Abstract

Detecting stress-related physiological changes, even before subjective awareness of these changes, allows for earlier intervention. Consumer wearables now make naturalistic stress research feasible at scale, yet many return coarse heart rate (HR) estimates in beats per minute rather than the millisecond-resolution interbeat intervals that established psychophysiological frameworks assume. In this feasibility study, we collected continuous Fitbit data over 12–23 weeks from 33 police academy trainees, paired with self-reports issued by a custom on-device application when heart rate exceeded an individualized threshold. We treated coarse HR/HRV data streams not as degraded signals but as a different measurement object, one suited to describing how cardiac activity is organized over time. We adopted a flexibility–structure framework that views cardiac dynamics as organized variability across timescales, operationalized through fourteen features in four theoretically motivated families. Using a two-stage analytic strategy that paired cross-validated prediction with explanatory modeling, we identified a signature we termed "constrained unpredictability" in the 90 minutes preceding stress reports: the operating band narrows, local unpredictability within it increases, and the series dwells longer in particular cardiac states. These findings suggest that cardiac dynamics can be informative in preceding stress experience among police trainees. We argue that the gap between consumer wearables and traditional psychophysiology is not a ceiling but a space for principled descriptive science, in which empirical questions and analytic methods are matched to the signal that is actually available.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Keywords: consumer wearables, heart rate dynamics, ambulatory stress monitoring, cardiac complexity, recurrence quantification analysis, police trainees

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

1 Introduction

Stress is ubiquitous in modern work, but in policing it assumes a unique intensity. Officers routinely make judgments under uncertainty—when to approach, when to use force, when to de-escalate—decisions that can have life-or-death consequences. In such moments, heightened physiological arousal may sharpen attention and reaction speed, yet excessive arousal can narrow perception, bias threat appraisal, and impair judgment, increasing the risk of escalation and error (Andersen, 2015; Hope, 2016; Nieuwenhuys & Oudejans, 2011). These acute reactions also occur against a backdrop of cumulative strain, including exposure to traumatic events and compounded by police officers having to operate under the watchful eye of both their departments and the public (Craddock & Telesco, 2022; Guy et al., 2014; Singletary, 2024). Over time, this combination of moment-to-moment stress reactivity and chronic load contributes to disproportionate rates of burnout, anxiety, depression, post-traumatic stress, substance use, and relationship difficulties among officers relative to other occupations (Andersen et al., 2023; Violanti, 2007).

As concerns over officer well-being and its implications for citizens grow, a central challenge is how to capture physiological markers of stress as they unfold in real-world policing environments (Furberg et al., 2017; Goodison et al., 2020). Indeed, this challenge extends to nearly all real-world occupational contexts. Laboratory paradigms have been essential for identifying the autonomic correlates of stress and related affective states, but their ability to generalize can be limited. Real-world stressors are often unpredictable in timing and magnitude, fluctuate across hours or days, and accumulate into chronic states that conventional laboratory tasks cannot fully reproduce (Epel et al., 2018; Ferreira et al., 2024; McEwen, 2017). Capturing

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

these processes requires methods that can follow stress-related physiology in situ, across the timescales at which it is experienced.

Recent advances in ambulatory monitoring and wearable sensing technologies have enabled researchers to move beyond constrained laboratory stressors and track physiology as people go about their daily lives (Martinez et al., 2022; Pinge et al., 2024). Wrist-worn devices have been particularly advantageous as they are affordable enough for deployment at scale, acceptable to users in demanding, high-stress occupations, and capable of continuous, passive tracking of cardiac activity and movement (Jetelina et al., 2022; Nelson et al., 2020). By embedding measurement into everyday routines, they introduce ecological validity that traditional lab-based approaches cannot easily match (Roos & Slavich, 2023; Sheikh et al., 2021). Current research suggests that, at minimum, wearables can reveal broad patterns of physiological change over time (Schwerdtfeger & Rominger, 2021; Radin et al., 2021) and how those changes may relate to subjective states such as perceived stress (Can et al., 2019). However, realizing this potential requires taking seriously the gap between what these devices provide or expose to researchers and what established psychophysiology assumes.

