NON-CONTACT CORE BODY TEMPERATURE MEASUREMENTS: PROPOSED PERFORMANCE STANDARD AND TEST METHODS

The IEC 80601-2-59:2017[6] standard (hereafter IEC) for a non-contact screening thermograph defines equipment requirements and the associated ISO/TR 13154[7] (hereafter ISO TR) defines operational requirements, or protocols, for core body temperature measurement (the combination of these is hereafter referred to as the "ISO standard"). However, some of these requirements fail to address important confounds affecting accuracy such that a screening system can be devised that fully meets the standard's requirements yet fails to meet the scientific requirements for determining core body temperature measurement.

During the COVID-19 pandemic, the increasing sale and use of devices that technically meet the ISO standard yet fail to detect real fevers has given rise to a growing awareness of limitations in the standard. One source of these limitations is a lack of awareness of the science of the skin-to-core physiologic offset, while another source is a lack of research on fundamental methods of thermographic measurements of variable temperature objects such as faces. The largest source however, is the unforeseen ease with which the existing performance standard can be subverted by simple algorithmic methods that mask inaccurate outputs while reporting apparently accurate outputs. The test methods defined in the standard rely on laboratory IR calibration equipment, due to the challenge of objectively manipulating actual human face temperatures to sufficient accuracy.

In this technical report, we propose a modification to the performance standard for thermographic febrile temperature detection and a simple test protocol that requires no special equipment and that can be implemented by any independent test laboratory. First, three essential performance characteristics are described in light of the limitations of the existing ISO standard. Next, a modification of the existing performance standard is defined. Next, the test protocol is defined. Finally, the science of thermographic body temperature measurement is described, along with notes on confounding factors such as humidity, background luminance and wind, which are described in a manner that may cause more implementation failures than if more concrete guidance were given.

1) AMBIENT TEMPERATURE AND PHYSIOLOGIC OFFSET - THE PHYSIOLOGIC CORRECTION

The physiologic offset is the difference in temperature between core body temperature (in this document, the reference method for core body is oral temperature) and the hottest skin location on the face. For a given ambient air temperature, this reduction due to heat flowing across a known skin location is widely considered stable from person-to-person and is the underlying method of thermographic, infrared and several types of contact thermometry[1,3,4,10,11]. IEC parts 201.5.3 and 201.7.9.3.1 specify the required ambient temperature operating range must be

between 20 and 24C. However, changes in the ambient temperature affect the amount of heat flowing through a skin location and thereby affect the skin surface temperature via the skin-to-core relation, which is linear to good approximation over the range 10-35C, and having a dependence of approximately 25% on change in the ambient temperature. The specified ambient temperature range is too large to use a fixed physiologic offset, whether using the 18-24C range in the IEC standard or the narrower 20-24C range specified in the ISO TR. The 20-24C range is sufficient to allow for a 1.2C range in inferred core temperatures (or conversely, a range of 1C in the appropriate febrile threshold), which is unacceptably large, as it is comparable to or greater than the 1C threshold most commonly considered (and referenced in the IEC rationale/introduction) for febrile detection. Unless an even narrower operating temperature range is specified, a parametric physiologic offset must be used. Implementing and maintaining a narrower temperature range is highly challenging and should not be relied upon to maintain the accuracy of the resulting system output. A changing parametric physiologic offset, or the equation used to calculate the physiologic offset between surface and core body temperature for a given ambient temperature, is termed the physiologic correction. This parametric physiologic correction must necessarily rely on continuous monitoring of the ambient air temperature.

Additionally, the laboratory temperature range specified in the ISO standard (34 to 39C) is not appropriate for revealing unacceptably strong biasing-towards-normal algorithms when the facial temperature is lower than expected. This is problematic because this can hide scenarios when the equipment is being operated incorrectly, such as being used on subjects who have recently been exposed to very cold temperatures or in windy conditions, as has been documented with some NCITs and thermographic systems. This can directly translate to reduced sensitivity because the system would not discourage improper usage, in contravention of ISO/IEC 60601-1-6 medical usability standards.

IEC parts 201.101.2.3, 201.102.2 provide no discussion of the threshold beyond stating the manufacturer should describe how the threshold is arrived at and provide a few examples of how one could create the summary statistic or output value used in comparison (e.g. average of 4 pixels, average of 16 pixels) but does not specify how to arrive at the threshold value this would be compared to. Further, IEC does not discuss the source or dependence of the physiologic offset on ambient temperature. ISO TR section 6.1 discusses the offset but also does not specify it, nor does it require the offset to react to ambient temperature - essentially, ISO TR implies the operator must determine a meaningful, actionable threshold by interpreting the output of the system and calibrating to measured (or assumed healthy) core body temperatures. This is only implied in the standard and not explicitly defined nor examples provided. An example is provided here that approximately matches the technique most commonly used in the field. The operator scans ten or more presumed healthy individuals, compute the average canthus measurement (for which the surface temperature will be a few degrees below 37.0C) and then add 1.0C to this surface temperature which becomes the presumed febrile threshold for detecting core body temperatures above 38.0C. Ignoring the ambient air temperature problem momentarily, this threshold-setting method is problematic

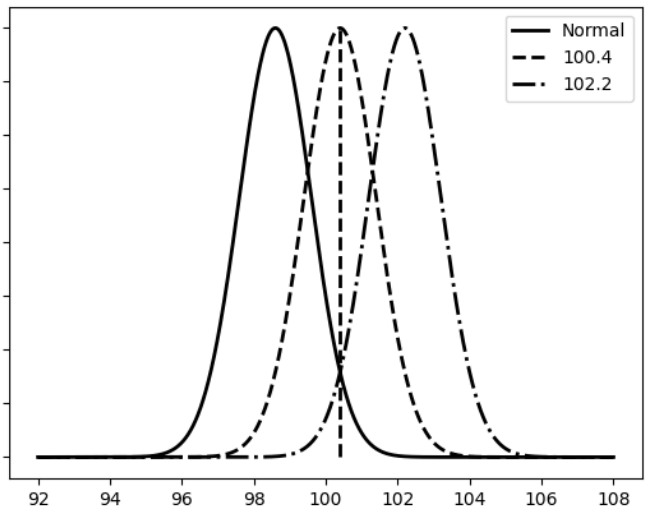
because it can lead to large variability in sensitivity over time due to variability in body temperatures and insufficient control over the previous exposure of each individual to significantly different temperature environments. Additionally, implementing this method does not result in a 38.0C core body threshold, but rather a 38.2C core body threshold due to the extrapolation from surface to core. While this is not a large difference from 38C, mistakes such as this increases the likelihood of the application failing. The source of this mistake in the standard and the source of many implementation failures is a lack of awareness of the skin-to-core dependence on ambient temperature.

