classified correctly. Based on few cases, this is a strong showing. Both HR_{acgih} and HRR_{120} had sensitivities of about 0.66, which has weak protection identifying only 2/3 of T_{re} Cases.

Metrics that used HR and/or T_{sk} to predict T_{re} Cases did less well. The AUCs ranged from a weak 0.72 with HRR and combinations of HR and T_{sk} to a useless 0.52 for MTA5 and 0.53 for $\Delta T_{re}T_{sk}$.

Heart Rate

As with T_{re}, metrics using HR to predict HR Cases did very well, and the outcome was trivial. The MTAs were developed to help smooth out heterogeneous activities and look at trends. Because most of the cases were from the TLPs with steady work, MTAs do not model the case definition well and the AUCs were somewhat degraded over different averaging times with no consistent pattern. PSI was at 0.96, which seem to suggest that it is better at predicting cardiovascular limits than core temperature. The neural network based on T_{re} + T_{sk} was 0.79. Otherwise T_{re}, T_{sk}, T_{re} surrogates, and T_{re}+T_{sk} did not predict HR Cases well with AUCs ranging from 0.7 to 0.5. This suggested a general independence of T_{re} Case and HR Case outcomes.

HR_{acgih} had a sensitivity of 1.0 based on 32 cases. This is a strong showing considering that it is about 50% of age-predicted maximum heart rate. There is not much that can be offered about sensitivity for either HRR criterion because there were no HR Cases to be identified above the criterion level.

Fatigue

Fatigue Cases were subjective and independent of any of the physiological metrics; that is, did not use any physiological measurements in the case definition. Thus the trivial outcomes for T_{re} and HR are not found for this case definition. Of the three Primary Metrics, T_{re} provided the highest AUC at an unremarkable 0.74. However, combining T_{re} with HR and T_{sk} yielded AUCs in the vicinity of 0.85. PSI was only 0.69, which meant that a rebalancing of T_{re} and HR has some promise for predicting Fatigue.

None of the HR methods with prescribed thresholds in Table 6 had high sensitivity, which were in the range of 0.45 to 0.65.

Detection Problem and Performance Implications

The detection problem for personal monitoring is illustrated in Figure 1. The figure is for the Fatigue case definition because it was independent of the physiological metrics. As the figures clearly show, there was considerable overlap in the non-case and case distributions. The overlap made it difficult to find a cut-point that puts most of the non-cases on one side and the cases on the other side. The result is a low AUC.

Figure 1. Distribution of non-cases and cases for fatigue for the primary physiological metrics of a) rectal temperature (T_{re}) (AUC=0.74), b) heart rate (HR) (AUC=0.60), and c) skin temperature (T_{sk}) (AUC=0.71).

The AUCs presented above provided a general insight into the capability of a metric to distinguish between non-cases and cases. Two scenarios illustrate the implications of the AUCs. One is when there is perfect knowledge of the primary metric and it is used to classify outcomes of another case definition. The other is when a surrogate of the primary metric is used to distinguish between non-cases and cases with a case definition based on the primary metric; that is, how well the surrogate metric works.

When a primary metric is used to distinguish cases based on the primary metric, an AUC = 1.0 is expected, and the ability to distinguish is perfect. A more interesting scenario is asking how well T_{re} would distinguish between outcomes for a different case definition, such as Fatigue. The ROC curve (AUC=0.74) using T_{re} to distinguish Fatigue cases is shown in Figure 2a. Because the cut-point values are along the ROC curve, Figure 2b better illustrates the trade-off of sensitivity and specificity with increasing cut-points. To achieve a sensitivity of 0.95, the cut-point for T_{re} in Figure 2b would be 38.0°C. The resulting specificity would be 0.43, meaning that there would be a substantial number of false positive observations.

Figure 2. a) ROC curve for distinguishing Fatigue cases based on T_{re}; and b) illustration of the trade-off between sensitivity and specificity at different T_{re} cut-points.

The other scenario asks the very practical question: "How well does a surrogate metric substitute for a primary metric when the case definition is focused on the primary metric?". To illustrate this, a dataset of T_{re} and predicted T_{re} (pT_{re}) was examined [44]. The prediction method had little bias and a standard deviation of 0.3°C. The case definition was T_{re} > 38.0°C. An ROC curve and the cut-point trade-off curves were developed from the data and shown in Figure 3. The AUC was 0.72. For sensitivity at 0.95, the cut-point was 37.5°C with a resulting specificity of 0.25.

