The Significance of 37 °C: Temperature, Ethnicity, and the Architecture of Normality

Anaïs Lohier
Yale School of Public Health, Yale University, New Haven, CT, United States

Abstract

Background: The canonical value of 37 °C reflects a historical convention rather than a universal biological norm, yet it continues to be treated as one. Health-informatics systems reproduce similar conventions by encoding ethnicity as fixed administrative categories whose analytic meaning is uncertain. How such inherited labels relate to temperature measurements is largely unexamined, particularly with respect to the ways data structures themselves shape what appears as variation.

Objective: To assess whether administratively encoded ethnicity categories generate measurable temperature differences under highly standardized measurement conditions, and to evaluate what this indicates about their role and limitations within informatics classification systems.

Methods: A secondary observational analysis was conducted using publicly available infrared thermography data collected under tightly controlled conditions. The analytic sample was restricted to adults aged 18-30 and to the two most accurate facial temperature measures – full-face maximum temperature (T_max) and inner-canthus temperature (T_CEmax) – recorded using Infrared Cameras Inc. (ICI) infrared thermometry device. One-way fixed-effects ANOVA assessed mean differences across six encoded ethnic categories, with full assumption checks and effect-size estimation.

Results: Small but statistically reliable differences were detected for both inner-canthus temperature (T_CEmax) ($F(5, 882) = 6.73$, $p = 3.64\times10^{-6}$, $\eta^2 = 0.04$) and full-face maximum temperature (T_max) ($F(5, 882) = 4.15$, $p = 0.00099$, $\eta^2 = 0.02$). ANOVA assumptions were adequately met for both outcomes.

Conclusions: Ethnicity categories encoded in the dataset corresponded to modest temperature variation, not evidence of biological distinction. Instead, the findings illustrate how administrative categories can become statistically consequential under high measurement precision, highlighting the need for refined ontologies, personalized baselines, and multi-scale modeling in health-informatics design.

1. Introduction

1.1 Background

For more than a century, 37 °C has defined "normal" body temperature, yet this value persists more by convention than biological universality. Wunderlich's nineteenth-century measurements [1], drawn from a narrow clinical population and specific procedural norms, stabilized 37 °C as a benchmark even though contemporary evidence shows that temperature varies by age, sex, circadian timing, measurement site, and individual physiology. Temperature thus functions as a paradigmatic biometric: historically contingent, clinically foundational, and shaped by the instruments and standards that make it measurable. A similar logic governs ethnicity in health-informatics systems, where complex identities are compressed into administrative categories so they can enter data schemas and analytic pipelines. Once encoded, both temperature norms and ethnicity labels acquire an interpretive authority that exceeds their empirical grounding, influencing how biomedical systems classify bodies and render certain distinctions visible.

1.2 Scientific Problem

Contemporary informatics systems treat temperature and ethnicity as stable, self-evident fields, yet it remains unclear whether these encoded categories generate discernible structure in physiological measurements or merely reflect the architecture of the data model. This project tests whether administratively defined ethnicity categories correspond to measurable variation in facial temperature under tightly controlled imaging conditions. Several challenges complicate this question: ethnicity in informatics is an institutional abstraction with uncertain correspondence to phenotype or ancestry; surface temperature fluctuates with vasomotor tone, perfusion, evaporative cooling, and ambient conditions even under standardized protocols; and infrared thermography is sensitive to calibration, emissivity assumptions, and device performance. When environmental and procedural variability are minimized, however, any residual differences across encoded categories become analytically informative not as indicators of physiology, but as reflections of how data systems partition human variation.

1.3 Objective

The objective of this study is to determine whether the ethnicity categories encoded in a rigorously standardized thermographic dataset correspond to measurable variation in two validated facial temperature measures (T_max and T_CEmax). More broadly, the study examines how administratively defined demographic fields behave within health-informatics analyses and what their statistical patterns imply for classification, interpretation, and decision-support systems that rely on routinely collected demographic data. Because the dataset is publicly available and the analytic workflow is fully documented, the results are reproducible and available for independent verification.

1.4 Overview

This paper is organized to connect the historical development of thermometry with the informatics systems that encode human difference. Sections 2 and 3 together constitute the literature review. Section 2 traces how temperature became a clinical and epistemic object through evolving instruments, modern device standards, and the constraints of infrared thermography. Section 3 shifts to health informatics and examines how measurement practices and data models translate complex biosocial identities into computable categories. Section 4 describes the thermographic dataset and analytic constraints, and Section 5 outlines the methodological and philosophical rationale for using ANOVA to interrogate its categorical structure. Section 6 presents the empirical results, and Section 7 interprets them across clinical, informatics, epidemiologic, and philosophical domains. Section 8 offers representational and methodological recommendations, and Section 9 discusses limitations and future directions. The paper concludes that the small but detectable differences observed across encoded ethnic categories reflect not physiology but the representational and measurement structures through which contemporary informatics makes human variation legible.

2. Thermometry: Devices, Error, and the Interpretation of Temperature

2.1 Fundamental Origins

Thermometry is the quantitative measurement of human body temperature using instruments that register the thermal energy at the site of contact, providing an objective physiological signal for diagnosis and observation. Its validity depends on the location of measurement, the physics of heat transfer, and the accuracy and calibration of the device, since thermometers detect only local heat and are sensitive to technical and environmental factors [1-5].

From Sensory Judgment to Numerical Authority

The history of thermometry traces not merely the invention of an instrument but the transformation of medical knowledge itself from a qualitative observation to numerical authority. The idea that heat could reveal the hidden order of life long predated the thermometer. In antiquity, Galen had described organs as relatively “hot” or “cold,” classifying parts of the body by their perceived warmth, but only through the senses [2]. In the late sixteenth century, Joannes Haslerus’s De Logistica Medica (1578) transformed this qualitative legacy into calculation [6]. Haslerus argued that medicine must be restored to the precision of number and proportion, proposing that each body possessed a natural degree of heat or cold determined by latitude, season, and age, and could be computed mathematically – such that an inhabitant of Antwerp, for example, would be “of the first degree of cold” [2, 6]. A schematic from Haslerus’s system is reproduced in Appendix A.1. In essence, Haslerus helped reposition medicine toward numerical scale and mathematical regularity.

By the seventeenth century, this confidence in quantification had become a defining feature of natural philosophy. As Shryock observed, early iatromathematicians believed that “phenomena perceptible only to reason” could be revealed through measurement [7]. Santorio (1561-1636) exemplified this idea through early instruments aimed at quantifying hidden physiological processes [1, 2]. His experiments marked one of the first sustained efforts to measure such processes, aiming to render the body’s fluctuations legible (see Appendix A.2).

Early Experimental Thermometry

Across the eighteenth and early nineteenth centuries, thermometry shifted from a philosophical curiosity to a reproducible clinical method. Anton de Haen demonstrated that fever followed a daily course through repeated bedside measurements. In the nineteenth century, Jacques Breschet and Antoine Becquerel showed that inflamed tissues were warmer than healthy ones, and Gabriel Andral sought to “map out the courses of temperature” in disease, marking the transition from isolated readings to patterned, temporal observation [1, 2].

These cumulative investigations converged in the work of Carl Reinhold August Wunderlich, whose *Medical Thermometry and Human Temperature* (1871) systematized thousands of cases and millions of observations [1]. Wunderlich defined 37 °C as the mean temperature of health and 38 °C as the threshold of fever, asserting that “thermometry in disease is an objective, physical method of investigation.” His manual provided precise procedural instructions:

> “The instrument may be centigrade or Fahrenheit, but it must be accurate. If self-registering, the nurse can use it at stated times, and the physician can read it at the next visit. The bulb is to be inserted in the axilla, previously dried if moist from perspiration, just beneath the fold of the pectoralis major muscle… It is left in situ… for eight to ten minutes… the degree is then read off and recorded on a blank diagram.” [1]

Critical Reflection

Ironically, these meticulous directions reveal the persistence of subjectivity within an enterprise devoted to its elimination. As Engel later observed, “the observational act is a unitary deed of which our choice is an active subjective component” [8]. Temperature readings still depended on touch, timing, posture, and interpretation, and without a reliable external reference, precision offered little guarantee of accuracy. The mercury-glass thermometers of the period routinely drifted and could not be easily checked against a stable standard, leaving even “standardized” measurements open to quiet error [9].

Early Considerations of Ethnicity

In the 1850s, efforts to relate temperature to human variation brought ethnicity into view, but the evidence was limited and largely anecdotal. Wunderlich briefly suggested that observed differences might reflect factors such as race or climate, yet he provided almost no empirical support for this claim [1]. His only substantive evidence came from a passing remark in *Livingstone's Travels in South Africa*, where Livingstone recorded his own oral temperature at 100 °F and that of local inhabitants at 98 °F [10]. Wunderlich immediately undercut its authority, reminding readers that "a single observation of temperature is always an imperfect and unsatisfactory standard" [1]. The episode shows how tenuous the earliest attempts were to link body temperature to purported ethnic or racial differences. Despite later recognition of wide inter- and intra-individual variation [3, 4, 11, 12], Wunderlich's number has endured, shaped as much by the authority of his method as by the universality of his sample [1].

2.2 The Contemporary Thermometric System

Contact and Infrared Modalities in Modern Thermometry

Modern clinical thermometry relies on two broad modalities: contact instruments (such as mercury-in-glass and bimetallic thermometers) and optical systems, most prominently infrared thermography (IRT). As Raiko and colleagues note, each modality detects only the thermal signals available through its physical interface, thereby defining what can be measured directly, what must be inferred, and where systematic error may enter [13].

IRT operates differently from contact methods. Rather than measuring the temperature of internal tissues or even the probe-skin interface, it detects infrared radiation emitted from the skin surface, a signal shaped by skin emissivity, local blood flow, and the temperature gradient between body and environment [13]. Because emissivity expresses how closely a surface behaves like an ideal blackbody, the accuracy of any inferred temperature depends on its stability. Clinical IRT systems generally assume a skin emissivity of approximately 0.98, and even small deviations from this value can generate errors approaching 1 °C, a range large enough to alter fever screening outcomes [13].

Optical Pathways, Pigmentation, and the Stability of the Thermal Signal

A central question for surface-based thermometry is whether skin pigmentation alters infrared temperature readings. Empirically, the answer is clear: human skin emissivity does not differ meaningfully across pigmentation levels. In a controlled study of 65 adults representing a wide range of Fitzpatrick types, emissivity values clustered tightly between 0.96 and 0.99 with no significant group differences [14]. A larger study of 289 male volunteers (including Caucasian, Black, and individuals of mixed ancestry) likewise found that melanin content did not change emissivity or reflectivity when measurements were taken under standardized conditions with a clinical infrared thermometer [15].

Even so, small but systematic temperature differences still emerge across pigmentation groups, and these arise through mechanisms other than emissivity. In the 289-participant study, for example, the contrast between temple and wrist temperatures was slightly greater among participants with darker skin, despite uniform devices, fixed emissivity parameters, and constant measurement distance [15]. Similar patterns appear in more recent controlled cooling experiments, where identical thermal perturbations produced apparent temperature differences of nearly 1 °C between the lightest and darkest skin tones [17]. Because emissivity appears stable across skin tones [14], these discrepancies are more plausibly attributable to non-emissivity optical and device-processing factors than to intrinsic thermoregulatory physiology, though this dataset cannot distinguish these mechanisms directly [3].

These optical and device effects take on added significance in datasets that encode ethnicity rather than direct measures of pigmentation. Constitutive skin color is a composite trait shaped by melanin content, genetic ancestry, environmental exposure, and adaptive responses [16]. Although ethnicity categories do not capture this complexity, in many populations they correlate broadly with skin pigmentation. As a result, pigmentation-linked measurement artefacts can become mapped onto administrative ethnic labels, creating the appearance of group-level temperature differences even when core physiology is the same.

Surface Measurement, Core Physiology, and the Limits of Inference

Infrared thermography measures surface temperature, not the tightly regulated core temperature that clinical practice treats as a physiological constant. Core temperature remains near 37 °C through hypothalamic control of heat production and loss, whereas surface temperature varies with local perfusion, evaporative cooling, ambient conditions, and the geometry of the measurement site [3, 4]. Even under ideal protocols, the skin does not provide a direct reading of internal thermal state.

Technical factors introduce additional variability. Infrared cameras differ in calibration stability, sensor sensitivity, thermal resolution, and in the algorithms used to convert infrared radiation into a displayed temperature [4, 5]. Minor discrepancies in calibration or signal processing can produce shifts on the scale of tenths of a degree – small in absolute terms but large enough to complicate interpretation when researchers are examining subtle group differences. Even when emissivity is fixed and pigmentation effects are minimized, device behavior still contributes noise that is difficult to separate from natural physiological fluctuations.

Because surface infrared measurements integrate thermodynamic, environmental, and device-related influences, they do not directly represent core physiology. When group differences are small, their meaning becomes difficult to determine since they may reflect these influences rather than biological variation. Such differences must therefore be interpreted with caution and cannot be presumed to indicate true thermoregulatory change.

2.3 Standards, Calibration, and Data Structures

Early standards did little to address the fragility of clinical thermometry. As experimental approaches proliferated, instruments remained prone to drift, and twentieth-century regulations offered only modest safeguards. Standard mercury-glass thermometers were legally required to display 96-106 °F (a range far wider than clinicians needed) and restoring the column to a true baseline often required specialized equipment [9]. A 1975 evaluation found that seven of twenty-five devices (28%) exceeded allowable accuracy limits, substantiating the claim that national standards such as CS1-42 and CS1-52 constrained manufacturers without ensuring clinical reliability [9]. Numerical readings thus entered practice on the authority of instruments whose calibration and stability were never fully secured.

Modern regulations attempt to remedy these gaps. ISO 80601-2-56 specifies accuracy ranges for clinical thermometers, typically ±0.3 to ±0.4 °C, and ASTM E1965 defines permissible errors for infrared devices [5]. Yet achieving these limits in routine practice is difficult: infrared and electronic sensors must perform across heterogeneous conditions even though their accuracy depends on ambient temperature, distance, angle, emissivity, and component stability. All such instruments are susceptible to material fatigue and calibration drift and require periodic verification against reference standards [4]. Recent hospital evaluations show that while contact devices often meet expectations, over half of non-contact forehead and ear thermometers fail ISO/ASTM criteria even under controlled conditions, with uncertainty bands that exceed clinically meaningful thresholds [5].