Therefore, the physiologic reactivity of a system in converting a surface temperature to the extrapolated body temperature must be evaluated. The test must assess whether the system's threshold or extrapolated body temperature is reactive to changes in air temperature and that this reactivity is within 25% of a previously-determined physiologic correction, over the operating air temperature range specified by the device manufacturer. Note, this previously-determined physiologic correction was generated by reference to literature values and a 32-subject study in June 2020 and may be supplemented or replaced by some future published or consensus physiologic correction. Note, assessing the appropriateness of the exact physiologic offset used is outside the scope of this test, because the offset may incorporate device-specific covariates arising from effects that correlate with air temperature and which therefore may result in a different mathematical shape, slope or intercept of the physiologic offset but may or may not affect overall sensitivity of the system to detection of real fevers, which is the subject of another test.

2) NORMAL-BIASING AND TEST-DETECTION ALGORITHMS

The greatest challenge to the standards is the use of algorithms that rely on bias-to-normal temperatures and algorithms that detect IR calibration targets and subsequently configure the system to behave differently, both of which could prevent evaluation of the system's true performance. Therefore, it is essential to include a test using objectively-controlled elevated skin temperatures. This is possible using the physiologic correction, human volunteer subjects and two temperature-controlled rooms connected by a door. Briefly, the system under test is positioned and installed in one room while the other room is heated (or cooled) to a difference in air temperature that would provoke apparent fevers at meaningful levels if a subject being in a heated room for sufficient time to equilibrate, subsequently moved to the other cooler room for a thermographic scan. For example, the air would be heated in three separate sessions by 4C, 8C and 12C (7.2F, 14.4F and 21.6F) above the temperature of the cooler room, corresponding to elevated core temperatures of 1, 2 and 3C or 38.0C, 39.0C and 40.0C (100.4F, 102.2F, 104F). When the room is stable at each setpoint, each volunteer will first receive a scan outside the room at the lower ambient temperature and proceed to the heated room to wait for their skin to normalize to the heated room air temperature, at least 5 minutes. For efficiency, all subjects could get their first scan and then all subjects enter the heated room, leaving one-by-one, while monitoring the air temperatures in both rooms. Once the subjects leave the heated room, they

will receive a second scan within 30 seconds. This could be performed for one threshold, e.g. 1C, or for multiple thresholds as described above.



This information is sufficient to calculate estimated sensitivity via the true positive and false positive rates for each threshold tested. Therefore, the test must include means to calculate the sensitivity and specificity (to this end, a simple open-source script suitable for calculating sensitivity is included in Appendix B). Importantly, this test method is directly applicable to the actual use of the system. The figure to the left shows distributions of normal body temperatures corresponding to the variabiilty reported in large population studies of oral temperatures (e.g. standard deviation of 1F). By acquiring thermographic data at one air temperature after heating the separate environment air to an appropriate amount and using the known physiologic heat increase, it is possible to shift the entire normal distribution to desired values with sufficient precision to experimentally probe the detection statistics. The values of most interest for actionable febrile detection are the false positive detection level of the normal population and the true positive levels above the desired threshold for the normal population shifted to a fever level, such as 102.2F. Importantly, if the increase is only to the desired threshold, then a-priori we can expect only 50% sensitivity, since the distribution will remain normal and centered about that threshold, so it is essential to choose a reasonably elevated threshold in order to produce a meaningful estimate of the system's sensitivity.

3) CONFIRMATION WITH CLINICAL THERMOMETER

In IEC 201.1.1, Note 101 states the screening themograph class of device is intended only for detection of SKIN TEMPERATURE elevated above normal, and that this should be confirmed or ruled out using a clinical thermometer. This is challenging given the only other non-contact option available to-date is a non-contact infrared thermometer (NCIT), some of which have been found to be insensitive for this purpose, outputting a normal temperature for a wide range of simulated body temperatures. The requirement to rely on a clinical thermometer should explicitly disallow and warn about inaccurate thermometers.

Nevertheless, composing a system out of two independent devices increases the potential failure rate by compounding their individual failure rates. It would be useful to remove the reliance on a secondary thermometer, if the screening system were sufficiently trustworthy. At

present, the standard does not involve testing on the targets for which it is intended, human faces, but rather on laboratory references. Therefore, if a thermographic system is evaluated with a test protocol relying on human faces having normal and elevated temperatures and is found to have sufficient sensitivity and specificity, it may not be necessary to confirm with a secondary thermometer, rather it may be possible to confirm or rule out potential fever with a series of up to 3 repeated face scans, each acquired at least 5 minutes following the previous scan. Because human body temperature can vary by 1C in normal conditions, and can vary depending on recent exertion, stress, menstrual cycle, food, medication and autonomic dysfunction, if more than 50% of the scans are above the febrile threshold, the subject should be referred to trained medical staff for further evaluation. Relying on a secondary measurement when the primary measurement has been validated on human faces may actually increase system failure.