Figure 3. a) ROC curve for distinguishing T_{re} cases (T_{re} > 38.0°C) based on predicted T_{re} from a commercial device; and b) illustration of the trade-off between sensitivity and specificity at different predicted T_{re} cut-points.

Limitations

The case definitions were arbitrary. As pointed out by others, there is no present consensus on what constitutes a threshold for core temperature or heart rate to protect someone from excessive heat strain [18]. This paper chose reasonable criteria to offer some insight to the discrimination challenges awaiting current and future personal monitoring systems.

A limitation of this study was the convenience database. Most of the data were from studies based on the progressive heat stress protocol. Even though selected data were unsustainable heat stress 15 min beyond the critical condition, the participants were not near a conventional physiological limit. Thus, these data oversampled non-case trials.

Most of the cases were generated by the TLPs. There were cycles of different work (TLP 3) and periodic sitting (TLP 4) that begin to consider heterogeneous work patterns, while TLP 1 and TLP 2 were continuous exposures to one level of heat stress. While rigorous laboratory testing is necessary, the final evaluation needs to be done in the field.

This paper used a sensitivity of 0.95 in the examples. This is a commonly accepted goal but not necessarily a requirement. Professional judgment is required to select an acceptable sensitivity and resulting specificity.

Conclusions

The ability to determine physiological limits will likely be a continuously improving process. This paper encourages the use of ROC curves and the associated trade-off curves as a means to demonstrate effectiveness. This also encourages the development of standard case definitions on which to base the comparisons.

If physiological monitors are to perform as well as the WBGT-based exposure limits, the ROC AUCs should exceed 0.85 [7]. There is some promise that T_{re} surrogates can do this for T_{re}. When commercial wearables for HR are used to monitor for a fixed threshold, they are likely to do very well identifying HR Cases. A combined T_{re} and HR device that employs a neural network to develop the algorithm showed some promise as well. But what is notable was that the only noticeable improvement in ROC was for a fixed threshold HR Case definition.

There were two fundamental lessons. First, the core temperature measurement was rectal temperature, suitable only to laboratory measures. So, the performance was likely to be as good as it can be if the criterion is simply a T_{re} limit; and surrogate methods for core temperature could add more error and thus reduce the AUC. Heart rate can be reliably measured with commercial devices and thus the reliability was less of an issue than for core temperature. Fatigue, on the other hand, was subjective and much less amenable to direct assessment by a wearable.

Second, it was clear from these data that core temperature, heart rate and fatigue are likely and separate endpoints as pointed out by others [9,18]. That means that an effective personal monitoring program needs to address all three or acknowledge that it does not address all the reasonable end points. We recommend that if personal monitoring is used for limiting the exposure time, that an active query also be used. That is, the user should be reminded that fatigue may limit the exposure [21] and that the person be asked periodically if they would like to stop [45]. The absence of this appreciation presents a real risk for employees and employers.

References

[1] ACGIH. TLV for Heat Stress and Strain. Threshold Limit Values and Biological Exposure Indices for Chemical Substances and Physical Agents. Cincinnati, OH: ACGIH®; 2025.

[2] ISO. ISO7243: Ergonomics of the thermal environment — Assessment of heat stress using the WBGT (wet bulb globe temperature) index. Geneva: ISO; 2017.

[3] NIOSH. NIOSH criteria for a recommended standard: Occupational exposure to heat and hot environments. By Jacklitsch B, Williams WJ, Musolin K, Coca A, Kim J-H, Turner N. Cincinnati, OH: Department of Health and Human Services, Centers for Disease Control and Prevention, National Institute for Occupational Safety and Health (DHHS (NIOSH) Publication No. 2016-106); 2016.

[4] OSHA. OSHA Heat Stress Rule 2024 [cited 2024]. Available from: https://www.osha.gov/heat-exposure/rulemaking

[5] Dukes-Dobos FN, Henschel A. Development of permissible heat exposure limits for occupational work. Am Soc Heating, Refrig Air Cond Eng J. 1973;15:57-62.

[6] Lind AR. A physiological criterion for setting thermal environmental limits for everyday work. J Appl Physiol. 1963;18:51-6.