Many advances in wearable-based stress research rest on the capacity of wearables to capture cardiac activity through optical photoplethysmography (PPG), which estimates inter-beat intervals (IBIs)—the metric typically used in laboratory and research-grade field studies given its high temporal resolution—by detecting blood volume changes with each heartbeat. Under controlled conditions, PPG-derived IBIs can yield heart rate variability (HRV) metrics comparable to electrocardiogram gold standards (Hernando et al., 2018; Schäfer & Vagedes, 2013), enabling the detailed autonomic assessments that underlie much of the stress-monitoring literature (Laborde et al., 2017). While having this data is the ideal, it is also hard to come by for

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

researchers who aim to use off-the-shelf commercial wearables, such as Fitbit or Garmin devices. As is often the case, the raw PPG or millisecond-resolution IBI streams are not exposed to or shared with end consumers. What is made available are processed heart rate estimates, such as beats per minute (bpm), sampled at variable intervals (often 5-15 seconds) and shaped by proprietary algorithms that prioritize battery life and user experience (Fitbit, LLC, 2022; Kim & Baek, 2023; Nelson et al., 2020). As a result, the very devices that make large-scale, naturalistic monitoring feasible do not readily provide the data fidelity that established psychophysiological frameworks assume. Even research-grade wearables that do expose IBIs can yield streams too sparse for HRV in everyday wear, leaving researchers to work from global heart-rate features instead (Tutunji et al., 2023). Alternative analytical strategies and theoretical frameworks must therefore be considered (e.g., Fitabase, n.d.; Dell'Aversana et al., 2026; Kinoshita et al., 2024), yet more work is needed.

We take a novel approach to addressing this methodological and conceptual tension by adopting what we call a *flexibility–structure framework*, inspired from the physiological complexity literature. Here, cardiac dynamics can be construed as organized variability reflecting regulatory balance or regulatory flexibility (Goldberger et al., 2002), rather than as one-to-one indices of the specific autonomic mechanisms that canonical, IBI-based HRV measures are taken to reflect (Berntson et al., 1997). The core idea is that healthy biological systems exhibit structured variability across multiple timescales, a multiscale structure that is neither rigidly regular nor completely random but organized in ways that support adaptive responding. When regulation is compromised, variability may increase or decrease, but more importantly this structure can break down (Ivanov et al., 1999; Lipsitz & Goldberger, 1992). Our framework, by design, does not require millisecond-resolution IBIs; it requires sufficient temporal granularity to

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

assess how variability is patterned over time. We propose that sequences of bpm estimates sampled every few seconds are sufficient for such assessment, but ultimately will be determined by relating extracted features to stress events reported by participants. Such an approach is consistent with theoretical accounts that link heart rate variability to emotion/affect regulation (Beauchaine & Thayer, 2015; Porges, 2011), and extends it with more nuanced temporal patterning indicators.

1.1 Current Study

The present study uses this flexibility–structure perspective as an organizing principle for analyzing cardiac dynamics from consumer wearables in a real-world police training environment. We investigated this within a feasibility study conducted as the first step in a larger research program aiming to capture potentially problematic (e.g., sustained, inflexible, misallocated) stress (and ultimately behavioral) reactivity among police. We deployed wrist-worn devices with police academy trainees over 12–23 weeks of training, collecting continuous heart rate and activity level alongside in-situ stress reports. Rather than attempting to recover traditional HRV measures from the available bpm estimates, we constructed a feature set tailored to the temporal resolution and coverage provided by these devices. Specifically, we selected fourteen "flexibility–structure" features organized into four theoretically motivated families: Cardiac Baseline, Cardiac Complexity, Cardiac Dynamics, and Pattern Recurrence.

To link these features to stress experience, we combined continuous wearable monitoring with physiologically-triggered self-reports. A custom application running on the device continuously monitored heart rate and issued prompts when individualized thresholds of sustained elevation were exceeded. Participants then rated how stressed they felt at that moment

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

on a brief Likert scale. Features were then generated from heart rate across varying windows of time preceding the moment of each self-report (stressed vs. not stressed).