Surface Temperature Test

Materials Required:

- IR calibration target meeting the IEC equipment requirements for 80601-2-59:2017, in particular the compliance methods of sections *201.101.2.2* and *201.101.4*.
    - Here we specify one additional requirement:
        - Linearity with at most ±0.1C deviation from linear over the 32 to 40C (89.6 to 104F) range, taking into account the accuracy specification of ±0.5C.
    - Here we modify one requirement:
        - Emissivity of at least 0.98. This differs from that required by the standard, which requires 0.998. An emissivity of 0.998 is not always used in practice and is not required for this test. Furthermore, relying on such a high emissivity carries a risk of gradual change in emissivity over time, since producing emissivities very close to 1 is typically dependent on highly tortuous surface microstructure, which can be altered by adsorbed contaminants over time which reduce the actual emissivity - this is a risk for any calibrated surface emitter but industry sources consider the risk higher for the highest emissivities. The difference between a 0.998 and 0.98 emissivity calibration target changes by about 18mK per degree C change in background temperature (or 0.018F per 1F change). To reduce the impact of air temperature changes below 0.1C, the room temperature must be monitored and peak-to-peak differences throughout the test above 5C be used to either correct the target temperature or to invalidate the test. Alternatively, a lower emissivity target may be used, but the effect of a change in air temperature could be greater and as such, a lower air temperature range would be necessary. For a target having emissivity of 0.95, room air temperature should not vary by more than 2C during the test procedure.
- Fixture or experimental setup to ensure consistent distance to target

- Device under Test, or the febrile scanning system under consideration, configured to output either the surface temperature or estimated body temperature.
    - It must be noted in the test report whether the device has been configured to output surface temperatures or body temperatures.
- A controlled air temperature environment having a room air temperature within the range of 68F to 75F
- Room air temperature measurement

Test is as follows:

a) Set up continuous monitoring of room air temperature, with the maximum and minimum temperatures thus far observed during the test being saved and the maximum peak-to-peak air temperature delta calculated, with a threshold on this delta of 6F being an alarm condition notifying the test operator of potential environmental effect on the results.
b) Set up the calibration target according to the calibration target manufacturer's operating manual and enter a setpoint of 32C (89.6F).
c) Wait until the stabilization period has passed or the calibration target indicates it has stabilized if it has such a function.
d) Set up the device under test at the specified distance from and oriented towards the calibration target according to the device manufacturer's operating manual.
    i) This step may be performed in parallel with step a).
e) Wait until the stabilization period has passed or the device under test indicates it has stabilized if it has such a function.
f) Operate the device to acquire three output measurements of the calibration target, recording the average of the three measurements.
g) Increment the setpoint temperature of the calibration target by 1C and wait until the stabilization period has passed or the calibration target indicates it has stabilized if it has such a function.
h) Repeat steps e) through g) until the measurement of a 39C calibration target is complete, comprising 7 steps.

The recorded data output should be compared with the calibration target setpoints and a least-squares linear fit performed between the setpoints and the output data. The average standard deviation from this fit and the maximum deviation from this fit must be below 0.9F (0.5C). The test report should include the *plot of the output data and setpoints, the linear fit, a goodness-of-fit metric such as R2, and the average standard and maximum deviation from the fit*. The device is in compliance if the maximum deviation from the fit is less than 0.9F.

Ambient Air Reactivity Test

Materials Required:

- All materials required for Appendix A: Surface Temperature Test Protocol with one modification:
    - The device under test must be configured to output body temperature estimation and must not be configured to output surface temperature.

- A controlled air temperature environment having one setpoint of between 68 to 70F and a second setpoint of between 73 to 75F and a difference between the two of at least 5F. The initial temperature can be either setpoint.
    - Alternatively, a smaller temperature controlled chamber may be used, of sufficient size as to contain both the device under test and the calibration target
    - Alternatively, the calibrator may be located outside this temperature controlled chamber in a room having less variation in air temperature than the controlled chamber.
        - This variation must be noted as “Calibration target outside controlled chamber”.
- In the resulting test report, it must be noted whether any internal emissivity compensation for background is used by the calibration target.
- In the resulting test report, the actual emissivity of the calibration target must be noted.

Test is as follows:

a) Set up the calibration target according to the calibration target manufacturer’s operating manual and enter a setpoint of 34.5C (94.1F).
b) Wait until the stabilization period has passed or the calibration target indicates it has stabilized if it has such a function.
c) Set up the device under test at the specified distance from and oriented towards the calibration target according to the device manufacturer’s operating manual.
    i) This step may be performed in parallel with step a).
d) Wait until the stabilization period has passed or the device under test indicates it has stabilized if it has such a function.
e) Operate the device to acquire three output measurements of the calibration target, recording the average of the three measurements.
f) Adjust the room temperature to the second setpoint and wait until the room air temperature has stabilized.
    i) Stabilization can be estimated using time constant measurement and waiting at least 3.3 times this time constant for sufficient stabilization of below 0.1C.
g) Operate the device to acquire three output measurements of the calibration target, recording the average of the three measurements.

The effect of emissivity and background temperature on the results can be expected to introduce a difference of as much as 0.09C (0.05F) with the two setpoints used, depending on actual blackbody emissivity. This test ignores the effects of emissivity, which may be permitted as long as the background/air temperature changes by less than 5C and ideally less than 2C during the test. To ensure this is the case, the calibration target’s emissivity and the range of ambient air temperatures during the test must be reported in the test report.

The difference of the output data at the two setpoint air temperatures is recorded as the *body temperature change* and the difference in the two setpoint air temperatures is recorded as the *air temperature change*. The body temperature change should be scaled by the proportion of air temperature change to 1F (e.g. if the setpoints were 4.8F apart, the body temperature should be

multiplied by 1/4.8), and taken as the system reactivity to air temperature. Prior work has established that the physiologic correction is within 10% of 0.2 however the exact factor may incorporate emissivity and other confounds specific to the system. For the purposes of this test, we require only that the system reactivity to air temperature be within 0.1 and 0.3.

Sensitivity and Specificity Test

Materials Required:

- Validated clinical thermometer such as in-ear or oral thermometer
    - The use of a clinical non-contact IR thermometer (NCIT) is not allowed for this test
- Ten healthy volunteer human subjects having body temperatures below 99.5F as measured by the clinical thermometer
- Device under test, or the febrile scanning system under consideration, configured to output the estimated body temperature
- Two environmental air temperature monitoring or recording devices.
- A first stable enclosed environment for the device under test, providing stable air temperature within 1C of a given setpoint between 65 and 72F for the duration of the test and adequate space for human subjects to wait five minutes for equilibration and for the acquisition of body temperatures of these subjects.
- A second stable enclosed environment, providing stable air temperatures within 1C of a given setpoint 4C and 8C above the first environment's air temperature, and adequate space for human subjects to wait five minutes for equilibration to this second air temperature.