[7] Garzón-Villalba XP, Wu Y, Ashley CD, et al. Ability to discriminate between sustainable and unsustainable heat stress exposures. Part 1: WBGT exposure limits. Annals of Work Exposures and Health. 2017;(in press).

[8] Bernard TE, Wolf ST, Kenney WL. A Novel Conceptual Model for Human Heat Tolerance. Exerc Sport Sci Rev. 2024;52(2):39-46.

[9] Notley SR, Flouris AD, Kenny GP. On the use of wearable physiological monitors to assess heat strain during occupational heat stress. Applied Physiology, Nutrition, and Metabolism.

2018;43(9):869-881.

[10] Buller MJ, Delves SK, Fogarty AL, et al. On the real-time prevention and monitoring of exertional heat illness in military personnel. J Sci Med Sport. 2021;24(10):975-981.

[11] Niedermann R, Wyss E, Annaheim S, et al. Prediction of human core body temperature using non-invasive measurement methods. International Journal of Biometeorology. 2014;58(1):7-15.

[12] Notley SR, Meade R, Looney D, et al. Physiological monitoring for occupational heat stress management: recent advancements and remaining challenges. Appl Physiol Nutr Metab. 2025;50:1-14.

[13] ISO. ISO 9886:2004 Ergonomics — Evaluation of thermal strain by physiological measurements. Geneva: ISO.

[14] Friedrich J, Schick TS, Mess F, et al. Wearable device-based interventions in heat-exposed outdoor workers — a scoping review and an explanatory intervention model. 2025.

[15] Topazian RJ, Wec A, Ali J, et al. Ethical Use of Wearable Device Data in Occupational Settings: Perspectives From the Fire Service. J Occup Environ Med. 2025;67(9):e621-e629.

[16] Cannady R, Warner C, Yoder A, et al. The implications of real-time and wearable technology use for occupational heat stress: A scoping review. Safety Science. 2024;177:106600.

[17] Garzón-Villalba XP, Wu Y, Ashley CD, et al. Ability to Discriminate Between Sustainable and Unsustainable Heat Stress Exposures-Part 2: Physiological Indicators. Ann Work Expo Health. 2017;61(6):621-632.

[18] Notley SR, Meade RD, Looney DP, et al. Physiological monitoring for occupational heat

stress management: recent advancements and remaining challenges. Applied Physiology, Nutrition, and Metabolism. 2025;50:1-14.

[19] Bernard TE, Ashley CD. Heat Stress Performance of Four Prototype Protective Clothing Ensembles and a Control Ensemble under Dry Hot Desert Conditions and Three Prototype Ensembles in a Jungle Environment. Tampa FL: University of South Florida; 2007.

[20] Bernard TE, Ashley CD. Short-term heat stress exposure limits based on wet bulb globe temperature adjusted for clothing and metabolic rate. Journal of occupational and environmental hygiene. 2009;6(10):632-8.

[21] Bernard TE, Kenney WL. Rationale for a personal monitor for heat strain. American Industrial Hygiene Association Journal. 1994;55(6):505-514.

[22] Wan M. Assessment of Occupational Heat Strain. Tampa FL: University of South Florida; 2006.

[23] Bernard TE, Luecke CL, Schwartz SW, et al. WBGT clothing adjustments for four clothing ensembles under three relative humidity levels. Journal of occupational and environmental hygiene. 2005;2(5):251-256.

[24] Bernard TE, Caravello V, Schwartz SW, et al. WBGT clothing adjustment factors for four clothing ensembles and the effects of metabolic demands. J Occup Environ Hyg. 2008;5(1):1-5.

[25] Fletcher OM, Guerrina R, Ashley CD, et al. Heat stress evaluation of two-layer chemical demilitarization ensembles with a full face negative pressure respirator. Industrial health. 2014;52(4):304-12.

[26] Malchaire J, Kampmann B, Havenith G, et al. Criteria for estimating acceptable exposure times in hot working environments: a review. Int Arch Occup Environ Health.

2000;73(4):215-20.

[27] WHO. Health Factors Involved in Working Under Conditions of Heat Stress, Technical Report Series 412. Geneva: World Health Organization Technical Report Series 412; 1969.

[28] ISO. ISO 7933: Ergonomics of the thermal environment -- Analytical determination and interpretation of heat stress using calculation of the predicted heat strain. Geneva: ISO.