Analytically, we adopted a two-stage approach designed to balance predictive generalization with interpretability. Stage 1 asked whether the framework generalized to new individuals: we used a machine-learning approach in which models were trained on a subset of participants and evaluated on participants whose data the model had not seen (participant-level cross-validation), and used the results to identify which features and feature families contributed most consistently to prediction. Stage 2 asked how those features related to stress: we fit a single explanatory mixed-effects model using the most robust features to estimate effect sizes and directions for theoretical interpretation, while controlling for activity level and time of day. Across both stages, our focus was not on maximizing classification accuracy per se, but on characterizing how organized variability in coarse heart rate signals related to stress reports in a high-intensity occupational context.

In what follows, we briefly elaborate on the main theoretical and methodological contributions of this paper: (a) the novel feature families and the underlying theoretical framework that justifies their selection, and (b) the focus on anticipatory cardiac patterns based on "in-the-wild" physiologically-triggered stress assessments. We then describe our data collection, feature engineering, data quality assessment, and analytic strategy in detail, and finally report how patterns of organized variability differ across time windows preceding stressed versus non-stressed reports, and the possible implications.

1.2 The *Flexibility–Structure Balance* Framework

Work examining physiological complexity shows that adaptive biological systems exhibit structured variability across scales, whereas dysregulation may involve a loss or alteration of that

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

structure (Goldberger et al., 2002; Lipsitz & Goldberger, 1992). In this tradition, the key distinction is not simply between "high" and "low" variability, but between variability that remains organized and variability that becomes overly rigid, noisy, or poorly coordinated. We draw on this stance in approaching our analyses derived from consumer wearables, which provide coarse, processed heart rate estimates rather than researcher-controlled beat-to-beat intervals. Accordingly, we do not treat the present features as direct markers of sympathetic or parasympathetic activity, nor as substitutes for canonical HRV indices. Instead, we use them as descriptive measures of how cardiac activity is organized over time. Because prior literature was developed with finer-grained physiological signals, we treat our coarse-bpm feature set as an ecologically adapted extension of it. The conceptual claim we carry forward is that organized, multiscale variability indexes adaptive capacity.

A second body of work links cardiac dynamics to broader processes of cognitive and emotional self-regulation. Neurovisceral integration models propose that cardiac variability reflects, in part, the functioning of distributed systems supporting attention, affect regulation, inhibitory control, and adaptive responding (Smith et al., 2017; Thayer & Lane, 2000). Related reviews have argued that heart-rate variability is informative because it covaries with regulated emotional responding and other top-down self-regulatory capacities (Appelhans & Luecken, 2006). Again, although this work is largely formulated around beat-to-beat HRV and cardiac vagal control, its central premise opens up exploration of whether coarse cardiac time series carry information about how regulation is organized before a participant reports feeling stressed.

These perspectives also align with psychological work emphasizing regulatory flexibility. In such accounts of coping and emotion regulation, adaptation to stress is recast less as the deployment of any single optimal response and more as the capacity to adjust responses to

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

situational demands, draw on a diverse repertoire of strategies, and update behavior in light of feedback (Aldao, Sheppes, & Gross, 2015; Bonanno & Burton, 2013; Specker, Sheppes, & Nickerson, 2024). In parallel, allostatic accounts characterize regulation as anticipatory and predictive, with physiological systems reallocating resources in relation to expected demands rather than merely reacting after perturbations occur (Sterling, 2012). Consistent with this account, the period preceding a stress report may contain physiological signs of anticipatory resource allocation.

Together, these ideas motivate our current extraction of feature families as a pragmatic measurement framework with theoretical implications, rather than as a mechanistic theory under direct test. Features indexing central tendency and range capture the operating band within which cardiac activity is fluctuating (*Cardiac Baseline*). Features indexing entropy and scaling properties capture local unpredictability and longer-range temporal organization (*Cardiac Complexity*). Features indexing successive changes characterize the frequency and asymmetry of moment-to-moment adjustments (*Cardiac Dynamics*). Recurrence-based features capture the extent to which the series revisits, dwells in, or rapidly diverges from prior states (*Pattern Recurrence*). Similar efforts to use nonlinear or ecologically adapted cardiac features for real-world stress detection have appeared in prior work, including studies of real-life stress detection using nonlinear HRV features and methods for identifying psychologically meaningful episodes in everyday life (Brown et al., 2018; Melillo et al., 2011). Our contribution is to test whether these features provide a principled and interpretable description of organized variability in consumer-wearable heart-rate data leading up to perceived stress.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