Test is as follows:

a) Based on the environmental temperatures of the first environment, set the second environment's air temperature to 4C above the air temperature of the first environment and wait for the second environment's air temperature to stabilize within 0.5C of the set point.
b) Initiate monitoring of each environment's air temperature, with the maximum and minimum temperatures observed during stable measurement and equilibration portions of the test being saved and the maximum peak-to-peak air temperature delta calculated, with a threshold on this delta of 1F being an alarm condition notifying the test operator of potential environmental effect on the results.
c) Set up the device under test according to the device manufacturer's operating manual in the first environment to allow acquisition of body temperatures from human subjects.
    i) This step may be performed in parallel with step a).
d) Wait until the stabilization period has passed or the device under test indicates it has stabilized if it has such a function.
e) Direct the human subjects identified for this test to remain within the first environment for the first equilibration phase consisting of at least five minutes in said first environment while refraining from exertion, stress or excessive movement.
f) During this first equilibration phase, measure the subject's body temperatures with the clinical thermometer and record these body temperatures.

g) After the first equilibration time has passed, acquire a set of thermographic body temperatures with the device under test at both the closest operating distance and the furthest operating distance and record these outputs as the *normal outputs (close-range and far-range)*.
h) Direct the human subjects to move to the second environment, occlude any large openings (e.g. close the door to the room) and remain within that enclosed environment for the second equilibration phase consisting of at least five minutes in said second environment while refraining from exertion, stress or excessive movement.
i) After the second equilibration time has passed, direct the subjects one at a time to move from the second environment to the first environment via the doorway, closing the door between each subject, and directing the subject to the device under test and acquire at the closest operating distance and the furthest operating distance to receive a second set of thermographic body temperature measurement and record these outputs as the *100.4F simulated outputs (close-range and far-range).*
j) Set the second environment's air temperature to 8C above the air temperature of the first environment and wait for the second environment's air temperature to stabilize within 0.5C of the set point.
k) Repeat steps e) through i), recording the clinical thermometry and thermographic measurements acquired in the first environment as repeat control measurements, and the measurements acquired after equilibration to the second environment as the *102.2F simulated outputs (close-range and far-range)*.
l) Calculate summary statistics of the *averages* and *standard deviations* of reported normal, 100.4F and 102.2F simulated thermographic body temperatures and the *percentage* of outputs being over the threshold of 100.4F, in each case computing separately for *close-range* and *far-range* measurements and combined *close-range* and *far-range* measurements, which the separate close-range and far-range statistics used to compute distance sensitivity and not used further. The difference between close-range and far-range in each condition (normal, 100.4F simulated and 102.2F simulated) shall be computed and the maximum difference shall be recorded as the *distance sensitivity*.

Expected outputs: average and standard deviation of combined ranges of normal, 100.4F simulated, and 102.2F simulated (6 metrics), percent over threshold of combined ranges of close-range and far-range normal, 100.4F simulated and 102.2F simulated (3 metrics) and distance sensitivity (1 metric). The standard deviations shall be pooled by vector summation to produce a single sample standard deviation to arrive at 8 total metrics retained for further analysis.

The percentage of subjects with a reported body temperature above the 38C (100.4F) threshold for each of the three conditions (equilibrated at ambient, equilibrated at +4C above ambient, equilibrated at +8C above ambient) provides an estimate of the specificity and sensitivity of the system to meaningful simulated body temperatures. The percentage of *normal outputs* above 100.4F provides an estimate of the false positive rate, or the percentage of healthy individuals incorrectly inferred as febrile. The percentage of *102.2F simulated outputs* above 100.4F provides an estimate of the true positive rate for detection of moderate fevers. The true positive

rate for 100.4F simulated outputs is not as meaningful, because it can be expected to be close to 50% at best - however it is also useful at determining whether the device has any sensitivity at detecting mild fevers. This is necessary because some devices appear to be configured to have a nonlinearly increasing sensitivity that only begins to be relevant at higher actual temperatures, and the sensitivity reported by the true positive rate for *102.2F simulated output* may not be reflective of the sensitivity to mild fever temperatures.

APPENDIX B: CODE EXAMPLE FOR SENSITIVITY AND SPECIFICITY CALCULATION

The following code is tested with Python 3.7.7, numpy 1.20.1, and scipy 1.4.1. The user enters the following information:

1) the total number of subjects scanned in ambient-equilibrated and +8C-equilibrated conditions,
2) the standard deviation assumed for the population (nominally 0.56C),
3) the number of ambient-equilibrated subjects producing outputs above the 38C/100.4F threshold, and
4) the number of +4C-equilibrated subjects producing outputs above the 38C/100.4F threshold,
5) the number of +8C-equilibrated subjects producing outputs above the 38C/100.4F threshold,
6) the average output produced for the ambient-equilibrated scans,
7) the average output produced for the +4C-equilibrated scans, and
8) the average output produced for the +8C-equilibrated scans.