The limitations of these devices became widely visible during the COVID-19 pandemic, when the World Health Organization (WHO) endorsed thermal imaging systems and handheld infrared thermometers that were rapidly deployed for mass screening [18]. Many were installed without blackbody reference targets and relied on inconsistent facial sites, which produced large variability and frequent false negatives [18, 19]. Subsequent large-scale IRT evaluations showed that this instability was not a failure of particular models but a predictable consequence of using precision instruments outside the tightly controlled conditions required to maintain reliable calibration [20].

These device-level constraints have direct analogues in the informational substrate into which clinical measurements are encoded. Medical Event Vectors (MEVs) store “symptoms, signs, laboratory treatments, diagnoses, and virtually all other medical data” in formats optimized for durability and retrieval rather than analytic depth, compressing clinical nuance and allowing systematic errors to persist [21]. Similar losses occur when continuous variables, such as temperature and blood pressure, are recoded into ordinal labels (“mild,” “moderate,” “severe”) that facilitate documentation but discard granularity [19]. And in laboratory medicine, where nearly 80 percent of results pass through automated systems without direct human interpretation, the structure of data fields often exerts more influence on downstream analysis than the original biological signal [22].

Calibration, whether mechanical or semantic, governs how closely a measurement reflects the phenomenon it is meant to represent. A thermometer may meet formal accuracy limits yet still blur physiologically meaningful variation, and a coding scheme may satisfy documentation requirements while removing distinctions that matter for analysis. Bias, drift, or compression introduced at either level carries forward into downstream interpretation and shapes estimates of normality, fever thresholds, and the apparent structure of demographic categories. Devices and data structures form a measurement architecture that makes temperature data usable while also defining the limits of what can be inferred from it.

2.4 How Thermometry Constructed Normality and Medical Evidence

The clinical importance of thermometry is not limited to detecting fever; it lies in how temperature became a vital sign that organizes medical judgment. Contemporary practice treats body temperature alongside heart rate, blood pressure, and respiratory rate as one of four principal indicators of “fundamental body functionality and efficiency” [3]. Normal body temperature (normothermia) is understood as a prerequisite for proper physiological function, whereas sustained hyperthermia or hypothermia perturbs metabolism, impairs cognition, and can cause tissue damage [3]. Yet, as these historical foundations suggest, the very idea of a “normal” temperature is not inherent to physiology; it is a historical construct shaped by the instruments and conventions that first defined its boundaries.

Wunderlich’s benchmark was not merely descriptive; it was normative. Statistical averages became criteria against which individuals were judged. What counted as “normal” depended on the specific instruments and measurement sites that supplied the data. As Canguilhem noted, instruments “give a result,” but diagnostic value comes only through interpretation [23]. Temperature becomes informative when clinicians connect a numerical value to thermoregulatory physiology, site selection, the patient’s condition, and the question at hand. Yet the 37 °C convention persisted because it offered an apparently objective standard at a time when medicine sought numerical clarity.

Clinical reviews emphasize that infrared thermography nonetheless occupies a distinctive role in modern temperature assessment. IRT can detect very small thermal differences and visualize spatial patterns linked to inflammation, perfusion changes, neuropathy, and other surface-manifesting processes [24]. Its interpretive value, however, depends on recognizing its limits: surface temperature is shaped by radiation, convection, local blood flow, and ambient conditions, whereas core temperature is kept within a narrow hypothalamic range. Real-time readings support screening and monitoring, but isolated values provide limited physiological insight without contextual information about measurement conditions [3]. Each thermometric device reveals only one layer of the body's thermal physiology, and clinical meaning arises only when that device's constraints are taken into account. Through this a paradox emerges, being "the more facts we learn the less we understand," notably when complexity grows faster than the models used to interpret it, a caution that applies equally to high-resolution thermal data [25].

Thermometry thus performed a double operation: it rendered the body legible as a sequence of numerical values and organized those values into distributions that could be labeled "normal," "febrile," or "hypothermic." From Galen's qualitative assessments of heat and cold to the threshold-based rules of contemporary screening, temperature became something that could be calculated, compared, and categorized. The authority of numerical precision emerged not only from physiology but from the instruments and practices that made temperature measurable, creating an architecture of normality in which devices produced readings, standards defined acceptable error, and thresholds converted those readings into judgments. In this way, measurement practices do not simply record normality; they help construct it.

3. Measurement, Representation, and the Ontology of Human Difference in Health Informatics

3.1 Introduction

The same measurement logic that transformed bodily variation into numerical form now underlies contemporary informatics. This section examines how that logic structures categories of human difference, a process illustrated in Appendix Figure B.1. Because measurement practices stabilize distinctions, tracing their logic clarifies how informatics inherits and amplifies the classificatory assumptions embedded in earlier forms of quantification.

3.2 The Datafication of Human Diversity and Resulting Ontological Constraints

Genetics was one of the earliest attempts to translate human variation into a computable form. Even after population genetics showed that most diversity lies within groups [26], researchers continued to rely on broad labels such as Black, White, and Asian – categories applied "without definition" yet carrying "powerful ramifications beyond the domain of science" [27, 28].

When ethnicity is formalized as a data field, its lived complexity is reduced so it can function within structured analytic systems. In electronic health records, clinicians select from a narrow menu of categories that must substitute for far broader identities and contexts. This simplification reflects the logic of data infrastructures, which rely on stable, computable entries to function

[29-32]. But these same structures narrow the space of interpretation, directing attention toward predefined distinctions rather than those that emerge in practice.

The core issue is not simply that diversity becomes a variable, but that the variable acquires ontological authority. Once formalized, categories become part of the conceptual terrain the system treats as self-evident [27], shaping how future distinctions are built and limiting which questions can even be asked. As Engel, drawing on Einstein notes, moving to a broader framework does not erase earlier structures; it reveals their scope and limits [8]. In this way, informatic ontologies do more than encode categories, they naturalize them, embedding prior classificatory assumptions into the foundations of new models of human variation.

3.3 The Politics of Representation in Normality

Informatics does not just generalize from populations rather it generalizes from those populations that are successfully encoded. Administrative categories function as gates to analytic visibility, determining which bodies can contribute to norms and which are structurally absent from their construction. As Birney notes, genetics continues to operationalize racialized groupings even while rejecting their biological coherence, because such categories remain indispensable to data organization and comparison [33]. Once embedded, these groupings stand in for populations they cannot fully represent, yet the statistical patterns they yield circulate as if they were population truths. Norms produced in this way are therefore not collective baselines but artifacts of selective inclusion, shaped by sampling asymmetries, category design, and institutional convenience. Canguilhem's critique of the norm takes on a new register here: informatics does not merely confuse averages with ideals, it stabilizes partial representations as reference points that govern interpretation downstream [23]. Engel's concern thus reappears not at the level of clinical judgment, but at the level of data architecture itself, where what cannot be cleanly encoded fails to register as difference at all [8, 31].

Critical Reflection

Population genetics complicates this. Bryc found over 99% correspondence between self-reported and genetic ancestry in a large U.S. dataset [34]. Social categories can align with biological and historical lineages at the population level, much as population-level temperature averages can meaningfully describe trends. But correspondence does not confer biological essentialism; a mean does not make a norm [23]. Informatics must therefore navigate categories that are socially constructed and statistically patterned without collapsing one into the other.

Representational Stakes

Contemporary informatics thus operates through representational choices that shape which forms of human difference can appear as analytic structure. Rather than resolving these questions at the level of theory, they can be examined by observing how encoded demographic categories behave within a concrete dataset. The thermographic dataset analyzed here serves as a case through which to trace how administratively defined ethnicity functions once it is fixed as a computational field. The following section introduces the data and materials used to examine how these representational commitments become visible in empirical analysis.

4. Data and Materials

4.1 A Note on Secondary Data Analysis

Secondary analysis imposes its own methodological constraints. As Sun argues, such work must be guided by a prespecified analytic plan, since flexible exploration risks data dredging and inefficient inference [35]. Within that framework, the empirical question here is intentionally narrow: do the encoded ethnicity values produce any measurable temperature differences under highly controlled conditions. Unlike primary studies, secondary analyses must accept the assumptions, variable definitions, and measurement decisions already built into the dataset. This project therefore adopts a transparent and tightly bounded analytic strategy, not to estimate population effects, but to examine how the dataset's representational choices structure the patterns that emerge under controlled analytic conditions.

4.2 About the Original Data

This dataset originates from a clinical thermographic study evaluating the accuracy of infrared thermographs (IRTs) for detecting elevated body temperature. Researchers collected facial thermal and visible images alongside oral temperature readings from more than 1,100 participants using two systems, a FLIR device (IRT-1) and an Infrared Cameras Inc. (ICI) device (IRT-2). All data were de-identified and collected under protocols approved by the U.S. Food and Drug Administration (FDA) and Institutional Review Boards (IRBs) [19].

The study was motivated by ongoing uncertainty about the calibration and clinical accuracy of IRT systems. Although international standards such as IEC 80601-2-59:2017 provide guidance for laboratory evaluation of fever-screening devices, there are no consensus methods for assessing clinical accuracy, and existing technical reports (for example, ISO/TR 13154:2017) outline deployment practices rather than validation procedures [19]. The investigators therefore implemented a standardized imaging protocol designed to reflect real-world variability in devices, subjects, and environmental conditions.

Strict quality controls governed data collection. Participants acclimated indoors for 15 minutes, and measurements were taken within a narrow ambient temperature range. Subjects were excluded if their two oral temperature readings differed by more than 0.5 °C or if only one reading was recorded – common indications of motion, improper probe placement, or recent ingestion of hot or cold substances [20]. Of 1,115 enrolled individuals, 6 had incomplete records and 56 were removed due to inconsistent oral temperatures. Image-level exclusions for motion artifacts yielded final samples of 1,020 subjects for IRT-1 and 1,010 for IRT-2 [19].

Facial regions of interest were fixed across participants (Appendix C.1), reducing variability from inconsistent measurement sites and supporting direct comparison of the temperature variables analyzed here.

4.3 Why this Dataset?

This dataset is well suited to the scientific problem because it isolates measurement noise to an unusual degree: participants were acclimated, imaging conditions were standardized, and the two most accurate temperature variables (T_max and T_CEmax) demonstrated strong agreement with core temperature. By minimizing environmental, device, and procedural variability, the dataset limits non-physiological sources of temperature variation. This controlled context therefore allows any observed differences across encoded ethnic categories to be interpreted chiefly as features of the dataset's classificatory structure rather than artifacts of inconsistent measurement.

A further advantage is that the dataset is fully publicly accessible through PhysioNet, a widely used National Institute of Health (NIH)-supported repository for open clinical data [20]. Public availability ensures that all analyses presented here can be independently reproduced or extended, and allows the methodological choices of this project – including variable selection, preprocessing, and statistical modeling – to be verified directly from the original source. This openness strengthens the scientific validity of the study by ensuring transparency and facilitating replication, both of which are essential for evaluating whether the observed patterns arise from the data itself or from analytic interpretation.

4.4 Variables and Preprocessing

Variable Selection (T_max and T_CEmax)

The analysis focuses on two validated facial temperature variables: T_max, the full-face maximum temperature, and T_CEmax, the maximum temperature in the inner canthus region. These measures showed the strongest agreement with oral temperature in the original evaluation ($r \approx 0.78$–0.79; $AUC \approx 0.95$–0.97) [19], with diagnostic performance summarized in Appendix C.2. Both variables were averaged across four sequential imaging rounds to reduce random noise and improve reliability.

The choice of these sites is also supported by physiological and thermometric considerations. Surface temperature varies widely across the body, and each measurement site reflects different patterns of conduction, convection, radiation, and evaporative heat loss [4]. The inner canthus is less exposed to ambient variation and is perfused by vessels that track core temperature more closely, which makes it one of the most stable anatomical sites for infrared measurement [3]. In contrast, forehead-based readings are more sensitive to emissivity changes, camera angle, distance, and environmental conditions, factors that have been shown to reduce accuracy in non-contact thermometry [5]. Restricting the analysis to T_max_mean and T_CEmax_mean therefore minimizes site-specific noise and aligns with established evidence on which facial regions yield the most dependable thermal information under controlled conditions.

Device Selection (IRT-2)

Although both infrared thermographs performed well, IRT-2 showed higher accuracy, better calibration stability, and lower spatial noise, particularly for the T_max and T_CEmax variables (for example, T_max AUC = 0.968 for IRT-2 vs. 0.951 for IRT-1) [19]. Because the analysis depends on detecting small between-group differences, selecting the more precise device reduces the likelihood that observed patterns reflect instrument variability rather than structure in the data.

The dataset provides no calibration records for either device, so calibration quality cannot be independently assessed. The analysis therefore relies on the performance metrics reported in the original evaluation, which identified IRT-2 as the more accurate system under the study's standardized conditions [19]. External comparisons support this choice. Independent studies find that ICI devices outperform comparable FLIR systems in accuracy, precision, and the detection of fine-grained spatial temperature differences [36]. Using IRT-2 therefore provides the strongest available measurement basis for the present analysis. Full device-performance comparisons appear in Appendix C.2.

Age Restriction (18-30)

Because the dataset is overwhelmingly composed of young adults, the analysis is restricted to participants aged 18 to 30. This reflects the demographic structure of the sample rather than a theoretical boundary because roughly 94 percent of participants fall within this range, while very few are older than 40 (Appendix C.3), and the authors note that the cohort is not representative of the general population [20]. Age influences thermoregulation in that older adults tend to show slightly lower and more variable core temperatures, but the small number of older individuals in this dataset makes it impossible to model such differences reliably [37]. Restricting the analysis to the age range the dataset meaningfully represents therefore reduces physiologic heterogeneity and prevents instability from underpowered subgroups, which helps preserve statistical validity and interpretability.