1.3 Physiologically-triggered stress assessment and anticipatory cardiac patterns

A persistent challenge in ambulatory stress research is that the correspondence between physiological activity and subjective stress is highly variable, both between and within individuals. An elevated heart rate or a demanding event does not necessarily translate into an experience of stress, and conversely, a person may report feeling stressed without showing the physiological profile a researcher would expect. In real-world settings where stressful events cannot be induced or controlled, one of the most direct ways to identify when stress occurs is to simply ask via ecological momentary assessments (EMAs). What constitutes *perceived* stress is often left to the individual's intuitive understanding. More formally, it can be operationalized as the experience of demands (physiological, psychological, and contextual) that exceed available coping resources (Folkman, 2013).

EMAs, especially when paired with ambulatory physiological sensing, have proven informative for capturing stress as it is perceived in daily life, and have been deployed across a range of occupational and ambulatory contexts (de Vries et al., 2022; Smets et al., 2018). However, across these designs, prompts are issued at random times or at fixed clock-time intervals (daily, hourly) and ask participants to recall stress at the present moment or over a recent window. Their timing is therefore decoupled from the physiological state that motivates the assessment in the first place.

An alternative approach, and the one we adopt here, is to let physiology itself serve as the timing signal. Wearable sensors continuously monitor heart rate and issue prompts for self-report when individualized physiological thresholds are exceeded (Hoemann et al., 2020; Lumley et al., 2014; Schwerdtfeger & Rominger, 2021). Lumley and colleagues piloted this design across a single workday with female managers, and we extend it here to a long-duration occupational

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

deployment in police trainees. This design departs from fixed-schedule assessment in an important way, by ensuring that each prompt coincides with a period of sustained physiological elevation; crucially, however, it does not equate elevated heart rate with stress. The threshold serves only to identify moments when something physiologically notable is happening; the participant's self-report then determines whether that episode corresponds to felt stress or not.

This design also creates a natural opportunity to examine how cardiac activity is organized in the period *leading up to* a stress report. As argued in Section 1.2, allostatic accounts characterize regulation as anticipatory rather than purely reactive, and if such anticipatory adjustments occur, the temporal organization of heart rate during the preceding window may carry information about the impending state. Our feature families provide a descriptive vocabulary for characterizing such anticipatory shifts, should they exist, without requiring us to specify in advance which features will shift or in which direction. The question is thus empirical: Do any aspects of how heart rate is organized over the preceding period reliably differentiate stressed from non-stressed reports?

To address this question, we extract features from the 30, 60, and 90 minutes of heart-rate data preceding each physiologically-triggered prompt. These three windows serve complementary purposes. Shorter windows (30 minutes) capture any organizational differences that emerge close in time to the stress report. Longer windows (60 and 90 minutes) provide a broader temporal context in which more gradual shifts in cardiac organization, if present, have more time to manifest. The choice of multiple windows is also pragmatic. With heart rates sampled approximately every 5–15 seconds, longer segments yield more data points and more stable feature estimates, particularly for measures of temporal structure and recurrence that depend on having sufficient series length.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

2 Method

All procedures were conducted in accordance with APA ethical guidelines and approved by the university's Institutional Review Board.

2.1 Participants

Data were collected at the Phoenix Police Regional Training Academy, where recruits completed a 26-week training program involving orientation, lecture instruction, field and scenario exercises, firearms training, and physical conditioning. Individuals were recruited from five consecutive training cohorts. Members of the research team presented the study to each cohort in a classroom setting during the initial two weeks of the Academy ("Pre-Academy"), as coordinated with training leadership (Sergeant, Recruit Training Officer). Recruits were informed that participation was voluntary and that they could withdraw at any time. Interested and willing recruits provided consent for the questionnaire and physiological data portions of the study (i.e., separate consents; those agreeing only to complete one-time questionnaires are not reported here). All recruits were paid as employees throughout training (i.e., as part of their enrollment in the Academy) but research participants did not receive additional compensation for study participation.