[29] ATS/ACCP. ATS/ACCP Statement on cardiopulmonary exercise testing. Am J Respir Crit Care Med. 2003;167(2):211-77.

[30] Staes M, Gyselinck I, Goetschalckx K, et al. Identifying limitations to exercise with incremental cardiopulmonary exercise testing: a scoping review. Eur Respir Rev. 2024;33(173).

[31] 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.

[32] Bernard TE, Ashley CD, Wolf ST, et al. Distribution of upper limit of the prescriptive zone values for acclimatized and unacclimatized individuals. J Appl Physiol (1985). 2023;135(3):601-608.

[33] Malchaire J, Piette A, Kampmann B, et al. Development and validation of the predicted heat strain model. Annals of Occupational Hygiene. 2001;45(2):123-35.

[34] Tanaka H, Monahan KD, Seals DR. Age-predicted maximal heart rate revisited. J Am Coll Cardiol. 2001;37(1):153-6.

[35] Moran DS, Shitzer A, Pandolf KB. A physiological strain index to evaluate heat stress. Am J Physiol. 1998;275(1):R129-34.

[36] Brouha L. Physiology in industry: evaluation of industrial stresses by the physiological reactions of the worker. Pergamon Press; 1960.

[37] Fuller FH, Smith JPE. Evaluation of heat stress in a hot workshop by physiological measurements. American Industrial Hygiene Association Journal. 1981;42(1):32–37.

[38] Pandolf KB, Goldman RF. Convergence of skin and rectal temperatures as a criterion for heat tolerance. Aviat Space Environ Med. 1978;49(9):1095-1101.

[39] Sawka MN, Cheuvront SN, Kenefick RW. High skin temperature and hypohydration impair aerobic performance. Exp Physiol. 2012;97(3):327-32.

[40] Nunneley SA, Antuñano MJ, Bomalaski SH. Thermal convergence fails to predict heat tolerance limits. Aviat Space Environ Med. 1992;63(10):886-90.

[41] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology. 1983;148(3):839-43.

[42] Bobbitt Z. How to Create a ROC Curve in Excel (Step-by-Step) 2021 [updated 9 August 2021;27 June 2025]. Available from: https://www.statology.org/roc-curve-excel/

[43] Lanoue S. Calculating AUC with the Trapezoidal Rule 2025 [updated 22 April 2025;27 June 2025]. Available from: https://www.thebricks.com/resources/how-to-calculate-auc-in-excel?utm_source=bricks&utm_medium=resources&utm_campaign=how-to-calculate-auc-in-excel#calculating-auc-with-the-trapezoidal-rule

[44] Yang S, Chen N, Callihan M, et al. Evaluation of a Wearable Core Body Temperature (CBT) Device's Accuracy in Indoor and Outdoor Environments: A Pilot Study. (in review). 2025.