Ethnicity Variable

Ethnicity is analyzed exactly as recorded in the dataset, which lists six administratively defined categories: Asian, Black or African American, Hispanic/Latino, White, American Indian or Alaskan Native, and Multiracial. Because the original documentation does not describe how these labels were defined or collected [20], the analysis treats them as fixed administrative codes rather than validated sociocultural or phenotypic constructs. Full distributions appear in Appendix C.3. Infrared temperature measurements can be influenced by phenotypic factors such as skin pigmentation [15, 17], yet the dataset includes neither Fitzpatrick skin type nor any direct measure of skin tone [20]. Without these variables, pigmentation-related optical effects cannot be evaluated independently and would be absorbed into the administrative categories if they exist. For this reason, any observed group differences should be understood as features of the dataset's representational structure rather than evidence that the encoded labels reflect physiologically meaningful distinctions.

A Note on Gender

Gender is retained exactly as recorded in the dataset (Female and Male), but it is not incorporated as a primary analytic variable. Although gender can influence baseline temperature through hormonal, vascular, and circadian factors [1, 5], including it here would introduce additional physiological variability without advancing the central analytic question. The goal of this study is to assess how ethnicity, as encoded in the dataset, behaves under tightly standardized thermographic conditions. Stratifying the analysis by both gender and ethnicity would multiply subgroup combinations and substantially reduce stability within several already small ethnic categories, as shown in Appendix C.3. For these reasons, gender is preserved descriptively but excluded from the main between-group comparisons.

4.5 Sample Characteristics

After applying the analytic restrictions – using IRT-2 measurements, retaining only T_max and T_CEmax, averaging across four imaging rounds, and limiting the sample to participants aged 18-30 – the final dataset included 888 individuals with complete demographic information.

Ethnic composition of the analytic subset showed heterogeneous but unevenly distributed representation (Table D.1). Just over half of participants identified as White (50.23%), followed by Asian (25.00%), Black or African American (13.96%), Hispanic/Latino (5.74%), Multiracial (4.73%), and American Indian or Alaskan Native (0.34%). These proportions provide necessary context for interpreting between-group comparisons, particularly for categories with limited sample sizes.

Mean temperature values showed modest variation across ethnicity groups for both infrared measures. Inner-canthus temperatures (T_CEmax_mean) ranged from approximately 35.57 °C to 35.88 °C, with standard deviations between 0.14 °C and 0.68 °C. Full-face maximum temperatures (T_max_mean) showed a similar pattern, ranging from 35.99 °C to 36.19 °C, with standard deviations between 0.05 °C and 0.66 °C. Because mean values can obscure extreme observations, minimum and maximum values were also examined to evaluate potential outliers. Across all groups, the observed ranges fell within physiologically plausible bounds (roughly

34-39 °C), and no group exhibited disproportionately high or low extremes. The narrow within-group ranges support the interpretation that no anomalous measurements materially influenced the descriptive statistics. Detailed numerical summaries for each group are provided in Appendix D.2.

Data completeness was also evaluated both before and after deriving participant-level temperature means. Before averaging, the proportion of missing values across the four measurement rounds ranged from 5.35% to 13.08% (Table D.3). After averaging, all analytic variables contained 0% missing data (Table D.4), indicating complete data integrity in the final subset used for ANOVA.

5. Methods

5.1 Study Design: Philosophical, Theoretical, and Scientific Rationale
This analysis uses a comparative-effectiveness, secondary observational design to test whether the dataset's administratively encoded ethnicity categories correspond to detectable variation in a temperature signal measured under tightly standardized conditions. Because the original study was designed to evaluate infrared thermography rather than human difference, the present analysis builds on its controlled ambient temperature, repeated imaging rounds, and validated facial sites [19, 20]. Within a comparative-effectiveness framework, the goal is estimation rather than causal inference, emphasizing effect sizes and confidence intervals to characterize the magnitude and precision of within-dataset differences [35]. One-way ANOVA is therefore used as a descriptive inferential tool to assess whether the dataset's predefined categories exhibit measurable structure beyond random variation. Analytic validity depends on identifying the constraints of the dataset and evaluating whether those constraints give rise to detectable patterns in the observed measurements.

Philosophical Logic: Why Group Comparison is the Right Structure
The decision to compare groups reflects a deeper philosophical issue: once encoded as variables, complex biosocial identities become fixed computational objects that invite statistical comparison. Group-based analysis does not assume biological kinds; rather, it interrogates how categories gain stability through measurement. Canguilhem warns that numerical distinctions are easily mistaken for natural ones when categories are treated as if they merely record reality rather than constitute it [23]. Fabrega similarly emphasizes the gap between lived identity and its administrative representation in clinical systems [38]. In this light, ANOVA operates not as a search for inherent physiological differences but as a way to examine how informational categories make certain distinctions visible.

Theoretical Logic: Why ANOVA Suits the Data and Question

The dataset includes six discrete, non-overlapping, non-ordered ethnic labels. The analysis is therefore concerned with the specific categories encoded by the system, not with sampling from a broader population of possible labels. These factors function as fixed effects, consistent with Ståhle's observation that fixed-effects ANOVA is appropriate when the investigator is interested in the particular levels of a factor rather than an underlying superpopulation [39]. The dependent variables — T_max_mean and T_CEmax_mean — are continuous, and averaging multiple imaging rounds yields stable participant-level estimates. These properties mean that the theoretical structure of ANOVA aligns directly with the structure of the dataset: discrete categorical predictors paired with continuous outcomes.

Scientific Logic: Why This Method Advances Health Informatics

Health informatics inevitably reduces complex human variation into computable categories, and ANOVA offers a way to test whether those categories correspond to distinct underlying distributions or simply reflect representational design [38]. Under tightly standardized physiological conditions, the method evaluates whether the encoded ethnic labels carry measurable informational content or whether they function as inherited artifacts of documentation. As Cimino argues, controlled vocabularies are useful only when their categories correspond to discernible differences rather than convenience [40]. ANOVA therefore becomes a direct test of the informational value of these administrative fields, clarifying which classifications support reliable inference and which risk misleading downstream analytics. In doing so, the analysis treats the ontology of the dataset itself as an object of scrutiny and makes visible how data structures shape and constrain what analytic systems can detect.

5.2 Statistical Framework and Assumptions

Statistical significance was evaluated using a two-sided α level of 0.01. This threshold was selected given the large sample size and the multiple group comparisons involved, ensuring that only the more pronounced differences were flagged as statistically significant. Effects with $p < 0.01$ were interpreted as significant, whereas $p \geq 0.01$ was not considered evidence of a difference. In line with Sun's emphasis on estimation in comparative-effectiveness research, all hypothesis tests are accompanied by effect sizes and confidence intervals to characterize the magnitude and precision of observed differences rather than relying solely on binary significance [35]. This approach maintains coherence with the study's broader analytic aim: evaluating whether the dataset's encoded categories correspond to measurable structure in the temperature distribution.

Because the purpose of ANOVA is to determine whether observed between-group differences exceed what can be attributed to random error, its validity depends on whether the model's assumptions correctly characterize that error structure. Ståhle emphasizes that ANOVA's F-ratio is interpretable only when its pooled residual variance is a faithful estimate of underlying noise; if groups differ in variance or if residuals deviate substantially from normality, the F-statistic can reflect artifacts of dispersion rather than genuine differences between means [39]. This matters directly for the present study's analytic aim, which asks whether encoded categories correspond to meaningful structure in the temperature distribution: such structure cannot be inferred unless the noise model is sound. Therefore, residual normality and homogeneity of variances were evaluated using Q-Q plots, the Shapiro-Wilk test, and the Brown-Forsythe version of Levene's test. Across all analyses, effect sizes and confidence intervals accompany p-values to maintain an emphasis on estimation rather than binary significance.

T_CEmax_mean ANOVA Assumption Validation

For T_CEmax_mean, residual diagnostics indicated that the ANOVA assumptions were reasonably satisfied. The Q-Q plot showed approximately normal residuals with modest tail deviations consistent with large sample sizes. Although the Shapiro-Wilk test was statistically significant ($W = 0.90$, $p < 2.2\times10^{-16}$), this reflects the test's sensitivity to minor departures from normality rather than substantive distortion of the residual distribution. Variance equality across ethnicity groups was supported by the Brown-Forsythe test ($F = 0.57$, $p = 0.72$). These diagnostics, presented in Appendix E.1, indicate that the normality and homoscedasticity assumptions for one-way ANOVA were adequately met for T_CEmax_mean.

T_max_mean ANOVA Assumption Validation

For T_max_mean, residual diagnostics indicated that ANOVA assumptions were adequately met. The Q-Q plot showed approximately normal residuals with minor tail deviations typical of large samples, and although the Shapiro-Wilk test was statistically significant ($W = 0.91$, $p < 2.2\times10^{-16}$), this reflected test sensitivity rather than substantive non-normality. Variance equality across ethnicity groups was supported by the Brown-Forsythe test ($F = 0.55$, $p = 0.74$). The corresponding diagnostics, provided in Appendix E.2, demonstrate that the normality and homoscedasticity assumptions were reasonably satisfied for T_max_mean.

Because both temperature measures met ANOVA's normality and variance assumptions, the one-way ANOVA models were considered appropriate for evaluating mean temperature differences across ethnicity groups. The subsequent analyses therefore proceed using standard ANOVA without the need for variance-robust or permutation-based alternatives.

6. Results

6.1 ANOVA Results

A one-way ANOVA indicated a statistically significant effect of ethnicity on inner-canthus temperature, $F(5, 882) = 6.73$, $p = 3.64\times10^{-6}$. Although the absolute temperature differences were modest, the effect size was nonzero ($\eta^2 = 0.04$, 95% CI: 0.02-1.00), meaning that roughly 4% of the variance in T_CEmax_mean was associated with the encoded categories. Because η^2 is a bounded parameter and the true variance explained is extremely small, the confidence interval expands markedly, a pattern noted in methodological discussions of fixed-effects ANOVA when effects approach the lower boundary [39].

The pattern of group means shows a subtle but consistent gradient. Asian and American Indian/Alaska Native participants were clustered at the lower end of the temperature distribution, whereas Hispanic/Latino, Multiracial, and White participants tended to exhibit slightly higher mean inner-canthus values. Black or African-American participants fell near the middle of the distribution. Importantly, the group confidence intervals overlap substantially, indicating that the categories do not demarcate sharp underlying boundaries and should not be interpreted as evidence of intrinsic biological differentiation, even though their aggregated means differ statistically. Complete statistical output and the corresponding visualization are provided in Appendix F.1.

A second ANOVA tested whether full-face maximum temperature (T_max_mean) differed across groups. This model also revealed a significant effect of ethnicity, $F(5, 882) = 4.15$, $p = 0.000993$, with a small effect size ($\eta^2 = 0.02$, 95% CI: 0.01-1.00). The ordering of groups means broadly mirrored that of the inner-canthus measure, with Hispanic/Latino and Multiracial groups exhibiting the highest observed values, American Indian/Alaska Native at the low end, and other groups distributed between these extremes. Again, the magnitude of the differences was small, and group intervals overlapped, but the directional consistency across two independent facial temperature metrics suggests that the encoded categories impose a reproducible structure on the dataset. Complete results and the associated plot appear in Appendix F.2.

Although between-group differences reached statistical significance, the underlying temperature distributions remained tightly bound. Across all six groups, the total spread in mean inner-canthus temperature was only about 0.30 °C, and full-face maximum temperatures varied by roughly 0.20 °C from lowest to highest. Within-group variability was similarly compact, with standard deviations typically well under 0.7 °C, indicating that each group was narrowly clustered around its mean.

7. Discussion

7.1 Clinical Lens

The statistically significant differences observed across encoded ethnic categories were small in magnitude and fall well below thresholds of clinical concern. The narrow confidence intervals, supported by the large sample size, show that although ANOVA detects structure in the data, the effect sizes ($\eta^2 = 0.02$-0.04) indicate that ethnicity, as defined by the dataset's administrative categories, accounts for very little variance in the measured temperature signal under controlled conditions. A single temperature reading carries low information density and often provides limited clinical insight without contextual or repeated measurements, since surface temperature does not directly reflect core thermoregulation [1, 3]. Clinically, thresholds such as “fever” function as decision rules that balance risk, uncertainty, and expected utility rather than as strict numeric boundaries [41]. Importantly, the observed between-group differences lie close to the uncertainty margins of non-contact infrared systems (approximately 0.3 °C), making differences of this scale difficult to distinguish reliably from ordinary measurement variability even under standardized imaging conditions [5]. In this context, the modest group differences observed here should not be interpreted as clinically meaningful.

Real-world practice further reduces any potential clinical significance. Holtzclaw notes that thermometers detect only the heat present at their interface and that readings vary substantially across sites because temperature has meaning only in relation to the region being measured and the mode of heat transfer involved [4]. In routine care, clinicians use different sites, techniques, and devices, and patients may not be positioned, acclimated, or prepared in consistent ways [4, 11]. Device performance also varies in practice, even in hospital settings, further amplifying uncertainty around small temperature differences. These sources of variability make it unlikely that small, statistically detectable differences (on the order observed in this dataset) would hold diagnostic value in real clinical environments.

7.2 Informatics Lens

From an informatics perspective, these findings show how data models can manufacture apparent structure by stabilizing both identity and uncertainty. Encoding ethnicity as a discrete field converts fluid, context-dependent identity into a durable computational object, so variation is organized and compared along boundaries the schema itself defines. As Cimino notes, controlled vocabularies create an appearance of conceptual clarity because they require unambiguous entries [40]. At the same time, clinical evaluations show that infrared and non-contact devices vary widely in accuracy and often fail to meet ISO and ASTM standards, with readings shaped by the measurement interface and subject to calibration drift [4, 5]. When such measurement uncertainty is ingested into informatics pipelines, it does not remain noise. Instead, aggregation, stratification, and reuse can stabilize small calibration or site-dependent fluctuations as consistent group differences. The resulting patterns reflect not inherent biological separation, but the joint imprint of representational schemas that discretize identity and numerical abstractions that render temperature comparable even as its uncertainty remains unresolved.

The directional pattern of group means, where Asian and American Indian or Alaska Native participants tended to fall slightly lower and Hispanic or Latino and Multiracial participants slightly higher, may reflect several non-physiological mechanisms made visible through the dataset’s representational structure. Prior controlled studies show that skin emissivity does not differ meaningfully across pigmentation levels [14, 15], yet report small apparent temperature differences under identical imaging conditions [15, 17], implicating device-specific signal processing, surface reflectance, and radiance-to-temperature conversion rather than intrinsic thermoregulation. A second possibility lies in the administrative categories themselves, which vary in size and internal heterogeneity and can influence the stability of aggregated means once encoded as fixed analytic fields. Because the dataset includes no direct measures of pigmentation, ancestry, or environmental context, these mechanisms cannot be evaluated independently. Their consistency across both thermographic measures therefore points to the joint effects of measurement and categorical encoding, illustrating how informatics ontologies can render modest, non-specific variation statistically visible without implying biological distinction.