Our sample comprised 33 trainees who participated in the Fitbit monitoring portion of the study and had usable physiological data (both heart rate and activity level, per below). There were 15, 10[1], 1, 6, and 2 trainees per cohort, respectively, reflecting differences in overall cohort size enrollments (due to the Academy's shift to smaller, more frequent cohort enrollments).

---

1 These per-cohort counts sum to 34 because one trainee changed cohorts mid-study and is therefore counted in both Cohort B and Cohort C. This trainee contributed the majority of the data during the Cohort B weeks and minimal data during the Cohort C weeks, minimizing person-based dependency in the data. Models accounted for and were robust to this cohort shift.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Participant demographics are in Table 1. Of the 33 trainees, 28 were male and 5 were female; participants were predominantly in their 20s, consistent with broader trends in police academy enrollment. Race/ethnicity reflected the broader Phoenix area demographic, with multiple backgrounds represented but predominantly white and Latino/a.

Table 1

Participant Demographics

Variable	*M*	*SD*
Age in years (range 20 - 50)	25.9	6.6
Years of education (range 12 - 18)	14.6	1.7
	N	%
Gender / sex		
Male	28	84.8
Female	5	15.2
Nonbinary	0	0.0
Ethnicity / race		
Non-Hispanic White	12	36.4
Hispanic / Latino (-a)	10	30.3
Black / African American	3	9.1
Asian / Asian American / Pacific Islander	1	3.0
Biracial / Multiracial	5	15.2
Participant specifications:		
White / Asian	2	
White or Caucasian / Spanish	2	
White / Hispanic/Latino / Native Amer.	1	
Missing data	2	6.1
Marital status		
Single	16	48.5
Married or cohabiting	9	27.3
In a relationship not living together	6	18.2
Missing data	2	6.1
Smoker status (endorses smoking)	0	0.0
Overall subjective health rating		
Excellent	8	24.2

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Very Good	19	57.6
Fair	4	12.1
Poor	0	0.0
Missing data	2	6.1

2.2 Physiological Data Collection

Heart rate and activity data were collected using Fitbit Versa 3 devices and a modified version of the companion Fitbit application installed on each participant's personal smartphone (see below). The research team set up Fitbit devices on-site immediately after participants provided consent and returned periodically to synchronize devices (and at the middle and end of the academy to collect questionnaires and additional physiological measures for some cohorts; questionnaire data not reported here also were collected from an initial cohort to characterize social and emotional attributes of the target population pre- and post-Academy before introducing the Fitbit component of the study). Accounts were assigned and managed by the research team. Each trainee was assigned a unique identifier to maintain anonymity. Height, weight, age, and gender were entered into the Fitbit application when prompted during set-up for calculation of preset Fitbit metrics (e.g., metabolic parameters). Each participant wore their device continuously during on-academy and off-academy hours—including while sleeping and on weekends—to ensure comprehensive monitoring in naturalistic contexts. Participants were encouraged to take occasional "wrist breaks" to charge their devices. Monitoring duration varied between 12 and 23 weeks, depending on scheduling and logistical constraints, with data collection lasting 13, 23, 12, 18, and 13 weeks for participants in each of the five respective cohorts. Data collection from each cohort began and ended at different times based on their Academy start date.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Fitbit signal measurement. The Fitbit devices collected heart rate data through their optical photoplethysmography (PPG) sensors, which detect pulsatile changes in blood volume at the wrist. Internally, the device processes these optical signals to derive estimates of cardiac timing, but rather than exposing raw inter-beat intervals, Fitbit makes accessible only processed heart rate values in beats per minute (bpm) at 5–15 second intervals. Additionally, the devices calculated resting heart rate values daily, which served as personalized baseline measurements for each participant. Metabolic equivalents (METs) were also collected at 1-minute intervals. All biometric data were stored locally on the device, synced to a paired smartphone via Bluetooth, uploaded to Fitbit's secure servers, and accessed for research purposes through Fitabase, a platform widely used in wearable research for managing device enrollment, monitoring compliance, and exporting Fitbit data in research-ready formats (Fitabase, n.d.). As a verification step, heart rate data were also retrieved directly from Fitbit's API and matched against the Fitabase exports.