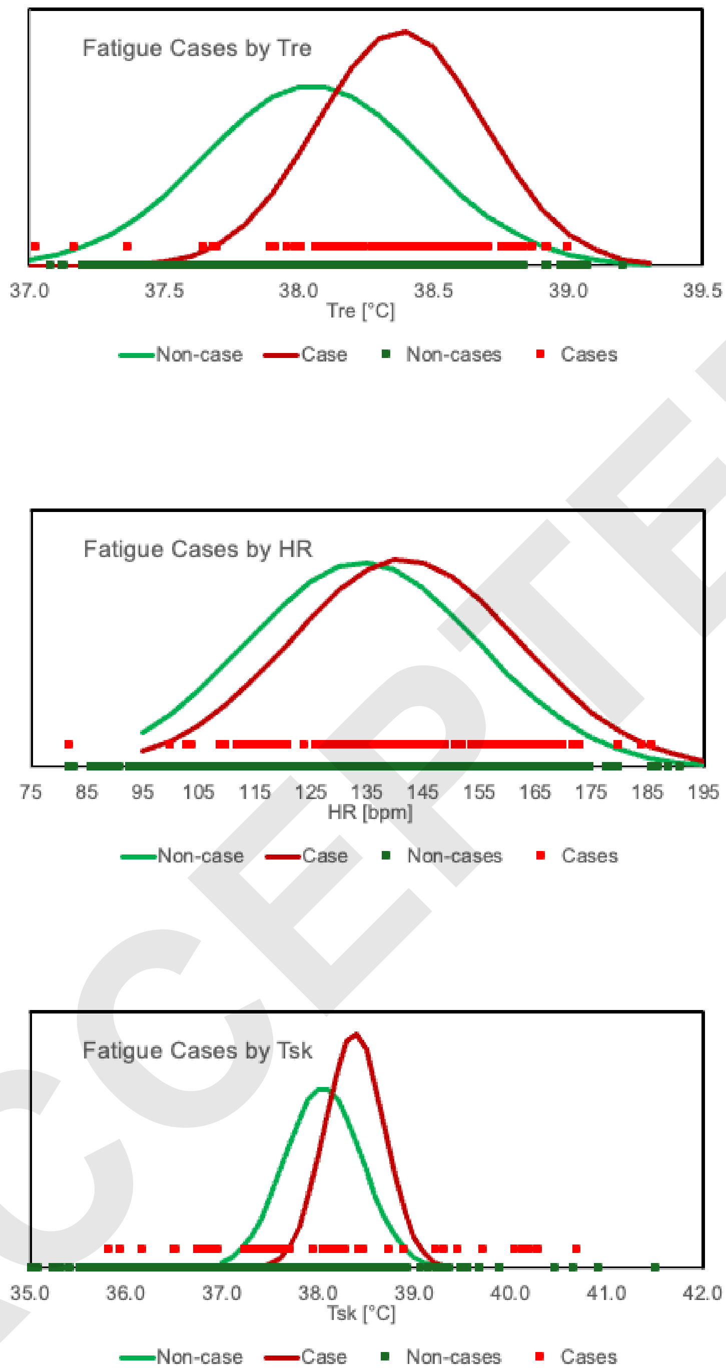
[45] Bernard TE, Kenney WL, O'Brien JF, et al. Heat Stress Management Program for Power Plants (EPRI NP-4453-L). Palo Alto: Electric Power Research Institute; 1991.

Figure Legends

Figure 1. Distribution of non-cases and cases for fatigue for the primary physiological metrics of a) rectal temperature (T_{re}) (AUC=0.74), b) heart rate (HR) (AUC=0.60), and c) skin temperature (T_{sk}) (AUC=0.71).

Figure 2. a) ROC curve for distinguishing Fatigue cases based on T_{re}; and b) illustration of the trade-off between sensitivity and specificity at different T_{re} cut-points.

Figure 3. a) ROC curve for distinguishing T_{re} cases (T_{re} > 38.0°C) based on predicted T_{re} from a commercial device; and b) illustration of the trade-off between sensitivity and specificity at different predicted T_{re} cut-points.



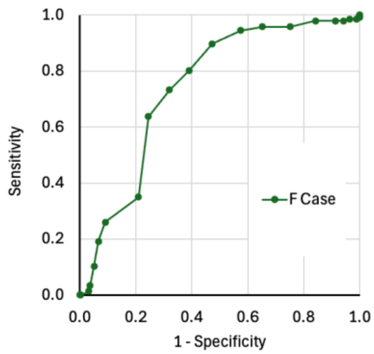
a.

b.

c.

Figure 1. Distribution of non-cases and cases for fatigue for the primary physiological metrics of a) rectal temperature (T_{re}) (AUC=0.74), b) heart rate (HR) (AUC=0.60), and c) skin temperature (T_{sk}) (AUC=0.71).

a.



b.