>>> BEGIN CODE LISTING >>>

```
#!/usr/bin/env python3
import matplotlib.pyplot as plt
import numpy as np
import scipy.stats as stats
plt.ion()

population_std = input('Assumed population variability (typically 1.0F)?')
# Theoretical distributions
normal_dist1 = stats.norm(loc=98.6, scale=population_std)
normal_dist2 = stats.norm(loc=100.4, scale=population_std)
normal_dist3 = stats.norm(loc=102.2, scale=population_std)
delta = 1e-4
detection_grid = np.arange(92,108,delta)
pmf1 = normal_dist1.pdf(detection_grid)*delta
pmf2 = normal_dist2.pdf(detection_grid)*delta
pmf3 = normal_dist3.pdf(detection_grid)*delta

thr_index=np.where(detection_grid>100.4)[0][0]
print('Theoretical FP rate for ambient-equilibration sample: ', np.sum(pmf1[thr_index:]))
print('Theoretical TP rate for 100.4-equilibration sample: ', np.sum(pmf2[thr_index:]))
print('Theoretical TP rate for 102.2-equilibration sample: ', np.sum(pmf3[thr_index:]))
```

```
pdf1 = pmf1/delta
pdf2 = pmf2/delta
pdf3 = pmf3/delta

# Experimental distributions
N = input('Total number of subjects?')
sample_std = input('Measured sample variability (vector sum across equilibration samples)?')
N_amb_febrile = input('Number of subjects above 100.4F in ambient-equilibration sample?')
N_100p4_febrile = input('Number of subjects above 100.4F in 100.4-equilibration sample?')
N_102p2_febrile = input('Number of subjects above 100.4F in 102.2-equilibration sample?')
T_amb = input('Average output temperature for ambient-equilibration sample?')
T_100p4 = input('Average output temperature for 100.4-equilibration sample?')
T_102p2 = input('Average output temperature for 102.2-equilibration sample?')

dist = stats.norm(loc=T_amb, scale=sample_std)
dist_amb = dist.pdf(detection_grid)*delta
dist = stats.norm(loc=T_100p4, scale=sample_std)
dist_100p4 = dist.pdf(detection_grid)*delta
dist = stats.norm(loc=T_102p2, scale=sample_std)
dist_102p2 = dist.pdf(detection_grid)*delta
print('Estimated FP rate for ambient-equilibration sample: ', np.sum(dist_amb[thr_index:]))
print('Estimated TP rate for 100.4-equilibration sample: ', np.sum(dist_100p4[thr_index:]))
print('Estimated TP rate for 102.2-equilibration sample: ', np.sum(dist_102p2[thr_index:]))
print('Measured FP rate for ambient-equilibration sample: ', float(N_amb_febrile)/N)
print('Measured TP rate for 100.4-equilibration sample: ', float(N_100p4_febrile)/N)
print('Measured TP rate for 102.2-equilibration sample: ', float(N_102p2_febrile)/N)

fig= plt.figure()
plt.plot(detection_grid,pdf1, 'k', label='Normal', linewidth=2)
plt.plot(detection_grid,pdf2, 'k--',  label='100.4', linewidth=2)
plt.plot(detection_grid,pdf3, 'k-.', label='102.2', linewidth=2)
vert=np.linspace(0, np.max(pdf1), 100)
thr=np.zeros_like(vert)+100.4
plt.plot(thr, vert, 'k--', linewidth=2)
plt.legend(loc='best')
plt.title('Test Protocol')
plt.show()
```

<<< END CODE LISTING <<<

APPENDIX C: SCIENCE OF THERMOGRAPHIC BODY TEMPERATURE MEASUREMENT

Accurate Thermographic Surface Temperature Measurement

Thermal imaging for surface temperature measurement is, in principle, an off-the-shelf component provided by nearly any calibration thermographic camera system. However, few off-the-shelf uncooled cameras provide accuracy better than 2C, and the accuracy required for inner canthus-based body temperature measurement is at most 0.5C. To achieve this accuracy, the only reliable method known at present is to incorporate an IR calibration target having

accuracy better than 0.5C into the system. Most IR calibration targets rely on resistive heating elements, often simply a cartridge heater inserted in the truncated peak of an Aluminum cone parallel to an inserted thermocouple located several millimeters off-axis, with the base of the cone being the emissive-coated calibration target surface. A PID controller is configured with parameters derived to maintain a stable temperature at the surface. Stability and uniformity can be nearly guaranteed by the geometry and thermal conductivity of the elements used and use of a radiative shield with lower-conductivity couplings such as phosphor bronze. The benefit of this arrangement is very low price, with a total IR calibration target cost (assembled, calibrated and tested) of under $40. However, this method has two major drawbacks: 1) the system cannot pull its temperature down if the controller overshoots either from slow drifts in thermocouple output or during changing environmental conditions and 2) the heat balance flow from the cartridge heater out to conductive and convective air can change significantly with small changes in the character of the air (low-level airflow, humidity, concentration of gases) which isn't taken into account by the embedded thermocouple. Improved calibration targets can be made using thermoelectric modules.

The Physiologic Offset

The use of appropriate thermographic equipment and procedure (see IEC[6] and ISO[7] standards and Izhar et al[1]) can obtain inner canthus or face maxima surface temperatures that can be used to identify whether individuals have febrile core body temperatures. However, the inner canthus or face maxima surface temperature is reduced from the core body temperature by the insulating effect of skin and film coefficient of still air surrounding the body. For a full treatment of bioheat transfer[2], including the effects of perfusion and different tissue layers, see Section 4 in Izhar et al[1]. In still air conditions typically encountered however, the effect reduces to that of static thermal equilibrium for the aforementioned thermal resistances of skin and air and thermal radiation between the skin and the environment. These effects are approximated well by a linear relationship, which has been documented in the literature over typical ambient air temperatures between 30F and 100F[3,4]. This relation has been confirmed by an internal study incorporating 32 volunteers performed in June 2020, where each individual's facial maxima was recorded while the subject had equilibrated at three stable ambient temperatures ranging from 45F to 85F and described further below (shared with researchers at CDRH and available on request).

This effect is correctable, if the ambient air temperature is known, and the physiologic calibration factor is known and applied correctly. This is dependent on the total effective resistance of the skin-to-core blood insulation, plus the skin-to-air combination of conductance, convection, and radiation to ambient air. (This skin-to-air resistance can also vary in ways that aren't monitored routinely today and is the subject of a planned follow-on technical research and development. For this report, skin-to-air resistance is assumed to be constant).