2.3 Self-Reported Stress Prompting and Assessment

To capture stress ratings in response to naturally-occurring events, a custom-built Fitbit application was developed and deployed via the Fitbit "clockface" (via the API) that enabled real-time, event-triggered stress reporting. The application continuously monitored heart rate streams locally on the watch and applied a thresholding algorithm: if current heart rate rose 35% above baseline for 120 consecutive seconds, the watch generated an in-situ notification prompting the participant to rate their stress. This thresholding approach adapts prior physiologically-triggered assessment paradigms (Hoemann et al., 2020; Jetelina et al., 2022; Lumley et al., 2014). The values used here are stricter than those in Lumley et al.'s pilot implementation of 20% above baseline (Hoemann et al. [2020] applied a prompting threshold of

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

±167 msec IBI change without a postural shift over eight seconds), reflecting the higher cardiac activity and more frequent extended departures from baseline expected for police (Anderson, Litzenberger, & Plecas, 2002) during multi-week academy training compared to a single sedentary workday. As in Lumley et al., parameters were finalized through internal testing before deployment, which supported raising the threshold from 120% to 135% of baseline to avoid prompting too frequently. Thresholds were not as high as those used in Jetelina and colleagues (2022; 50-80% for 10 minutes), given that we aimed to capture a range of stress reports but were not implementing an intervention.

Stress assessment consisted of a single item displayed on the Fitbit screen: "How stressed do you feel right now?" rated on a 4-point Likert scale (1 = no stress, 2 = a little, 3 = moderate, 4 = a lot).[2] Participants could respond immediately using the Fitbit's touchscreen interface or dismiss the prompt. Non-responses within 7 minutes were automatically coded as 0 in the database. To avoid notification fatigue, the system implemented a cooldown period of 30 minutes between successive prompts. Response rates varied among participants due to training schedules, environmental constraints, and individual engagement levels. Notification directly on the Fitbit was a novel feature of the present study and especially relevant as academy trainees are not permitted to carry phones during the training day. As with the physiological data, stress prompt responses were stored locally, transferred to the paired smartphone via Bluetooth, and uploaded to secure cloud storage.

2 The 1–4 rating was preceded, at the beginning of the earliest cohort, by a 5-point pull-down scale that was replaced with a 4-point single-screen scale as ratings of "5" were extremely rare; see Supplementary E.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Daily ecological momentary assessments (EMAs) were also administered via web-based surveys, but response rates were inconsistent and these data were not analyzed in the current study.

2.4 Integration of Data Streams

All data streams were merged into a unified time series by aligning timestamps. Stress prompt responses were linked to the closest preceding heart rate value within a 5-second tolerance, ensuring that physiological measures reflected states prior to, rather than following, self-reports. Each participant's daily resting heart rate estimate from Fitbit was joined onto the unified time series by calendar date, so every HR sample (and any stress prompt aligned to it) carried that day's resting HR as a covariate; for three participants missing these values, a population-norm default of 58 bpm was imputed.

Activity measures (steps and metabolic equivalents, recorded at 1-minute intervals) were synchronized with the 5-second heart rate series using backward matching and forward-fill interpolation, providing continuous contextual information. All alignment procedures were implemented to preserve temporal order, so that physiological and activity predictors always preceded or coincided with stress reports.