7.3 Epidemiologic Lens

From an epidemiologic standpoint, the internal validity of the analysis is strong but tightly bounded. Measurement conditions were standardized, exposures and outcomes were assessed uniformly across groups, and age-related physiological variation was reduced through restriction. The large overall sample size limits random error and yields narrow confidence intervals, making it unlikely that the observed associations are due to chance alone [42]. However, precision is uneven across strata. Several ethnicity categories contain relatively few participants, which inflates uncertainty around their group means and limits the stability of between-group comparisons despite the large total sample.

External validity is more constrained and hinges on construct validity rather than sampling alone. The cohort consists almost entirely of healthy young adults, but the more consequential limitation lies in the ethnicity variable itself. The dataset provides no information on how ethnicity was assigned, whether by self-report, administrative classification, or observer judgment, nor does it specify the criteria governing group membership. As a result, the exposure under analysis is an unverified administrative label whose correspondence to sociocultural identity, ancestry, or phenotype cannot be assessed.

In epidemiologic terms, this uncertainty weakens interpretability more than the cohort's narrow age range. Any statistically detectable differences across groups therefore describe the behavior of the encoded variable within this dataset rather than a property of populations beyond it. The findings are best understood as internally valid patterns generated under controlled conditions, with limited generalizability and no warrant for physiological inference across ethnic groups. This distinction is critical: the analysis supports inference about data structure, not about biological difference.

7.4 Philosophical Lens

These findings sharpen a broader philosophical concern: when complex forms of human difference are compressed into fixed computational labels, the resulting numerical distinctions can easily be mistaken for natural ones. Small statistical differences across encoded ethnic groups demonstrate not biological separation but the ease with which classificatory structures acquire the appearance of biological meaning. As Fabrega noted, medical categories often reflect sociocultural convention rather than intrinsic boundaries [38]; the present results exemplify how such conventions can stabilize into seemingly objective differences through their incorporation into measurement and analytic systems.

The “Multiracial” category makes this dynamic especially visible. It aggregates heterogeneous ancestries, phenotypes, and lived identities into a single residual bin, and the statistical imprecision surrounding this group reflects that indeterminacy. The wide confidence interval alone reveals the looseness of the category itself. When a label has no coherent boundary, the variability attached to it becomes an artifact of its construction rather than evidence of distinct thermal physiology.

The patterns in this dataset show how classification systems designed for operational simplicity can shape the differences they appear to detect. The issue is not flawed measurement, but representational scaffolding that imposes categorical boundaries with limited correspondence to lived or physiological variation. The next section turns to possible approaches for redesign.

8. Proposed Solutions

8.1 Ontology Refinement

A first direction for improvement lies in refining the representational structures through which health-informatics systems encode biosocial identity. Current ethnicity fields are rigid, mutually exclusive, and administratively defined, collapsing multidimensional social experience into single categorical bins that statistical models then treat as biologically meaningful. Cimino’s desiderata emphasizes concept orientation, polyhierarchy, and nonsemantic identifiers as ways to preserve nuance and support multiple analytic purposes without sacrificing interoperability [40]. Applying this logic would mean replacing monolithic ethnicity fields with multi-scalar identity structures that allow individuals to be represented across dimensions such as ancestry (probabilistic and multi-valued), sociocultural affiliation (e.g., migration history, linguistic community), and clinically relevant contextual factors. This approach prevents administrative abstractions from gaining unwarranted ontological weight and offers a representational solution that aligns with both informatics design principles and the philosophical concerns raised in the analysis.

8.2 Personalized Baselines

A second solution reorients the problem entirely by shifting focus from between-group comparisons to within-person physiology. As research in thermoregulation and longitudinal temperature measurement demonstrates, individual “normal” temperatures vary systematically and cannot be reduced to a single population mean [11, 12]. Precision medicine has long recognized that a patient’s baseline is more informative for decision-making than comparisons to externally defined norms, and clinical reasoning frameworks grounded in Bayesian decision analysis likewise treat thresholds as context-dependent estimates influenced by patient values and expected utilities rather than fixed biological cutoffs [41, 43]. A personalized-baseline approach would extend this logic to thermal data by generating individualized reference ranges through repeated or longitudinal measurements, interpreting deviations relative to an individual’s stable pattern rather than the population’s average. Such an approach directly addresses the philosophical concern raised by Canguilhem that statistical norms are not physiological ideals, and it offers a practical way to avoid overinterpreting small group-level differences that may simply reflect demographic or administrative structure [23]. Whether such an individualized system is feasible at scale is a separate question, but the conceptual logic remains sound.

8.3 Multi-Scale Modeling

A third avenue for future work involves adopting multi-scale computational models to overcome the limitations of purely statistical generalization. As An argues, statistical inference becomes unreliable when the denominator – the space of possible variation – is poorly specified, a condition common in human biological and social data [44]. Administrative ethnicity categories compress heterogeneity in ways that obscure the mechanisms linking identity, physiology, behavior, and environment; ANOVA can test for differences between groups, but not for the origins or coherence of those groups. Multi-scale modeling offers a theory-based alternative by mapping processes across biological, behavioral, and social levels, enabling researchers to represent how heterogeneity arises rather than forcing it into static bins. Such models can integrate diverse data types, capture dependencies across scales, and provide a structured framework for generalization that does not depend on fragile population averages. In doing so, they directly address the epistemological tension at the heart of contemporary informatics: how to build systems that preserve meaningful variation rather than flatten it for computational convenience.

9. Limitations and Future Work

9.1 Data Limitations

Although the study benefits from unusually controlled imaging conditions, several limitations constrain the inferences that can be drawn. The most substantial concerns calibration transparency. The dataset provides no calibration schedule, reference standards, uncertainty margins, or drift checks for the IRT-2 system, despite Zhou's description of adherence to consensus guidelines [19]. Without this metadata, small between-group differences cannot be confidently interpreted as features of the underlying signal rather than residual instrument behavior. This limitation is nontrivial given that infrared thermometers are sensitive to distance, angle, and ambient conditions, and even well-maintained devices drift over time and require verification against reference standards that many instruments fail to meet in practice [4, 5, 9]. Because modern devices typically operate within tolerances of approximately ±0.3 to ±0.4 °C, the observed differences fall close to the limits of instrumental resolution, complicating interpretation.

The dataset partially mitigates instrumental concerns through acclimation protocols and the exclusion of subjects with inconsistent oral temperatures [20], but these controls do not address a separate source of uncertainty: residual physiologic microvariation. Even under standardized imaging conditions, surface temperature fluctuates with vasomotor tone, circadian phase, emotional arousal, and local perspiration at magnitudes comparable to the observed group differences [3, 4, 11, 13]. These influences are rarely measured directly and cannot be retrospectively modeled in secondary analyses. As a result, part of the observed between-group pattern may reflect ordinary microphysiological variability rather than stable categorical structure.

Precision is also uneven across categories. Although the overall sample is large, several ethnic groups contain markedly fewer participants than others, which reduces the stability of their mean estimates and increases the uncertainty surrounding small differences [42]. Additional interpretive limits arise from missing phenotypic variables: the dataset includes no measure of skin pigmentation or Fitzpatrick type, preventing independent evaluation of pigmentation-linked optical effects that could influence infrared readings [14, 15, 17]. The analytic restrictions further narrow external validity. Because nearly all participants are between 18 and 30 years of age, the dataset does not represent the thermoregulatory characteristics of older adults or clinical populations.

A final limitation concerns construct validity. The dataset provides no information about how ethnicity was defined or collected, whether through self-report, administrative assignment, or institutional records. Any detected differences therefore reflect the behavior of an administrative variable rather than a validated measure of sociocultural or ancestral identity [34].

9.2 Methodological and Theoretical Limitations

Methodologically, the analysis is limited by the structure of ANOVA itself which cannot determine if differences reflect measurement features, unmeasured factors, or the categorical design of the dataset [23, 39]. Even when assumptions are met, the fixed-effects formulation treats ethnicity as a discrete, meaningful variable, mirroring the ontology built into the data model. In this sense, the method tests the coherence of the schema rather than the biological plausibility of the categories it contains.

The theoretical framework also has limits. Testing whether administrative categories map onto a controlled physiological signal cannot resolve whether small differences arise from category design, historical residue, or the effects of standardization. As a result, while the analysis probes the adequacy of existing categories, it cannot specify what alternative representational forms would more faithfully capture biosocial identity.

9.3 Directions for Empirical and Conceptual Future Work

Future work must therefore proceed on multiple levels. Empirically, a more representative thermal dataset spanning age, geography, device variability, and clinical contexts would allow for hierarchical models capable of separating individual variation from categorical structure. Longitudinal designs, following the logic of Obermeyer [12], could estimate individual baseline temperatures and test whether group-level differences persist once intra-individual variation and environmental context are modeled explicitly. More granular measurement, incorporating additional physiological variables such as vasomotor reactivity, circadian phase, or hormonal status could clarify whether subtle differences arise from physiology or sampling.

From an informatics perspective, alternative representational structures are needed to move beyond fixed administrative categories. This may include ontologies that encode multiscalar identity (ancestry, migration history, lived experience), probabilistic rather than discrete membership, or narrative-linked metadata, as suggested by work in structured narrative and semantic representation [29, 30]. Such approaches could preserve the interoperability required for computation while reducing the risk of granting biological weight to categorical abstractions.

Finally, future philosophical work should evaluate how measurement practices – whether thermometric or informatic – shape what counts as biological knowledge. As the history of thermometry demonstrates, numerical conventions can become epistemic anchors long after their empirical foundations have shifted. A reflexive informatics must therefore confront not only how to measure difference, but how measurement itself produces the categories through which difference becomes legible.

10. Conclusion

This study examined whether the ethnic categories encoded in a tightly controlled thermographic dataset corresponded to measurable variation in facial temperature. The analyses revealed statistically reliable but clinically modest differences, accounting for only a small fraction of total variability. These effects do not reflect intrinsic physiological distinctions; rather, they show

how categorical structures built into a dataset can generate the appearance of patterned differences under conditions of high measurement precision. Once complex identities are encoded as fixed computational fields, they acquire a stability and interpretive weight that exceed their sociocultural origins.

The findings highlight a broader epistemic point: measurement is not a passive mirror of reality but an active constructor of it. Statistical differences emerge from the representational choices that make them possible, and in health informatics those choices determine which forms of human variation become legible, actionable, or pathologized. Just as 37 °C became ‘normal’ through measurement practices rather than universal physiology, so too may the demographic categories that shape modern analytics. Recognizing this symmetry is essential: otherwise today’s informatics risks repeating the epistemic blind spots of 19th-century thermometry. In this sense, the architecture of our data systems becomes the architecture of our truths, and designing more responsive and inclusive informatics requires not only better instruments but categories capable of capturing human diversity without reifying it.

11. Computational Reproducibility

To support transparency, reproducibility, and independent verification, all data-processing steps used in this analysis are fully documented in the project’s GitHub repository:

https://github.com/anais-lohier/bis560-aml276/tree/main

Because several preliminary cleaning steps were performed in Excel, the repository includes three sequential versions of the dataset: (1) the original raw file, (2) the Excel-processed version with filtered variables and averaged temperature fields, and (3) the R-processed analytic dataset used for the final models. This provides a clear audit trail from raw input to analytic output.

All statistical analysis and data wrangling were conducted in R (RStudio), with preliminary variable filtering and averaging performed in Microsoft Excel. The repository contains all R scripts used for data cleaning, missingness evaluation, and summary statistics. Since both the original dataset and the full analytic workflow are publicly available, the results can be independently replicated without relying on undocumented steps or proprietary data.

Artificial Intelligence (AI) Acknowledgement
AI-assisted tools (Clarity and ChatGPTv5.2) were used in a supportive capacity during the preparation of this work. Their use was restricted to generating draft R code for data manipulation and visualization based on detailed author specifications, and to assisting with

formatting, language refinement, and workflow efficiency. AI tools did not conduct the research, identify or formulate the research question, select data sources, design the study, determine analytic methodology, perform statistical reasoning, or interpret results. All analytical decisions, modeling choices, interpretations, and substantive conclusions are the intellectual work of the author. All AI-assisted outputs were critically reviewed, re-edited, and validated by the author to ensure accuracy, coherence, and scientific integrity.

References

1. Wunderlich, C. A., & Seguin, E. (1871). Medical Thermometry and human temperature. [Pt. 1] By C. A. W. (abridged by E. Seguin). The New York Printing Company. https://books.google.com/books?hl=en&lr=&id=a6UNq33GPfIC&oi=fnd&pg=PP13&ots=Ed0GT4zTIG&sig=a66RX5tKsKiErEz6UreTYvTn-ME#v=onepage&q&f=false

2. Taylor, S. (1942). The origin of the thermometer. Annals of Science, 5(2). https://doi.org/10.1080/00033794200201401

3. Chen, W. (2019). Thermometry and interpretation of body temperature. Biomedical Engineering Letters, 9(1), 3–17. https://doi.org/10.1007/s13534-019-00102-2

4. Holtzclaw, B. J. (1998). New Trends in Thermometry for the Patient in the ICU. Critical Care Nursing Quarterly, 21(3), 12–25. https://doi.org/10.1097/00002727-199821030-00003

5. Shinsuphan Nikorn, Aphinan Phanthi, Theera Leeudomwong, & Tassanai Sanponpute. (2024). Assessment of the accuracy and reliability of clinical thermometers for body temperature measurements at the hospital: A laboratory study. Measurement Science and Technology, 35(11), 115007–115007. https://doi.org/10.1088/1361-6501/ad64f7

6. Haslerus, J. (1578). De Logistica medica. Google Books. Retrieved from https://books.google.com/books?id=W5qgc-vz7fQC&printsec=frontcover&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false

7. Shryock, Richard H. "The history of quantification in medical science." Isis 52, no. 2 (1961): 215-237. https://www.jstor.org/stable/228680

8. Engel, George L. "How much longer must medicine's science be bound by a seventeenth century world view?." Psychotherapy and psychosomatics 57, no. 1-2 (1992): 3-16.