The reason for targeting the inner canthus is the relative thinness and consistency across people of this insulative skin above the vasculature fed by the angular artery (fed by the facial artery, which in turn is fed by the external carotid), and the relative lack of autonomic vasoreactivity in the inner canthus. This region is considered the most reliable area of normally-exposed skin for non-contact core body temperature measurement. The forehead also

has arteries near the surface, but they are more variable in location, the thickness of skin at the forehead is more variable from person to person and finally the presence of significant autonomic vasoreactivity in the forehead. A recent clinical trial[5] involving regions of interest encompassing the forehead, inner canthi and whole face maxima revealed a large difference between forehead and inner canthi, with inner canthi significantly more reliable, and surprisingly, whole face maxima to be the most reliable for detection.

$$T_{surface} = T_{core} - 0.2 * (T_{core} - T_{ambient}) \qquad \text{Equation 1}$$

The physiologic calibration factor can be obtained using a suitable population of test subjects with core body temperature measurements and inner canthus surface temperature measurements in at least two stable ambient air temperatures after each subject being scanned has fully equilibrated to each ambient air temperature. Our physiologic calibration factor study was performed using a population of 32 individuals each scanned after equilibration in three different areas having ambient air temperatures of 45F, 70F and 85F and with oral thermometry performed once with a clinical heated probe oral thermometer (Welch Allyn SureTemp Plus 690) which had its accuracy tested in an immersion circulator water bath accurate to 0.1C traceable to NIST calibration standards (PolyScience SD7LR-20). For each scan, the ambient air temperature was recorded.

The physiologic calibration factor is obtained by fitting the surface temperature data to the oral temperature data, where the range in ambient temperatures must be sufficiently far apart to minimize the error of the resulting fit. We share our physiologic correction formula with a calibration factor that is sufficiently close for test and demonstration purposes to the factor we use. We caution researchers and engineers that the exact calibration factor should be independently obtained at one time with the equipment intended for use and referenced to traceable calibration standards at that time, because it is not unlikely there exist equipment-dependent factors, such as sensitivity to background emissivity, that could result in a slightly different calibration factor. Regardless, a study of this type is essential to confirming ability to react to ambient conditions accurately. Equation 2 is merely Equation 1 trivially rearranged for ease of use because in most conditions, the user is trying to extrapolate the core temperature from the surface and ambient air temperatures. The primary reason for presenting Equation 2 here is to demonstrate that the extrapolation from surface to core uses a larger factor, and to prevent accidental error from using the smaller 0.2 factor in scaling rather than the correct 0.25 factor that results from the rearrangement.

$$T_{core} = T_{surface} + 0.25 * (T_{surface} - T_{ambient}) \qquad \text{Equation 2}$$

These equations may be used for temperatures in any of Fahrenheit, Celsius or Kelvin. For example, in a room temperature of 68F, using the first formula, a core body temperature of 98.6F would correspond to 92.3F. Conversely, using the rearranged Equation 2, a surface temperature of 95F in room temperature of 68F would correspond to a core body temperature of 101.25F, which is above a mild fever threshold of 100.4F. If the room temperature were increased to 75F, a surface temperature of 95F would correspond to a core body temperature of

100F, and would be considered below (most) fever thresholds. In fact, if a correction was based on a single temperature anywhere in the ISO standard's 20-24C (68 - 75.2F) range, the *resulting core body temperatures would then be allowed to vary by as much as 1.8F* depending on the actual ambient air temperature at the time of scan.

Given the other sources of variance inherent in measurement systems and the fact we desire to use a threshold that is only 1.8F greater than the assumed normal core body temperature, the need to monitor and react to ambient temperature more closely is a *critical deficiency in the standard*. One could either control ambient carefully or monitor ambient and use the resulting information to react to these changes in ambient.

One method for compensating for changes in ambient temperature is to use a moving average over the last N subjects scanned. However, the effectiveness of this method is limited by the variability of normal body temperatures, the presence of potential actual abnormal temperatures in the last N subjects, any effects of air and wind on these last N subjects, and time passed since the first and last of these N subjects. If the number of subjects is increased, the effect of variability is decreased, but then the time and potential for air temperature to affect the system will increase. Being on guard for these effects increases the burden on the operator and increases the likelihood of errors that would not be introduced when using a calibrated physiologic correction. The system's sensitivity to ambient temperature should be tested for both healthy and lowest desired fever threshold temperatures.

The Use of Pseudo Methods for Biasing outputs to normal

There are now numerous non-contact fever detection systems that have been examined and found to be reporting core body temperatures that do not correspond to the surface temperatures and a realistic physiologic correction. Most concerningly, these systems report numbers closer to an expected normal core body temperature despite being tested with IR calibrators set to surface temperatures corresponding to healthy and febrile core body temperatures.

The most prevalent method appears to be a nonlinear, S-shaped curve that "pushes" numbers closer to 37C (98.6F), regardless of how inappropriate this is. This has the side effect of making a system appear accurate when tested against oral or core body temperatures when only people with non-febrile oral or core body temperatures are examined, making this technique nearly impossible to detect with the resources typically available to customers. A polite term for this type of technique is "pseudo-method". Note, some amount of filtering may be acceptable to reduce false positives while maintaining a sufficient level of sensitivity, but the system's overall sensitivity and specificity must be tested while operating with the filtering that would be used in the field. Critically, this filtering must be disclosed with an analysis of its effect on the sensitivity and specificity of the device's outputs when operated normally as specified.

We define pseudo-method as a method devised to accept a range of thermal sensor data and produce temperature outputs consistent with healthy body temperatures when tested in the general population, or any number between 96.8 and 99.5F (36 and 37.5C). One example pseudo-method could be to output a constant average human body temperature plus or minus some small random number matching the variability of the human population. Many other pseudo methods are possible, including ones which would linearly correlate with oral

temperatures in large populations yet fail to detect actual fevers. For example, the device could add 0.1 times the thermal sensor data plus 0.9 times 98.6F, which would correlate with oral temperatures, possibly significantly above chance, but fail to report above threshold for any body temperature likely seen in the real world.

Because there the variation in oral temperature measurements is near 1F in the majority of the population, a system devised with a pseudo-method can be undetectable from a system devised for accuracy unless the testing population includes wider variation in core body temperature due to fever, and furthermore, that the accuracy test is weighted to include equal contributions from core body temperatures further from the average. Without weighting the test to balance non-normal temperatures with normal temperatures, or intentionally separating into at least two groups (e.g. having oral temperatures above threshold and below), a large enough normal population will overwhelm any test statistic produced, with the result a pseudo-method could easily appear accurate by the test metric.