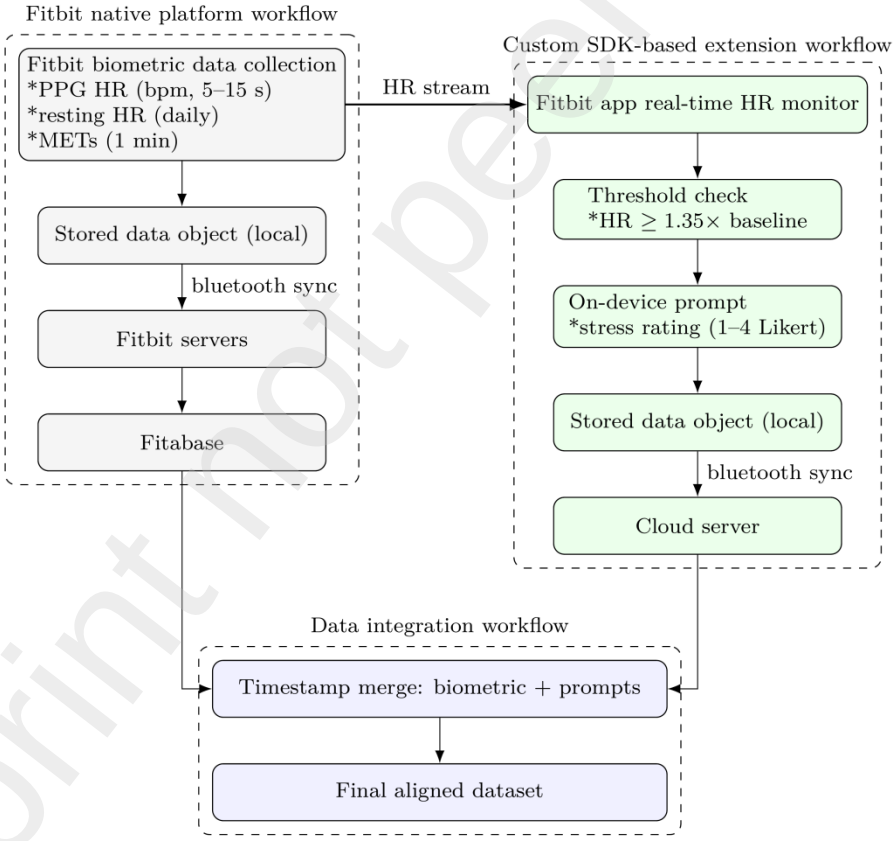
The complete workflow from physiological data collection to integration of data streams is summarized in Figure 1.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Figure 1

Workflow separation. Left: Fitbit native platform workflow—biometric data are collected on-device, stored locally, synced via Bluetooth to Fitbit servers, and exported through the Fitabase research platform. Right: Custom SDK-based extension workflow—the custom application monitors heart rate in real time, applies the threshold check, delivers stress prompts, and uploads responses to a separate cloud server. Bottom: data integration pipeline merging both streams into a single timestamp-aligned dataset. SDK = Software development kit. Asterisks mark the specific data streams or parameters within each stage.



Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

2.5 Heart Rate Segment Extraction

To examine physiological states leading up to stress, we extracted non-overlapping heart rate segments preceding each stress event. For every event where a participant reported stress (rating "2" or above), we collected the prior 30, 60, and 90 minutes of continuous heart rate data, yielding approximately 120–360 observations per 30-minute segment, 240–720 per 60-minute segment, and 360–1080 per 90-minute segment. Segments were retained only if they contained at least 90% of the target duration and had no data gaps longer than five minutes. Once a valid segment was identified, any other events within that time window were disregarded to ensure temporal independence. In other words, no extracted segment overlapped with another stress event.

For events marked as "1" (not at all stressed), we also extracted the preceding segment of heart rate data, provided no stress event occurred within that window. If a reported stress event did occur, it replaced the no stress reported event segment through a priority reset. This procedure ensured that the final dataset consisted of distinct, independent heart rate segments that reflected physiological states immediately before stress and non-stress experiences.

2.6 Heart Rate Data Quality Assessment

Consumer wearables enable continuous physiological monitoring in naturalistic settings, but this ecological validity comes at the cost of precision. Fitbit reports a heart rate estimate only at 5-15 seconds, and at that resolution artifacts can imitate the very patterns our features are designed to capture (e.g., a flatline can appear as rigidity, a spike as chaos). To counter this risk, we built a quality protocol that sorts every segment by how usable it is, then retains, repairs, or excludes it.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

We scored each segment and sorted it into one of three classes. We excluded as unsalvageable any segment that failed any of four critical checks: long sensor dropouts (extended flatlines), physiologically impossible jumps between successive readings, values outside a plausible heart-rate range, or so much spike-level noise that even a corrected segment could not be trusted. Segments that passed every check we treated as clean and analyzed as-is. The intermediate, salvageable segments cleared exclusion thresholds but carried one or more correctable problems. Each received only the corrections its problems triggered, i.e., detrending for slow drift, Hampel filtering for isolated spikes, and interpolation for short flat gaps. These were applied in a fixed order when more than one was needed. Supplementary D gives the exact thresholds and the trigger for each correction.