9. Beck, W. C., & Campbell, R. (1975). Clinical Thermometry. The Guthrie Journal, 44(4), 175–194. https://doi.org/10.3138/guthrie.44.4.175

10. Livingstone, D. (1858). Missionary Travels and Researches In South Africa. Abe Books. https://www.abebooks.com/book-search/title/livingstone%27s-travels-and-researches-in-south-africa/first-edition/

11. Sund-Levander, M., & Grodzinsky, E. (2009). Märtha Sund-Levander RNT PhD. International Journal of Nursing Practice, 15, 241–249. https://doi.org/10.1111/j.1440-172X.2009.01756.x

12. Obermeyer, Z., Samra, J. K., & Mullainathan, S. (2017). Individual differences in normal body temperature: longitudinal big data analysis of patient records. BMJ, j5468. https://doi.org/10.1136/bmj.j5468

13. Raiko, J., Koskensalo, K., & Sainio, T. (2020). Imaging-based internal body temperature measurements: The journal Temperature toolbox. Temperature, 1–26. https://doi.org/10.1080/23328940.2020.1769006

14. Charlton, M., Stanley, S. A., Whitman, Z., Wenn, V., Coats, T. J., Sims, M., & Thompson, J. P. (2020). The effect of constitutive pigmentation on the measured emissivity of human skin. PLoS ONE, 15(11), e0241843. https://doi.org/10.1371/journal.pone.0241843

15. auf der Strasse, W., Campos, D. P., Mendonça, C. J. A., Soni, J. F., Mendes, J., & Nohama, P. (2022). Forehead, Temple and Wrist Temperature Assessment of Ethnic Groups using Infrared Technology. Medical Engineering & Physics, 102, 103777. https://doi.org/10.1016/j.medengphy.2022.103777

16. Naik, P. P., & Farrukh, S. N. (2021). Influence of ethnicities and skin color variations in different populations- A Review. Skin Pharmacology and Physiology, 35(2). https://doi.org/10.1159/000518826

17. Sonenblum, S. E., Jordan, K., John, G. T., Chung, A., Asare-Baiden, M., Pelkmans, J., … Ho, J. C. (2025). Impact of Skin Tone, Environmental, and Technical Factors on Thermal Imaging. MedRxiv : The Preprint Server for Health Sciences, 2025.05.08.25327244. https://doi.org/10.1101/2025.05.08.25327244

18. Dell'Isola, G. B., Cosentini, E., Canale, L., Ficco, G., & Dell'Isola, M. (2021). Noncontact Body Temperature Measurement: Uncertainty Evaluation and Screening Decision Rule to Prevent the Spread of COVID-19. Sensors, 21(2), 346. https://doi.org/10.3390/s21020346

19. Zhou, Y., Ghassemi, P., Chen, M., Mcbride, D., Casamento, J., Pfefer, T., & Wang, Q. (2020). Clinical evaluation of fever-screening thermography: impact of consensus guidelines and facial measurement location. Journal of Biomedical Optics, 25(9), 97002–97003. https://doi.org/10.1117/1.JBO.25.9.097002.

20. Wang, Q., Zhou, Y., Ghassemi, P., Dwith Chenna, Chen, M., Casamento, J., Mcbride, D. (2023). Facial and oral temperature data from a large set of human subject volunteers. *Physionet.org*. https://physionet.org/content/face-oral-temp-data/1.0.0/

21. Gleser M, Young G, Woods D. A database built upon the Medical Event Vector. Methods Inf Med. 1979 Jul;18(3):131-7. PMID: 522663.

22. Jones RG, Johnson OA, Batstone G. Informatics and the clinical laboratory. Clin Biochem Rev. 2014 Aug;35(3):177-92. PMID: 25336763; PMCID: PMC4204239.

23. Canguilhem, George. (2012). On the Normal and the Pathological. Springer Science & Business Media. https://monoskop.org/images/b/b6/Canguilhem_Georges_The_Normal_and_the_Pathologic_1991.pdf

24. Liu, Q., Li, M., Wang, W., Jin, S., Piao, H., Jiang, Y., … Yao, H. (2025). Infrared thermography in clinical practice: a literature review. European Journal of Medical Research, 30(1). https://doi.org/10.1186/s40001-025-02278-z

25. Lazebnik, Yuri. "Can a biologist fix a radio?—Or, what I learned while studying apoptosis." Cancer cell 2, no. 3 (2002): 179-182.

26. Pearce, N., Foliaki, S., Sporle, A., & Cunningham, C. (2004). Genetics, race, ethnicity, and health. BMJ, 328(7447), 1070–1072. https://doi.org/10.1136/bmj.328.7447.1070

27. Sankar, P., Cho, M., & Mountain, J. (2007). Race and Ethnicity in Genetic Research. National Institute of Health (NIH).

28. Ali-Khan SE, Krakowski T, Tahir R, Daar AS. The use of race, ethnicity and ancestry in human genetic research. The HUGO Journal. 2011 Jul 7;5(1-4):47–63.

29. Johnson, Stephen B., Suzanne Bakken, Daniel Dine, Sookyung Hyun, Eneida Mendonça, Frances Morrison, Tiffani Bright, Tielman Van Vleck, Jesse Wrenn, and Peter Stetson. "An electronic health record based on structured narrative." Journal of the American Medical Informatics Association 15, no. 1 (2008): 54-64.

30. Lemieux, M., Bordage, G.: Propositional versus structural semantic analyses of medical diagnostic thinking. Cognitive Science 16(2), 185–204 (1992)

31. Engel, George L. "The need for a new medical model: a challenge for biomedicine." Science 196, no. 4286 (1977): 129-136.

32. Morreau, Michael, and Aidan Lyon. “How common standards can diminish collective intelligence: A computational study.” Journal of Evaluation in Clinical Practice, vol. 22, no. 4, 23 June 2016, pp. 483–489, https://doi.org/10.1111/jep.12585.

33. Birney E, Inouye M, Raff J, Rutherford A, Scally A. The language of race, ethnicity, and ancestry in human genetic research. 2021. https://doi.org/10.48550/arXiv.2106.10041

34. Bryc K, Durand Eric Y, Macpherson J Michael, Reich D, Mountain Joanna L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. The American Journal of Human Genetics. 2015 Jan;96(1):37–53. https://doi.org/10.1016/j.ajhg.2014.11.010

35. Sun, M., & Lipsitz, S. R. (2018). Comparative effectiveness research methodology using secondary data: A starting user’s guide. Urologic Oncology: Seminars and Original Investigations, 36(4), 174–182. https://doi.org/10.1016/j.urolonc.2017.10.011

36. Sagan, V., Maimaitijiang, M., Sidike, P., Eblimit, K., Peterson, K., Hartling, S., … Mockler, T. (2019). UAV-Based High Resolution Thermal Imaging for Vegetation Monitoring, and Plant Phenotyping Using ICI 8640 P, FLIR Vue Pro R 640, and thermoMap Cameras. Remote Sensing, 11(3), 330. https://doi.org/10.3390/rs11030330

37. Waalen, J., & Buxbaum, J. N. (2011). Is Older Colder or Colder Older? The Association of Age With Body Temperature in 18,630 Individuals. The Journals of Gerontology Series A: Biological Sciences and Medical Sciences, 66A(5), 487–492. https://doi.org/10.1093/gerona/glr001

38. Fabrega Jr, Horacio. "The Need for an Ethnomedical Science: The study of medical systems comparatively has important implications for the social and biological sciences." Science 189, no. 4207 (1975): 969-975.

39. Ståhle, L., & Wold, S. (1989). Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems, 6(4), 259–272. https://doi.org/10.1016/0169-7439(89)80095-4

40. Cimino, James J. "Desiderata for controlled medical vocabularies in the twenty-first century." Methods of information in medicine 37, no. 04/05 (1998): 394-403.

41. Pauker, Stephen G., and Jerome P. Kassirer. "The threshold approach to clinical decision making." The New England journal of medicine 302, no. 20 (1980): 1109-1117.

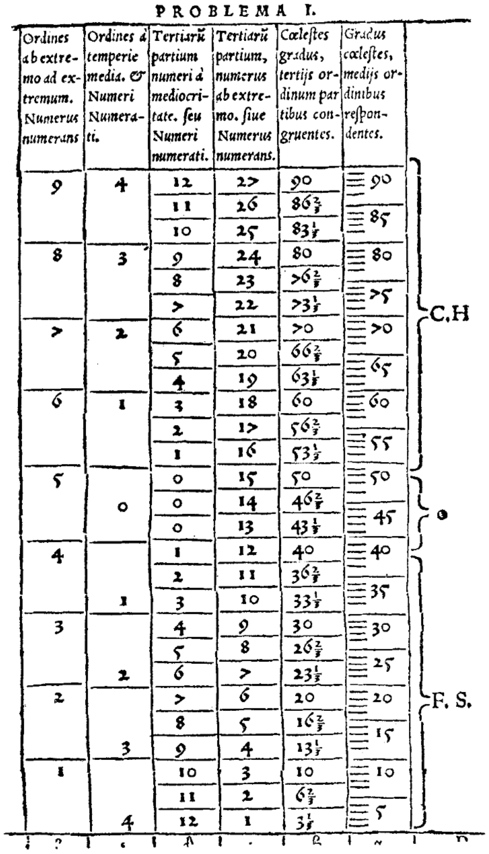
42. Ferrucci, L. “Lecture 18 - PUBH 508 Causation”. Yale University, Foundations of Epidemiology and Public Health

43. Feinstein, Alvan R. "Clinical judgment revisited: the distraction of quantitative models." Annals of Internal Medicine 120, no. 9 (1994): 799-805.

44. An G. The Crisis of Reproducibility, the Denominator Problem and the Scientific Role of Multi-scale Modeling. Bulletin of Mathematical Biology. 2018 Sep 7;80(12):3071–80.

Appendix A

Figure A.1 Early Conceptualizations of Temperature
Haslerus's table represents one of the earliest known numerical temperature scales. It arranges nine degrees of heat and cold symmetrically around a neutral midpoint (marked 0), corresponding to the "equal temperature" of the human body in Galenic theory. The rightmost column shows calculated celestial and terrestrial values, while the central columns record the numerical progression from the extremes (fourth degree of heat at the equator to fourth degree of cold at the pole). Each step corresponds to one-third intervals, linking climate, geography, and physiology.



PROBLEMA I.

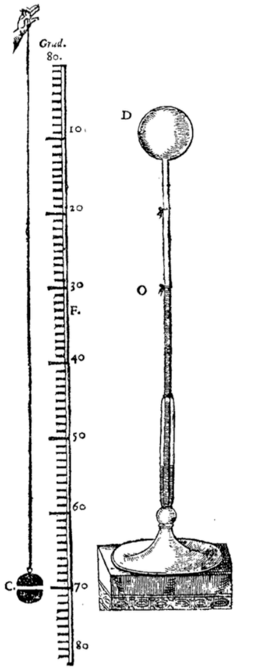
Ordines ab extremo ad extremum. Numerus numerans	Ordines à temperie media. & Numeri Numerati.	Tertiarũ partium numeri à mediocritate. ſeu Numeri numerati.	Tertiarũ partium, numerus ab extremo. ſiue Numerus numerans.	Cœleſtes gradus, tertijs ordinum partibus congruentes.	Gradus cœleſtes, medijs ordinibus reſpondentes.	
9	4	12	27	90	90	
		11	26	$86\frac{2}{3}$		
		10	25	$83\frac{1}{3}$	85	
8	3	9	24	80	80	
		8	23	$76\frac{2}{3}$		
		7	22	$73\frac{1}{3}$	75	C.H
7	2	6	21	70	70	
		5	20	$66\frac{2}{3}$		
		4	19	$63\frac{1}{3}$	65	
6	1	3	18	60	60	
		2	17	$56\frac{2}{3}$		
		1	16	$53\frac{1}{3}$	55	
5		0	15	50	50	
	0	0	14	$46\frac{2}{3}$		○
		0	13	$43\frac{1}{3}$	45	
4		1	12	40	40	
		2	11	$36\frac{2}{3}$		
	1	3	10	$33\frac{1}{3}$	35	
3		4	9	30	30	
		5	8	$26\frac{2}{3}$		
	2	6	7	$23\frac{1}{3}$	25	
2		7	6	20	20	F. S.
		8	5	$16\frac{2}{3}$		
	3	9	4	$13\frac{1}{3}$	15	
1		10	3	10	10	
		11	2	$6\frac{2}{3}$		
	4	12	1	$3\frac{1}{3}$	5	

Fig. 1.—Temperature scales (Haslerus, *De Logistica Medica*, 1578, p. 2).

Source: Joannes Haslerus, De Logistica Medica [6], reproduced in F. Sherwood Taylor [2].

Figure A.2 Early Tools for Experimental Thermometry

Santorio Santorio’s thermoscopium represents one of the earliest known applications of temperature measurement in clinical observation. Built around 1612-1626, his device consisted of a glass bulb and a long, narrow stem partially submerged in water or alcohol. As air inside the bulb expanded or contracted with temperature changes, the liquid in the stem rose or fell, allowing visible comparison rather than calibrated measurement. The accompanying pulsilogium (pendulum) shown on the left of the illustration was used to time the patient’s pulse and respiration, demonstrating Santorio Santorio’s effort to unify bodily observation with quantifiable rhythm. These instruments reflected a new epistemic ideal in seventeenth-century medicine, being that health and disease could be rendered measurable through instruments rather than perception alone.