It is possible to detect pseudo-methods in laboratory settings by testing whether the system is a) responsive to changes in ambient temperature and b) has an appropriate transfer function relating surface temperature to core body temperature estimate. These can be determined most effectively with an IR calibration source set to a range of temperatures encompassing the range of surface temperatures expected in the population of healthy and feverish individuals, as described in this document.

However, some systems are configured to not report a temperature in body mode when there is no face present, and furthermore, it is possible for object detection systems to be configured to detect calibration targets as a separate class than faces, and to behave differently when a face is detected versus when an IR calibration target is detected and thereby revert to an appropriate transfer function when tested with an IR calibration target but rely on an inappropriate transfer function when directed at faces. The manufacturer of the system under test could attest whether the system performs any face detection-specific adjustments to the temperature and describe the differential processing performed or enable a test mode that performs the same processing as face detection mode without performing a face detection. This is not necessary because it is possible to use different temperature environments and faces to experimentally adjust actual face temperatures, as described in the test protocol.

Clinical Validation

The final validation for devices seeking de novo status for non-contact core body measurement is a clinical trial enrolling individuals from the general population or a hospital population. This clinical trial should be designed to attempt to include a sufficient number of subjects having a range of core body temperatures extending to at least moderate fevers of 102F or above. The actual number of febrile subjects included will not be known initially but can be estimated in some cases, in particular a clinical internal medicine practice is likely to have an estimate of the percent of patients having febrile temperatures, and this estimate can be used to inform the study recruitment size. Most importantly, any clinical trial result should perform separate accuracy analyses for groups of healthy, mild fever, moderate fever and severe fever if possible, or at minimum healthy and feverish groups, in order to avoid swamping the accuracy analysis with a large number of healthy results. Alternatively, the data can be reweighted in the

analyses such that the errors from febrile subjects are equally weighted with errors from healthy subjects.

For practical implementation, two values are usually of most interest: what portion of fevers will be identified as fever (portion of fevers that are true positives) and what portion of healthy will be identified as fever (portion of healthy that are false positives). The first translates into exposure probability reduction (and hence risk reduction) and the second translates into likelihood of implementation failure. This can be measured as described in the test protocol, but it can also be estimated from the expected variability of human temperatures and the device's surface temperature accuracy performance.

Human core body temperatures measured with clinical oral thermometry have an approximately normal distribution centered on 37C (98.6F) with a width of 1F (0.5C). The false positive rate for oral thermometry can be estimated by summing the area under a normal distribution above 100.4F (or other selected threshold). For standard clinical oral thermometry, the false positive rate for 100.4F is estimated as 3.6% (to see other thresholds, an online false positive calculator can be used, entering a mean of 98.6, sigma of 1 and input a greater-than threshold). Non-contact methods are unlikely to perform as well as oral thermometry and should be expected to have a greater standard deviation. If a non-contact system exhibits a lower standard deviation than the oral thermometry (or the 0.9F standard deviation expected in large populations), it is likely the system is relying on some level of biasing towards a characteristic value and therefore may have less actual sensitivity. Therefore, it is useful to have an estimated standard deviation to expect, which we arrive at by vector-summing the estimated system surface temperature accuracy of 0.9F (0.5C) with the population mean of 0.9F (0.5C), which results in an expected non-contact system standard deviation of 1.3F (0.7C), resulting in a false positive rate of nearly 10%. Repeating the test will eliminate most of these false positives but carry the risk of a subsequent false negative, especially for actual core body temperatures that are near the threshold.

The true positive rate can be estimated by assuming that 1) fevers have a uniform distribution between 100.4 and 104F, and 2) the measurement standard deviation of 1.3F is a normal distribution convolved with the uniform, resulting in a blurred uniform distribution. Then, taking the area under the threshold over the total area gives us an estimate of detection sensitivity, or what portion of fevers the system could detect. This is an estimate with several assumptions but it provides a realistic way to compare systems. For an example system examined, the 100.4F threshold was not crossed until the IR calibration target was set to a simulated core body temperature of 103.6F. The relative area above the threshold is only 16.7%, meaning this example system (an off-the-shelf FDA-approved NCIT) is likely to detect only 16.7% of fevers. To have a true positive rate greater than 50%, the system must produce output core body temperature no greater than halfway between the 100.4 to 104F range, or 102.2F, meaning the system must have an actual (e.g. not just for 98.6 but over the full range) accuracy of no worse than 1.5F. We are unaware of any system not using a blackbody that can even approach this accuracy. With proper design, it is possible to achieve 0.9F total system accuracy and thereby detect the majority of fevers.