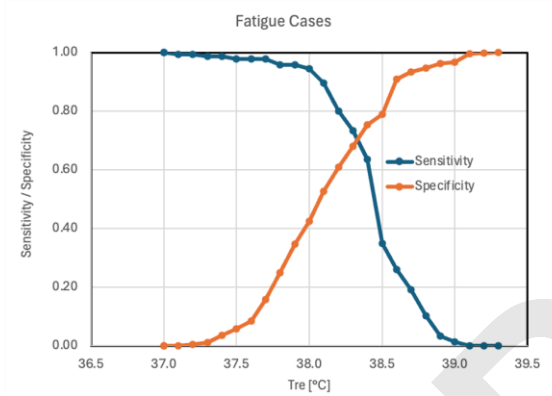
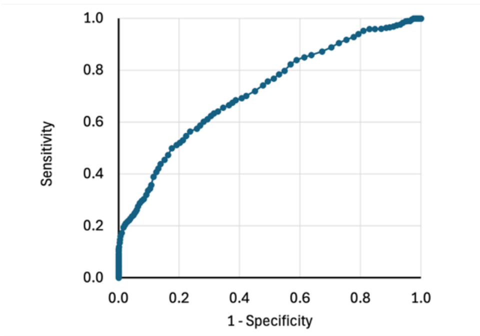
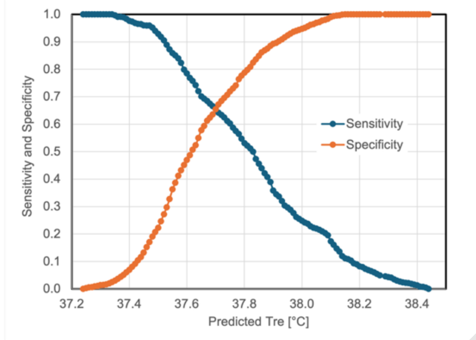


Figure 2. a) ROC curve for distinguishing Fatigue cases based on T_{re}; and b) illustration of the trade-off between sensitivity and specificity at different T_{re} cut points.



a.



b.

Figure 3. a) ROC curve for distinguishing T_{re} cases (T_{re} > 38.0°C) based on predicted T_{re} from a commercial device; and b) illustration of the trade-off between sensitivity and specificity at different predicted T_{re} cut-points.

Table 1. Participant characteristics by study

Study	Sex	N	Age [yr]	Ht [cm]	Wt [kg]
TLP 1	M	4	28	175	73
TLP 2	M	8	35	181	95
	F	4	28	160	66
TLP 3	M	31	41	177	84
	F	5	37	166	64
TLP 4	M	8	26	176	74
	F	4	26	157	64
PHS 1	M	9	29	183	97
	F	5	32	161	63
PHS 2	M	12	27	176	85
	F	4	23	165	64
PHS 3	M	6	25	181	73

Table 2. Metrics to predict cases for three case definitions.

Metric	T_{re} Case	HR Case	F Case
Primary Metrics			
Rectal Temperature, T_{re}		X	X
Heart Rate, HR	X		X
Mean Skin Temperature, T_{sk}	X	X	X
Surrogate Metrics			
Oral Temperature, T_{oral}	X	X	X
Ear Canal Temperature, T_{ear}	X	X	X
Disk Temperature, T_d	X	X	X
Combined Metrics			
T_{re} + HR			X
T_{re} + T_{sk}		X	X
HR + T_{sk}	X		X
T_{re} + HR + T_{sk}			X
Derived Metrics			
Physiological Strain Index (PSI)	X	X	X
HR Recovery at 1 min (HRR1) (dichotomous)	X	X	X
MTA (moving-time averages of HR) for 5, 10,20, 30 min	X	X	X
HR_{acgih} (dichotomous)	X	X	X
Difference between T_{re} and T_{sk} ($\Delta T = T_{re} - T_{sk}$)	X	X	X

Dark Grey cell marks a metric that is used for a case definition.

Light Grey cell marks a derived metric that contains a metric used for a case definition. The case definition is included because the derived metric is intended to be used in a combined fashion. Care in interpretation should be taken.

Table 3. Distribution of metrics (mean, standard deviation and count) by case definition for single observations. The cells highlighted in gray mark those combinations of case definition and metric that are not independent and those cells that have too few observations (<5) in one cell to compute an ROC.

Case Definition		T_{re}		HR		Fatigue	
Metric		Non-Case	Case	Non-Case	Case	Non-Case	Case
T_{re}	Mean	38.00	38.73	38.10	38.39	38.05	38.38
	Std Dev	0.34	0.18	0.41	0.27	0.41	0.31
	N	749	131	850	30	734	146
HR	Mean	134	146	134	175	134	142
	Std Dev	21	17	20	9	21	21
	N	752	131	851	32	734	149
T_{sk}	Mean	37.20	37.86	37.24	37.43	37.19	37.92
	Std Dev	0.86	1.18	0.90	0.99	0.86	1.12
	N	671	50	702	19	666	55
T_{oral}	Mean	37.55	38.28	37.68		37.45	37.98
	Std Dev	0.42	0.38	0.50		0.38	0.49
	N	23	5	28		16	12
T_{ear}	Mean	37.31	38.34	37.70		37.40	37.99
	Std Dev	0.56	0.23	0.68		0.75	0.46
	N	23	14	37		18	19
T_d	Mean	36.99	38.19	37.36	37.99	36.05	37.91
	Std Dev	1.46	0.47	1.38	0.61	1.66	0.76
	N	47	28	66	9	19	56
HRR1	Mean	119	132	124		138	119