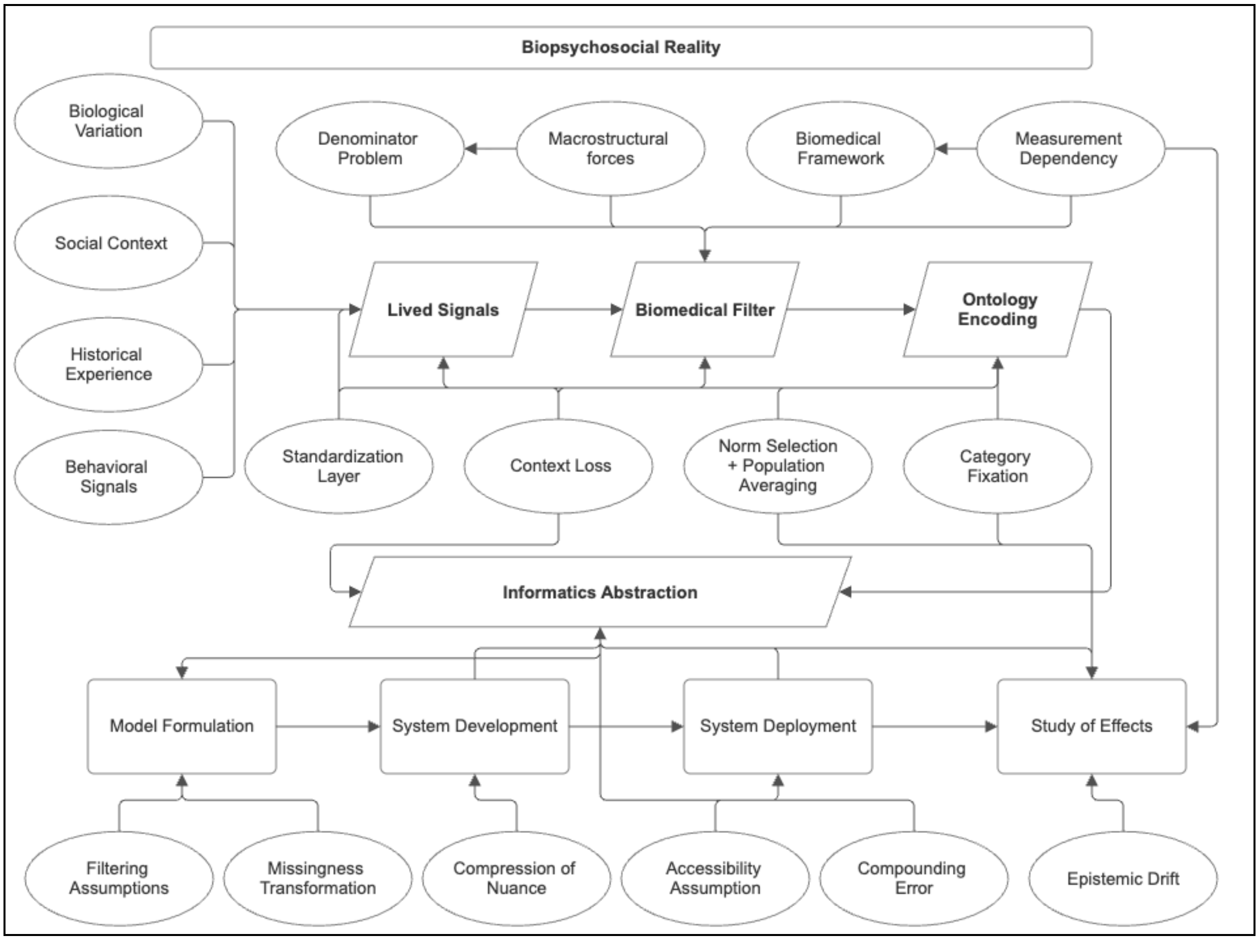


Fɪɢ. 3.—The thermometer of Sanctorius (1626, *op. cit.*, col. 22), with pulsilogium on left.

Joannes Haslerus, De Logistica Medica [6] as reproduced in F. Sherwood Taylor [2].

Appendix B

Figure B.1 Transformation of Biopsychosocial Reality into Computable Medical Data
This diagram illustrates how complex biopsychosocial reality is progressively transformed into standardized, computable data within health-informatics systems. Each step – from lived experience, to biomedical framing, to data abstraction – filters out nuance and embeds institutional assumptions into the resulting categories. The recursive structure of the diagram highlights that once categories are encoded, they continue to shape downstream interpretation, system design, and the patterns analytics can detect. This visual foregrounds the central argument of the paper: that measurement and representation co-produce the forms of human difference that become visible in clinical data.



Appendix C

Figure C.1 Facial Regions of Interest

Standardized facial regions of interest (ROIs) used in the thermographic study by Zhou et al. Temperatures were extracted from delineated zones – including the inner canthi, forehead, oral region, and whole-face maximum – to evaluate which sites most reliably track core temperature. This image demonstrates how anatomical standardization reduces measurement variability and why the two regions used in the present analysis (T_max and T_CEmax) are methodologically preferable.

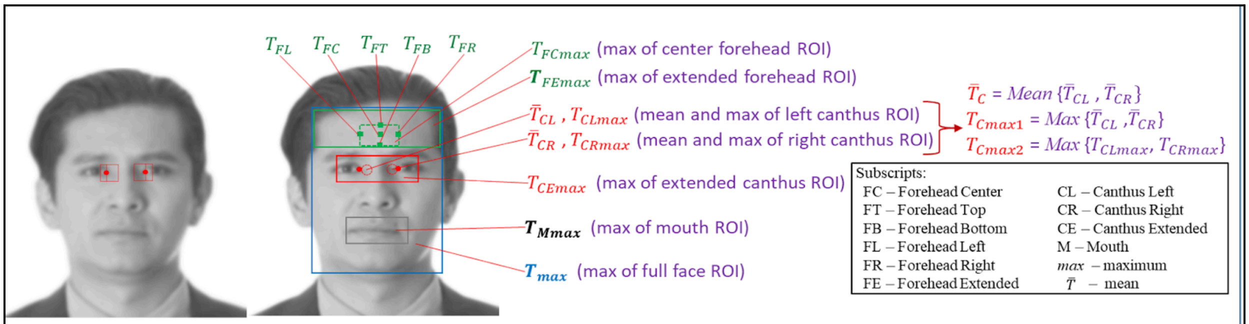

T_{FL}
T_{FC}
T_{FT}
T_{FB}
T_{FR}
T_{FCmax} (max of center forehead ROI)
T_{FEmax} (max of extended forehead ROI)
$\bar{T}_{CL}$, T_{CLmax} (mean and max of left canthus ROI)
$\bar{T}_{CR}$, T_{CRmax} (mean and max of right canthus ROI)
$\bar{T}_C$ = $Mean \{\bar{T}_{CL} , \bar{T}_{CR}\}$
T_{Cmax1} = $Max \{\bar{T}_{CL} , \bar{T}_{CR}\}$
T_{Cmax2} = $Max \{T_{CLmax}, T_{CRmax}\}$
T_{CEmax} (max of extended canthus ROI)
T_{Mmax} (max of mouth ROI)
T_{max} (max of full face ROI)
Subscripts:
FC – Forehead Center
FT – Forehead Top
FB – Forehead Bottom
FL – Forehead Left
FR – Forehead Right
FE – Forehead Extended
CL – Canthus Left
CR – Canthus Right
CE – Canthus Extended
M – Mouth
max – maximum
$\bar{T}$ – mean

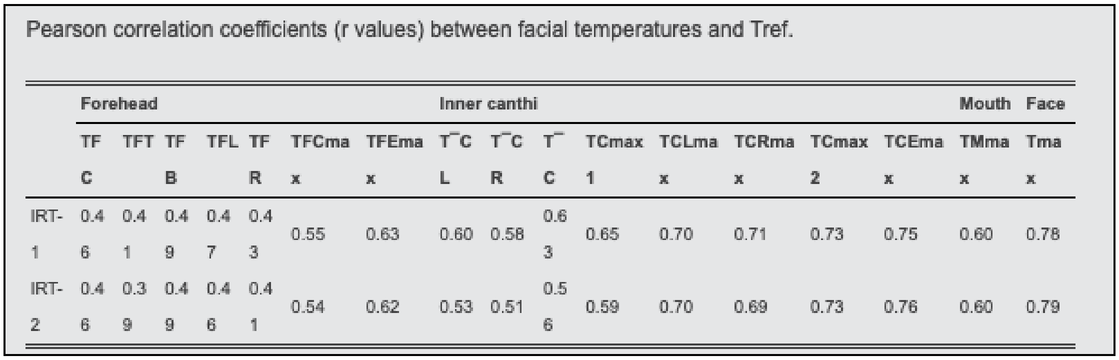
Figure 1. Delineated facial regions and critical points on thermal images: forehead regions and points (green), canthi region and points (red), mouth region (gray rectangle), and entire face (blue rectangle).

Note: The above image is a generic face (based on PowerPoint clip art: Insert > Icons > Cutout People >Alfredo) used for illustration purposes and not an actual participant in our study.

Data reproduced from Zhou [19].

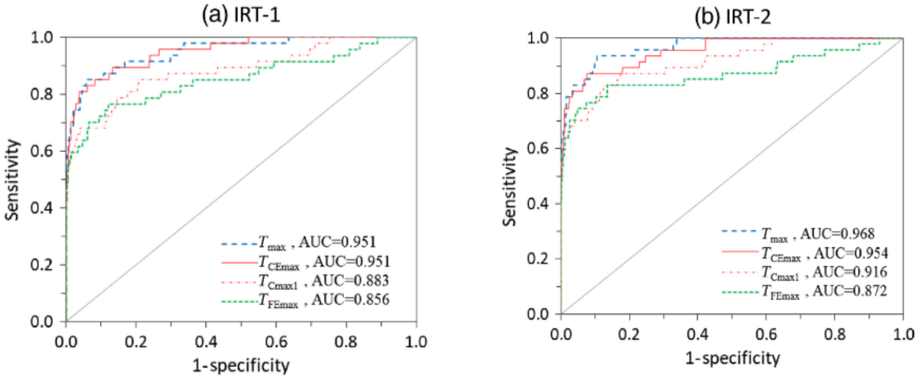
Figure C.2 Device Performance Comparison (IRT-1 vs. IRT-2)

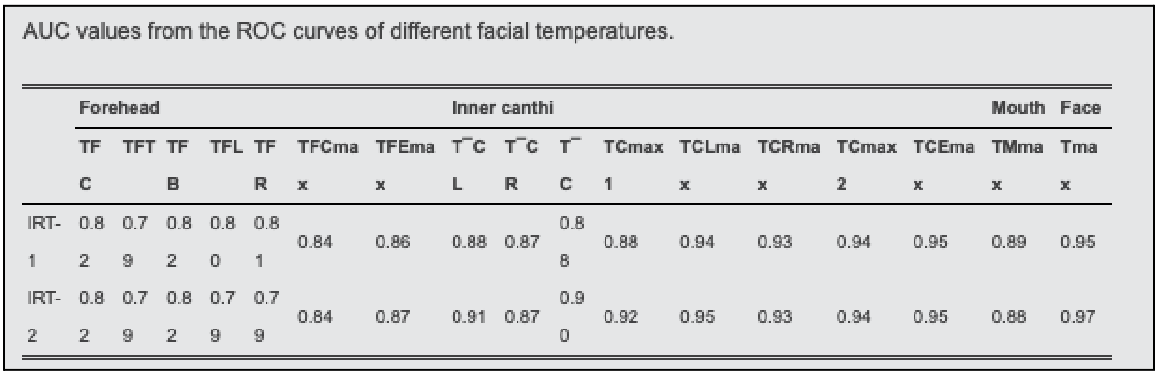
Comparative performance metrics for the two infrared thermography devices (IRT-1 and IRT-2) used in the original study. IRT-2 shows stronger correlations with core temperature and higher AUC values for both T_max and T_CEmax, especially under controlled conditions. This comparison motivates the analytic decision to restrict the present study to IRT-2 measurements, ensuring that observed differences are less influenced by instrument noise and more reflective of the dataset’s categorical structure.



Pearson correlation coefficients (r values) between facial temperatures and Tref.

	Forehead							Inner canthi								Mouth	Face
	TFC	TFT	TFB	TFL	TFR	TFCmax	TFEmax	T¯CL	T¯CR	T¯C	TCmax1	TCLmax	TCRmax	TCmax2	TCEmax	TMmax	Tmax
IRT-1	0.46	0.41	0.49	0.47	0.43	0.55	0.63	0.60	0.58	0.63	0.65	0.70	0.71	0.73	0.75	0.60	0.78
IRT-2	0.46	0.39	0.49	0.46	0.41	0.54	0.62	0.53	0.51	0.56	0.59	0.70	0.69	0.73	0.76	0.60	0.79





AUC values from the ROC curves of different facial temperatures.

	Forehead							Inner canthi								Mouth	Face
	TFC	TFT	TFB	TFL	TFR	TFCmax	TFEmax	T¯CL	T¯CR	T¯C	TCmax1	TCLmax	TCRmax	TCmax2	TCEmax	TMmax	Tmax
IRT-1	0.82	0.79	0.82	0.80	0.81	0.84	0.86	0.88	0.87	0.88	0.88	0.94	0.93	0.94	0.95	0.89	0.95
IRT-2	0.82	0.79	0.82	0.79	0.79	0.84	0.87	0.91	0.87	0.90	0.92	0.95	0.93	0.94	0.95	0.88	0.97

Data reproduced from Zhou [19].

Table C.3 Demographics of Study Subjects

This table presents the demographic composition of the original Zhou et al. dataset, including age range, gender distribution, and ethnic categories for both IRT systems. These distributions contextualize the analytic restrictions despite strong procedural control.