NOTES ON UNNECESSARY IMPLEMENTATION REQUIREMENTS

1) Confounds from nearby heat sources can be potential problems, but only if above some fairly permissive level. The level of concern can be approximated to first order by comparing the half-angle integrated temperature at the measurement site (the subject's face), and typically, due to the small proportion of solid angle subtended by most typical sources, most sources, even if hotter than 100C, can be ignored. For example, consider a 1m^2 hot water radiator with an average surface temperature of 70C, at 2m from the subject. The half angle surface area at 2m is 4*pi*2*2 m^2, or 25.2 m^2, which means the radiator fills just under 4% of the half angle emitting to the measurement site. Assuming the rest of the environment is radiating at the room temperature of 20C, this would increase the background radiation from 20C to 22C. The impact of this increased radiance on a 98% emissive inner canthus would be an apparent increase of 36mK. A large radiator at that high of a temperature, which should be noted is impermissible in common areas in commercial facilities due to potential for harm, is unlikely to be present in most scenarios. Thus, while it is certainly conceivable one could construct a scenario that leads to excessive and uncorrected background radiance, it is a far lower likelihood concern than other confounds. Mid-wave IR sensors may have additional challenges with certain window types, particularly if the windows in question are very large and exposed to direct sunlight. Therefore, the current guidance is probably sufficient but could use additional clarity to enable scenarios with some heat sources that are unlikely to cause confounds.
2) Confounds from wind are also potential problems, but again there isn't a clear solution and the confound may only be relevant at wind speeds higher than a few m/s. Walking at a slightly elevated speed of 2m/s versus standing could produce a significant change in the total heat transfer from the sides of the face (note, heat transfer incorporates conduction, convection and radiation, although in most references, and here, heat transfer specifically excludes radiative transfer). However, data on the actual heat transfer of skin versus air speed is lacking for low speeds such as walking. In our modeling work, we use 8.35 W/m^2 for the viscous film layer heat transfer coefficient in still air. The difference in radiative heat transfer from a 34C body to ambient is approximately 6W/m^2 (linear approximation, sufficient for 10-35C ambient and 34C skin), which will not change with wind speed. With increasing air speed, the film layer effectively reduces in thickness. In real-world conditions in mostly still air, the heat transfer is likely on the order of 10W/m^2, and while people are walking this may rise to 20W/m^2 at the sides of the head (h = 12.12 - 1.16*v + 11.6*sqrt(v), not valid below 2m/s), but the inner canthus is protected from the increase by location and head geometry, possibly by a factor of two or more, thus perhaps an increase from 10 to 15W/m^2. Fluid flow models are beyond the scope of this work, but could help in further narrowing uncertainty. Specifically, there may be a wind speed beyond which the flow transitions to turbulent within the canthus which then breaks down the protection of the inner canthus, resulting in a nonlinear increase in total heat transfer. Therefore, the guidance to avoid noticeable drafts should stand. However, the use of wind speed anemometers or other one-time measurements during setup will be blind to changes in drafts due to forced air systems and other changes to the environment that can and

often do happen in real-world conditions. An additional recommendation would be that the operator survey for noticeable drafts during the operation of the device, which can be achieved by checking for noticeable drafts on the operator's face once per hour during the course of a day, repeating this monthly or during major weather changes or when the building's environmental monitoring indicates major shifts in the forced air handling systems.

3) The ISO standard notes humidity has an effect on the transmission of long-wavelength infrared, which is true, and can be calculated based on the measured extinction coefficient of air, which increases with relative humidity (RH). This can vary because most facilities attempt to maintain RH within 20 to 60% for comfort. However, the scale of this effect is negligible at distances up to 2 meters. In fact, the extinction coefficient for 8-14 micron IR has been measured to rise from 0.1 km^{-1} at low relative humidity to 0.4 km^{-1} for high relative humidity in normal conditions at 296K (IDA reference below). Another reference measured (see Table 1 in Zhan 2016) for 8-12 microns at 1km coefficients ranging from 0.004 in dry air, to 0.16 at 10% RH, to 0.75 at 50% and reaching 1.5 at 100% RH. At 50% relative humidity and 5 meters distance the reduction is less than a percent, reducing the measured temperature by 130mK, which does indeed produce a significant effect. Below 2 meters, this effect is negligible (less than 0.1C) for even the largest RH changes. Depending primarily on the operating distance range of the equipment, a correction for relative humidity may or may not be necessary.

These calculations imply negligible real-world impacts of the challenging implementation requirements of the standard that most sites will struggle to meet in all conditions. It is true that humidity and other ambient conditions could affect the face temperature, but it should be made clear in the standard that these are not likely to affect the equipment itself and further, that providers of such equipment must not excuse equipment failures in realistic implementations. Otherwise, providers of thermographic systems will (and routinely do today) rely on implementations not meeting these unrealistic and unnecessary requirements to abrogate actual performance failures in their equipment. A good use of consensus could be to arrive at meaningful and actionable criteria of what can be safely ignored and what the maximum impact could be of some potential confound, to help implementation sites determine what is and is not likely to be causing a problem in febrile screening.

References

1. Izhar, LI and Petrou, M, "Chapter 2: Thermal Imaging in Medicine", Advances in Imaging and Electron Physics, Volume 171, ISSN 1076-5670, DOI:10.1016/B978-0-12-394297-5.00002-7.
2. Wissler EH, "Pennes' 1948 paper revisited", J Applied Physiology, 81:1, 35-41, 1998, https://doi.org/10.1152/jappl.1998.85.1.35
3. Tan, JH et al, "Infrared thermography on ocular surface temperature: A review". Infrared Physics & Technology 52(4):97-108. DOI: 10.1016/j.infrared.2009.05.002

4. Slettedal, J and Ringvold, A, “Correlation between corneal and ambient temperature with particular focus on polar conditions”, Acta Opthalmologica, 93:5, 2015, 422-426, https://doi.org/10.1111/aos.12657
5. Zhou, Y et al, “Clinical evaluation of fever-screening thermography: impact of consensus guidelines and facial measurement location”, J. of Biomedical Optics, 25(9), 097002 (2020). https://doi.org/10.1117/1.JBO.25.9.097002
6. IEC 80601-2-59:2008, Medical electrical equipment – Part 2-59: Particular requirements for basic safety and essential performance of screening thermographs for human febrile temperature screening
7. ISO/TR 13154:2009, Medical electrical equipment – Deployment, implementation and operational guidelines for identifying febrile humans using a screening thermograph.
8. Roberts RE, Biberman LM, Selby JEA, Infrared Continuum Absorption by Atmospheric Water Vapor in the 8-12 um Window. April 1976, Institute for Defense Analysis, Science and Technology Division, Virginia
9. Y-c. Zhang, Y-m. Chen, X-b. Fu, C. Luo, The research on the effect of atmospheric transmittance for the measuring accuracy of infrared thermal imager, Infrared Physics & Technology (2016), doi: http://dx.doi.org/10.1016/j.infrared.2016.06.025
10. Fujishima H, Toka I, Yamada M, Sato N, Tsubota K, Corneal Temperature in Patients with Dry Eye Evaluated by Infrared Radiation Thermometry, British Journal of Ophthalmology 1996; 80: 29-32.
11. Kessel L, Johnson L, Arvidsson H, Larsen M, The Relationship Between Body and Ambient Temperature and Corneal Temperature. Investigative Ophthalmology & Visual Science, December 2010, Vol 51, 12: 6593-6597.