	Std Dev	26	19	23		33	18
	N	9	6	15		4	11
MTA5	Mean	142	145	139	164	150	143
	Std Dev	17	18	15	10	10	18
	N	34	23	47	10	3	54
MTA10	Mean	140	145	138	161	146	142
	Std Dev	18	18	16	11	8	18
	N	34	23	47	10	3	54
MTA20	Mean	137	141	135	152	148	138
	Std Dev	16	16	15	12	9	16
	N	34	22	46	10	3	53
MTA30	Mean	130	136	129	148	143	132
	Std Dev	16	16	15	13	10	16
	N	34	22	46	10	3	53
PSI	Mean	5.38	7.16	5.57	7.94	5.50	6.37
	Std Dev	1.27	0.89	1.32	0.53	1.37	1.14
	N	749	131	850	30	734	146
ΔT_{re}-T_{sk}	Mean	0.78	0.76	0.78	0.84	0.81	0.40
	Std Dev	0.84	1.22	0.87	0.99	0.84	1.15
	N	671	50	702	19	666	55

Table 4. Crosstab of case definitions (T_{re} and HR by Fatigue)

	T_{re}		HR	
Fatigue	Noncase	Case	Noncase	Case
Noncase	644	90	713	21
Case	108	41	138	11

Table 5. AUCs for each metric and case definition. Shaded AUCs represent a lack of independence between the metric and case definition.

Metric	Case: T_{re}	Case: HR	Case: Fatigue
Single Metrics			
Primary Metrics			
T_{re}	0.99	0.71	0.74
HR	0.68	0.98	0.60
T_{sk}	0.68	0.61	0.71
Surrogate Metrics			
T_{oral}	0.89		0.79
T_{ear}	0.96		0.74
T_d	0.78	0.61	0.83
Derived Metrics			
HRR1	0.72		
MTA5	0.52	0.93	
MTA10	0.62	0.96	
MTA20	0.56	0.81	
MTA30	0.63	0.90	
PSI	0.87	0.96	0.69
$\Delta T_{re}T_{sk}$	0.53	0.49	0.40
Combined Metrics			
Logistic Model			
T_{re} + HR	0.99	0.98	0.75
T_{re} + T_{sk}	0.99	0.73	0.81
HR + T_{sk}	0.72	0.98	0.72
T_{re} + HR + T_{sk}	0.99	0.99	0.81

Neural Network			
T_{re} + HR	0.99	0.99	0.85
T_{re} + T_{sk}	0.98	0.79	0.84
HR + T_{sk}	0.73	0.98	0.69
T_{re} + HR + T_{sk}	0.98	0.98	0.85

Table 6. 2x2 tables for the dichotomous metrics (HR_{acgih}, $HHR1_{110}$, and $HHR1_{120}$)

		T_{re} Case	
HR_{acgih}	Case	Non-case	Sens/Spec
High	87	339	0.66
Low	44	413	0.45
HRR>110	Case	Non-case	Sens/Spec
Yes	6	4	1.00
No	0	5	0.44
HRR>120	Case	Non-case	Sens/Spec
Yes	4	3	0.67
No	2	6	0.33
		HR Case	
HR_{acgih}	Case	Non-case	Sens/Spec
High	32	394	1.00
Low	0	457	0.46
HRR>110	Case	Non-case	Sens/Spec
Yes	0	10	--
No	0	5	0.67
HRR>120	Case	Non-case	Sens/Spec
Yes	0	7	--
No	0	8	0.47
		Fatigue Case	
HR_{acgih}	Case	Non-case	Sens/Spec
High	82	344	0.55
Low	67	390	0.47
HRR>110	Case	Non-case	Sens/Spec
Yes	7	3	0.64
No	4	1	0.75
HRR>120	Case	Non-case	Sens/Spec
Yes	5	2	0.45
No	6	2	0.50

Visual Abstract