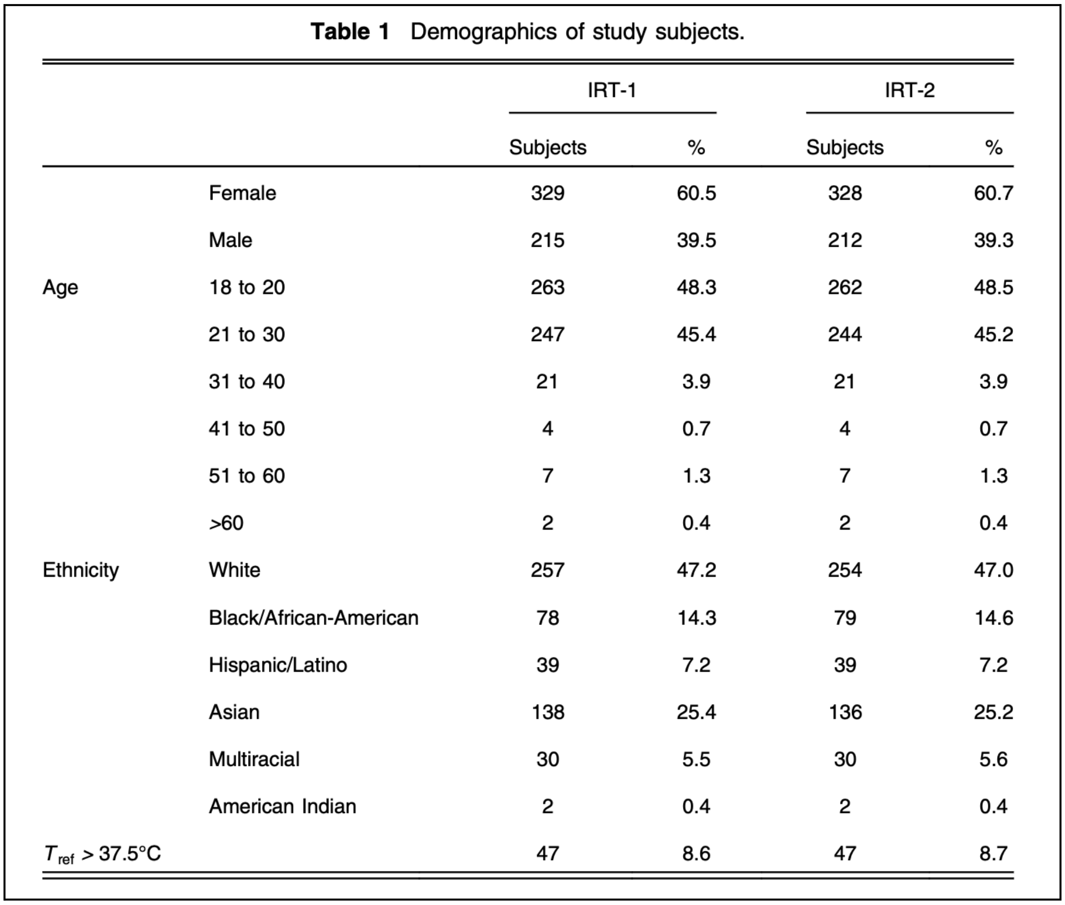


Table 1 Demographics of study subjects.

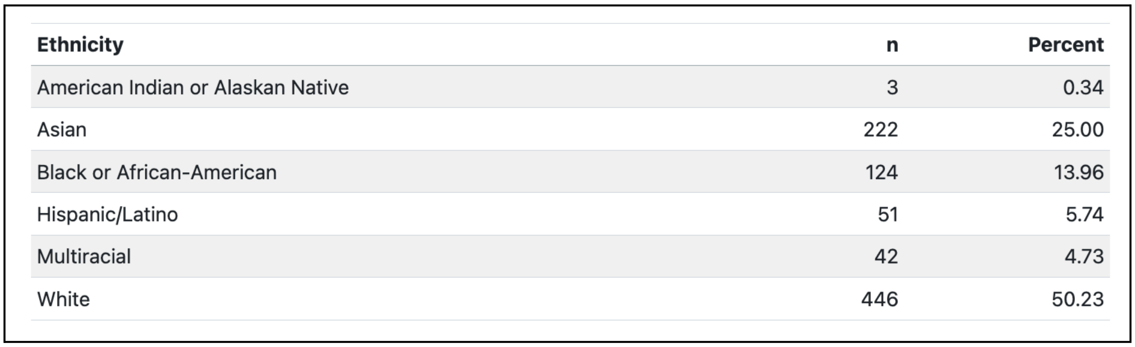
		IRT-1		IRT-2	
		Subjects	%	Subjects	%
	Female	329	60.5	328	60.7
	Male	215	39.5	212	39.3
Age	18 to 20	263	48.3	262	48.5
	21 to 30	247	45.4	244	45.2
	31 to 40	21	3.9	21	3.9
	41 to 50	4	0.7	4	0.7
	51 to 60	7	1.3	7	1.3
	>60	2	0.4	2	0.4
Ethnicity	White	257	47.2	254	47.0
	Black/African-American	78	14.3	79	14.6
	Hispanic/Latino	39	7.2	39	7.2
	Asian	138	25.4	136	25.2
	Multiracial	30	5.5	30	5.6
	American Indian	2	0.4	2	0.4
$T_{\text{ref}} > 37.5$°C		47	8.6	47	8.7

Data reproduced from Zhou [19].

Appendix D

Table D.1 Ethnicity Distribution of Subset

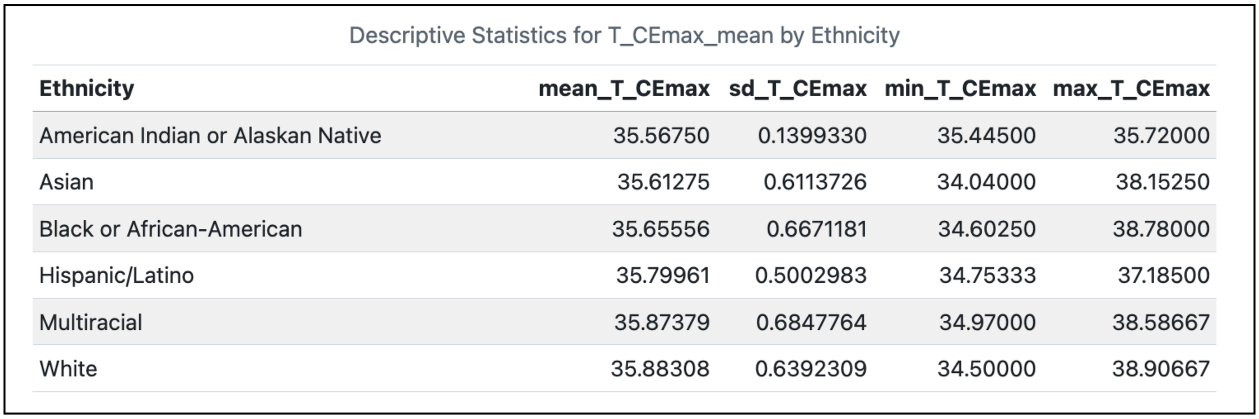
The following table summarizes the ethnic composition of the analytic subset. Counts and percentages are presented to document the distribution of participants retained after age filtering and variable selection.



Ethnicity	n	Percent
American Indian or Alaskan Native	3	0.34
Asian	222	25.00
Black or African-American	124	13.96
Hispanic/Latino	51	5.74
Multiracial	42	4.73
White	446	50.23

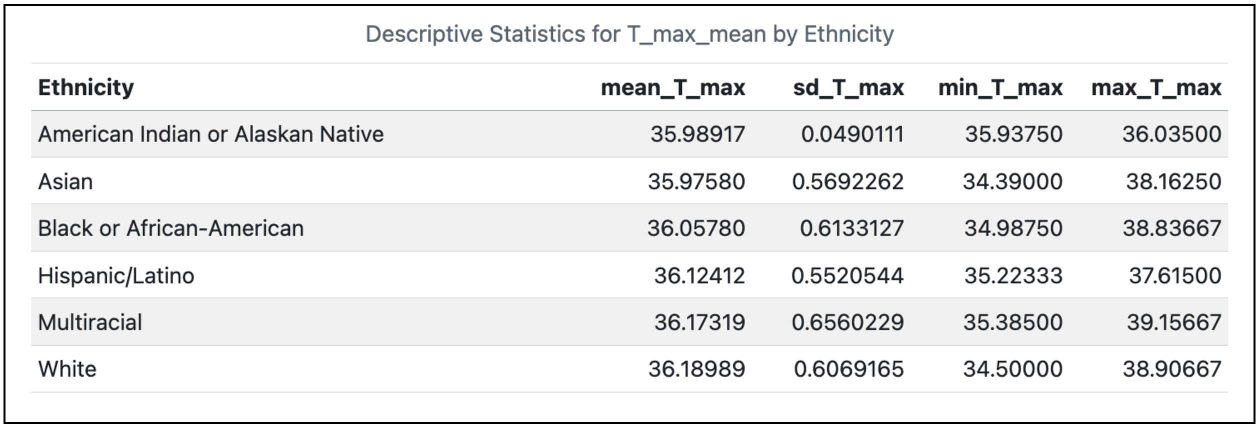
Table D.2 Descriptive Statistics

The following provides summary statistics for both temperature measures across ethnicity groups. Mean values and standard deviations are shown to document the central tendency and variability of the analytic variables used in the ANOVA models.



Descriptive Statistics for T_CEmax_mean by Ethnicity

Ethnicity	mean_T_CEmax	sd_T_CEmax	min_T_CEmax	max_T_CEmax
American Indian or Alaskan Native	35.56750	0.1399330	35.44500	35.72000
Asian	35.61275	0.6113726	34.04000	38.15250
Black or African-American	35.65556	0.6671181	34.60250	38.78000
Hispanic/Latino	35.79961	0.5002983	34.75333	37.18500
Multiracial	35.87379	0.6847764	34.97000	38.58667
White	35.88308	0.6392309	34.50000	38.90667

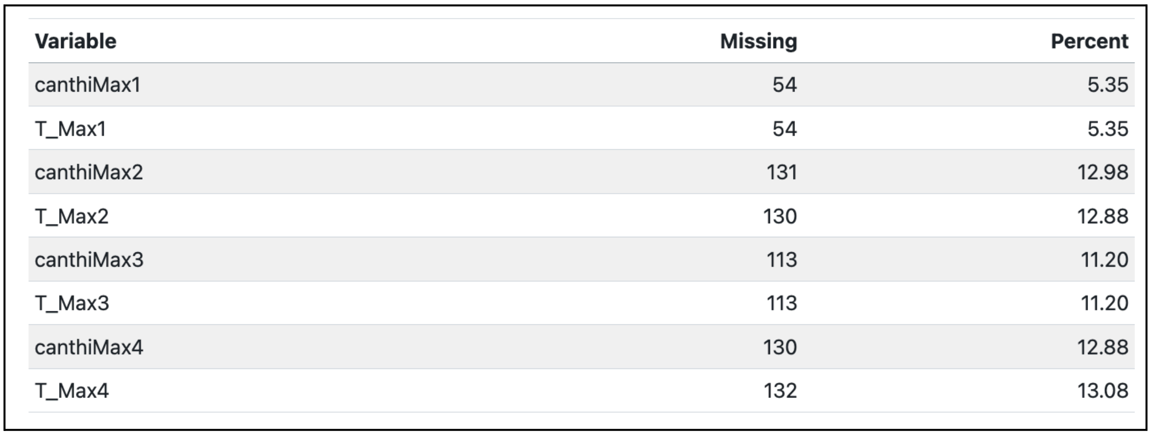


Descriptive Statistics for T_max_mean by Ethnicity

Ethnicity	mean_T_max	sd_T_max	min_T_max	max_T_max
American Indian or Alaskan Native	35.98917	0.0490111	35.93750	36.03500
Asian	35.97580	0.5692262	34.39000	38.16250
Black or African-American	36.05780	0.6133127	34.98750	38.83667
Hispanic/Latino	36.12412	0.5520544	35.22333	37.61500
Multiracial	36.17319	0.6560229	35.38500	39.15667
White	36.18989	0.6069165	34.50000	38.90667

Table D.3 Missingness across Rounds

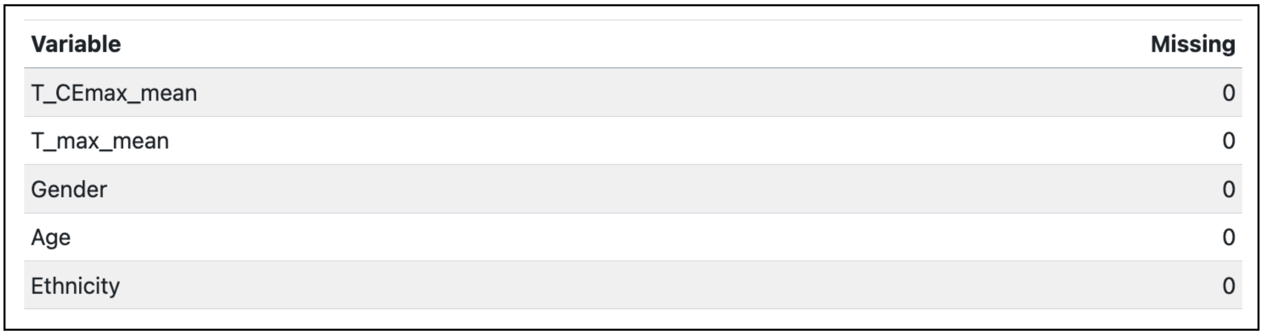
This table reports the number and percentage of missing observations for each of the four measurement rounds for both temperature variables. This table provides transparency about completeness prior to averaging round-level values.



Variable	Missing	Percent
canthiMax1	54	5.35
T_Max1	54	5.35
canthiMax2	131	12.98
T_Max2	130	12.88
canthiMax3	113	11.20
T_Max3	113	11.20
canthiMax4	130	12.88
T_Max4	132	13.08

Table D.4 Missingness after Averaging

The following shows missingness for the final analytic variables after round-level temperatures were averaged. All retained variables have complete data, confirming that the analytic subset used for ANOVA contains no missing observations.

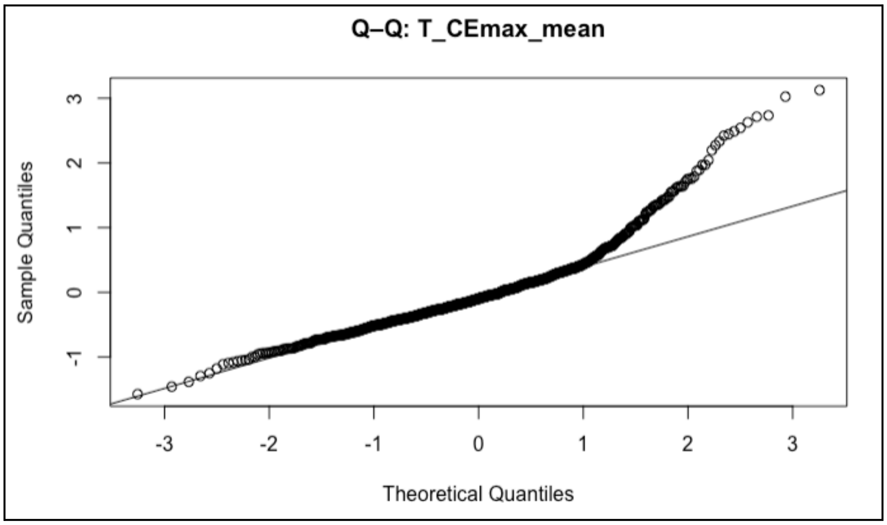


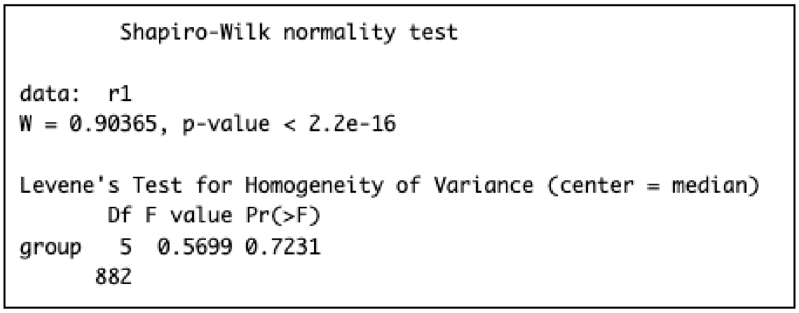
Variable	Missing
T_CEmax_mean	0
T_max_mean	0
Gender	0
Age	0
Ethnicity	0

Appendix E

Figure E.1 Statistical Assumptions for T_CEmax_mean

This figure summarizes the ANOVA diagnostic checks for T_CEmax_mean, including a Q-Q plot of residuals, the Shapiro-Wilk normality test, and the Brown–Forsythe test for homogeneity of variances. These outputs document whether the model meets the required assumptions.

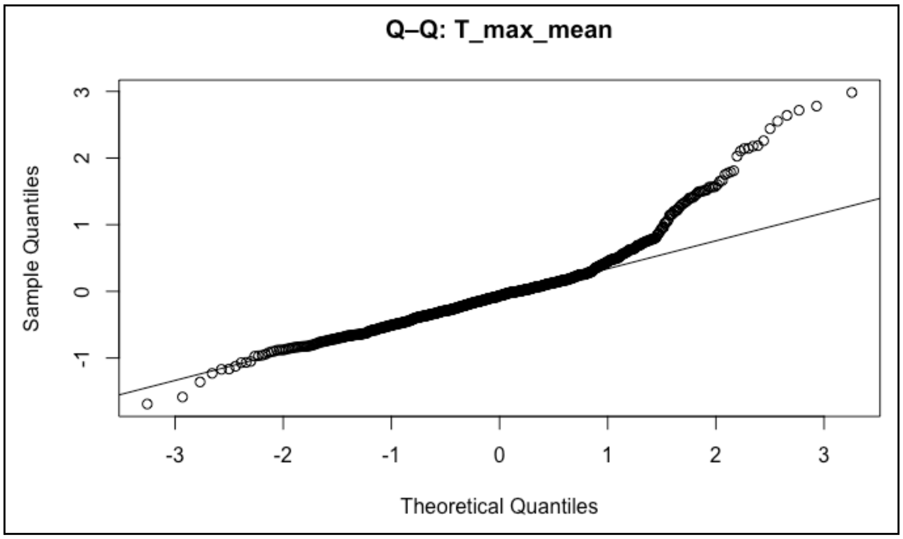


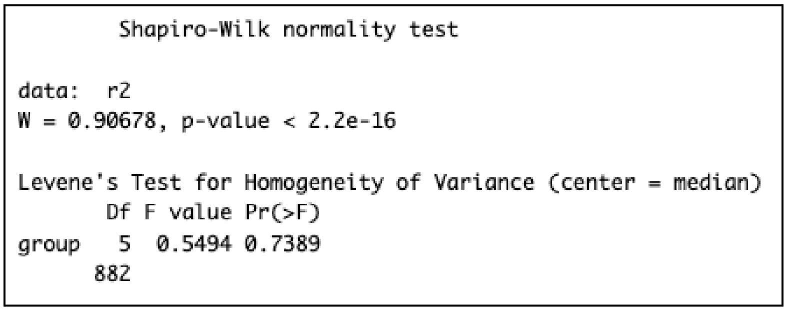


```
        Shapiro-Wilk normality test

data:  r1
W = 0.90365, p-value < 2.2e-16

Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  0.5699 0.7231
      882
```

Figure E.2 Statistical Assumptions for T_max_mean

This figure summarizes the ANOVA diagnostic checks for T_max_mean, including the residual Q-Q plot, Shapiro-Wilk normality test, and Brown-Forsythe test for homogeneity of variances. These outputs document whether the model satisfies the assumptions required for a one-way ANOVA.

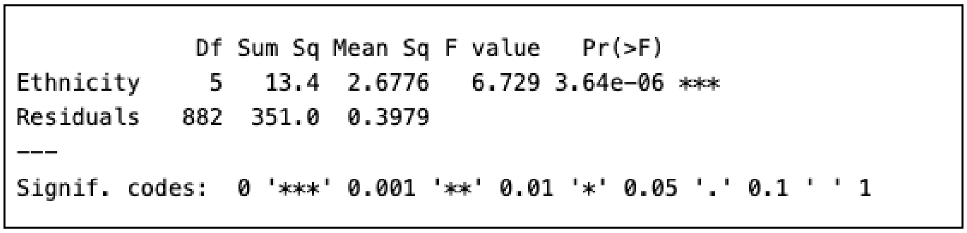




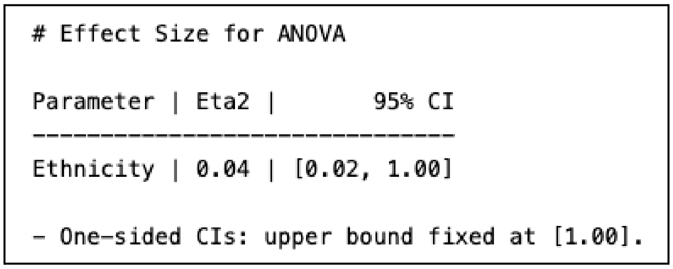
```
        Shapiro-Wilk normality test

data:  r2
W = 0.90678, p-value < 2.2e-16

Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   5  0.5494 0.7389
      882
```

Appendix F

Table F.1 ANOVA Analysis and Visualization for T_CEmax_mean
This section provides the complete ANOVA results for T_CEmax_mean, including the F-statistic, p-value, and η² effect size, followed by a visualization of group means with 95% confidence intervals. Together, the statistical output and plot document the underlying results and illustrate the magnitude and direction of between-group differences in inner-canthus temperature.



```
              Df Sum Sq Mean Sq F value   Pr(>F)    
Ethnicity      5   13.4  2.6776   6.729 3.64e-06 ***
Residuals    882  351.0  0.3979                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
# Effect Size for ANOVA

Parameter | Eta2 |       95% CI
-------------------------------
Ethnicity | 0.04 | [0.02, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```

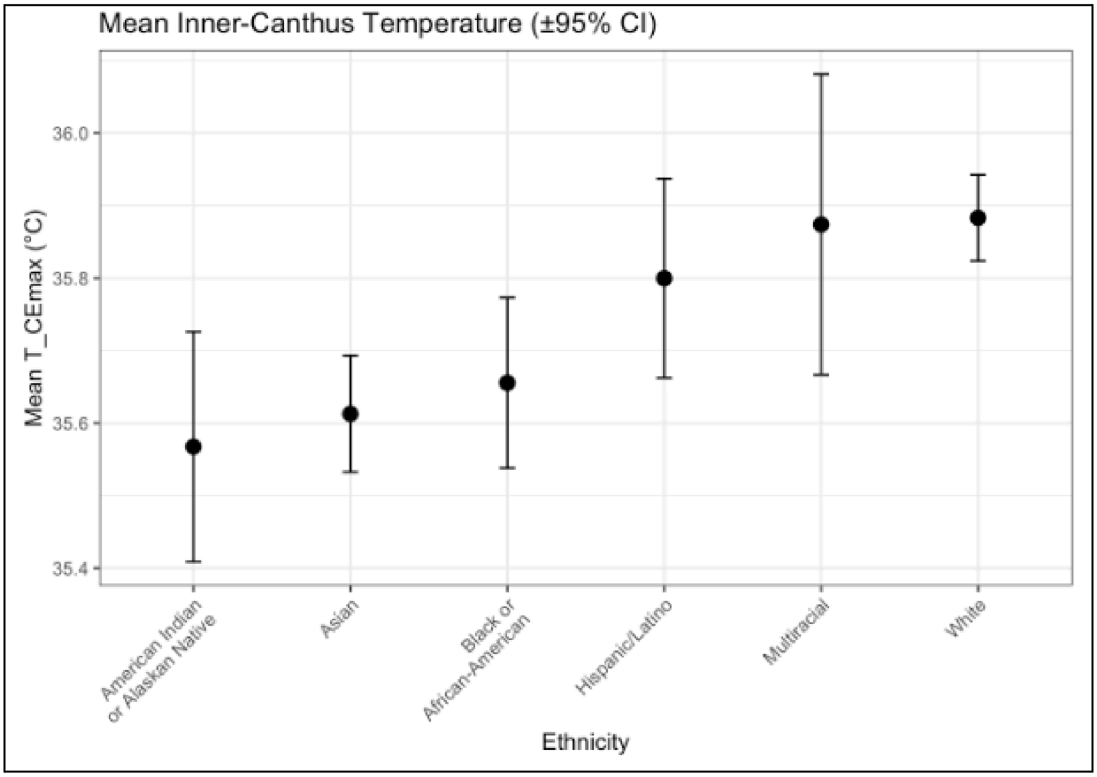
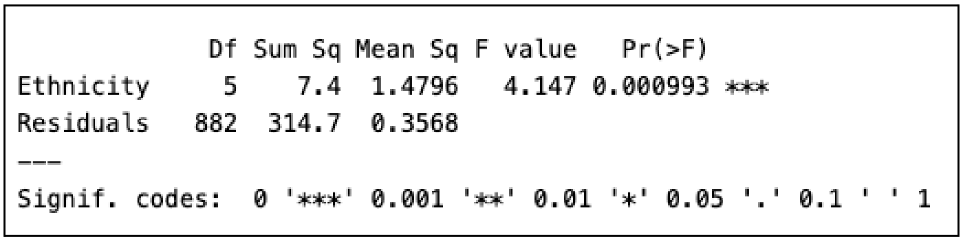
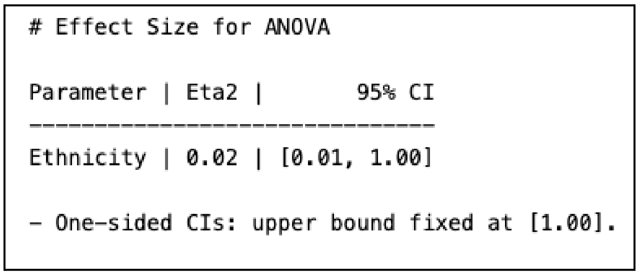


Table F.2 ANOVA Analysis and Visualization for T_max_mean

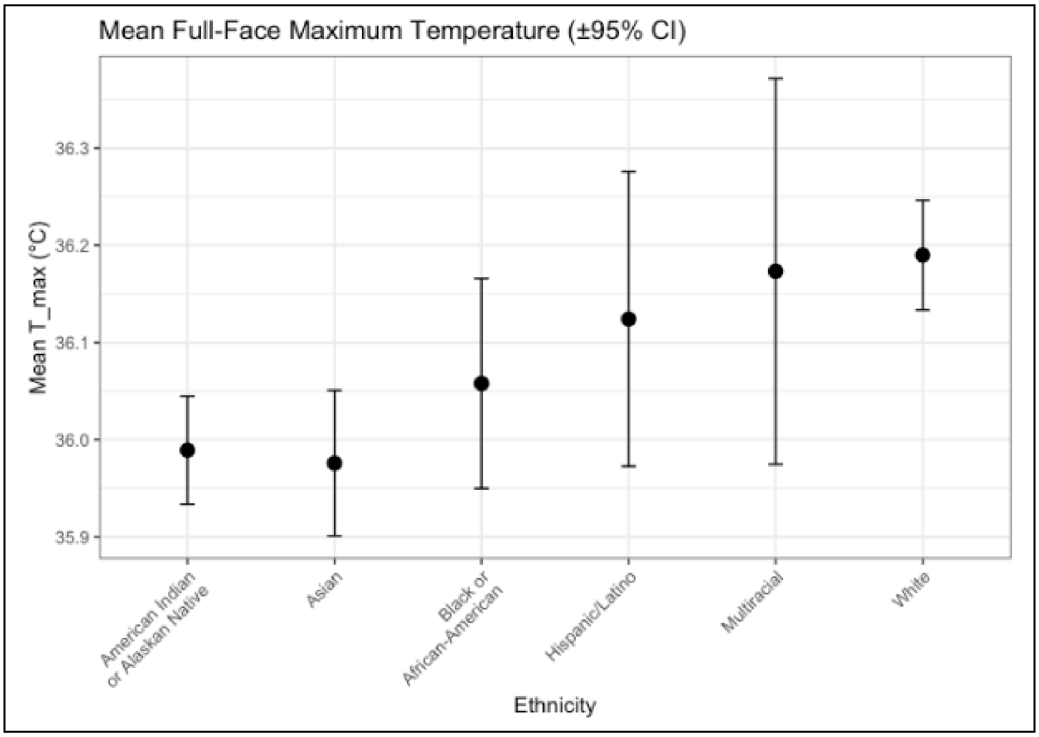
This section provides the complete ANOVA results for T_max_mean, including the F-statistic, p-value, and η² effect size, followed by a visualization of group means with 95% confidence intervals. Together, the statistical output and plot present the detailed results and illustrate the pattern and magnitude of between-group differences in full-face maximum temperature.



```
               Df Sum Sq Mean Sq F value   Pr(>F)    
Ethnicity       5    7.4  1.4796   4.147 0.000993 ***
Residuals     882  314.7  0.3568                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
# Effect Size for ANOVA

Parameter | Eta2 |       95% CI
-------------------------------
Ethnicity | 0.02 | [0.01, 1.00]

- One-sided CIs: upper bound fixed at [1.00].
```


Mean Full-Face Maximum Temperature (±95% CI)
Mean T_max (°C)
American Indian or Alaskan Native
Asian
Black or African-American
Hispanic/Latino
Multiracial
White
Ethnicity