Table 2 reports the number of segments for each window and how they were classified. Across the 30-, 60-, and 90-minute windows we extracted 2,153, 1,500, and 1,159 segments and retained 1,990, 1,426, and 1,100 after excluding the unsalvageable ones. The proportion of clean segments fell modestly as the window lengthened (from 64.6% at 30 minutes to 59.3% at 90), likely reflecting longer windows having more opportunities to accumulate artifacts, while the unsalvageable rate stayed low throughout (4.9–7.6%). Retention per participant was substantial but uneven at every window; at 90 minutes, for example, the median participant contributed 26 segments (range of 1 to 113). See Supplementary E for the full per-participant distributions across windows.

We validated the protocol in two ways. The first asked whether our corrections preserved the moment-to-moment variability that the features depend on. We summarized that variability as the standard deviation of successive heart-rate differences (ΔHR - SD), the quantity our complexity, recurrence, and dynamics features are designed to respond to, and compared it

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

before and after correction. We illustrate the result at the 90-minute window, our longest and most artifact-prone. Across its 413 corrected segments, the post- to pre-correction ratio held near one, and 97% of segments stayed within $\pm$ 10% of their original value.

The second validation asked whether segment quality was entangled with stress. Quality class was modestly but significantly related to stress at all three windows (χ^2 tests, all ps < .05; Supplementary E). In every case the direction of this association was reassuring in that unsalvageable segments preceded proportionally fewer stress reports than clean ones (18.6% versus 32.0% at 90 minutes). Removing unsalvageable segments therefore discarded mostly non-stress data, supporting the idea that the exclusion works against detecting a stress effect.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Table 2

Per-Window Segment Yield and Quality Classification

	30-min	60-min	90-min
Extracted segments	2,153	1,500	1,159
Non-stress (rating $=$ 1)	1,633 (75.9%)	1,099 (73.3%)	827 (71.4%)
Stress (rating $\geq$ 2)	520 (24.1%)	401 (26.7%)	332 (28.6%)
Quality classification			
Clean (no correction)	1,391 (64.6%)	939 (62.6%)	687 (59.3%)
Salvageable (corrected)	599 (27.8%)	487 (32.5%)	413 (35.6%)
Unsalvageable (excluded)	163 (7.6%)	74 (4.9%)	59 (5.1%)
Retained for analysis	1,990	1,426	1,100
Per-participant: median [min, max]	46 [1, 192]	35 [1, 136]	26 [1, 113]

Note. All counts based on 34 cadets with complete baseline resting heart-rate and activity (METs/steps) data. The three windows are independent analyses (segments extracted as 30-, 60-, and 90-minute non-overlapping windows preceding each stress-prompt response), not nested. Clean segments proceed directly to feature extraction; salvageable segments receive ordered corrections (polynomial detrending, delta-based Hampel filtering, linear interpolation; see Supplementary D); unsalvageable segments are excluded. Per-participant retention counts exclude unsalvageable segments.

2.7 Feature Selection and Generation

We created 14 features organized into four theoretically-motivated families, each an index of temporal organization. Across these families, the features describe how wide the heart-rate band is (Cardiac Baseline), how predictable short patterns are (Cardiac Complexity), how often and how asymmetrically the signal makes step changes (Cardiac Dynamics), and whether it revisits or dwells in similar local configurations over time (Pattern Recurrence). What follows is

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

a high-level description of each index, a more detailed explanation is provided in Supplementary A.[3]

2.7.1 Cardiac Baseline

This family captures the range and variability within which cardiac activity operates. Its features require minimal timing assumptions and are robust to our sampling resolution. From a flexibility–structure perspective, constrained systems may exhibit a narrower operating band and reduced variability compared to more flexible systems. We computed three measures from each filtered HR segment:

- *standard deviation* of all HR values, capturing absolute variability magnitude
- *range*, max HR minus min HR, defining the full span of the operating band
- *coefficient of variation*, the standard deviation of HR divided by mean HR, providing scale-independent variability that enables cross-participant comparison regardless of baseline HR level. Unlike standard deviation or range alone, this measure expresses spread relative to each segment's typical operating level, so two segments with equal absolute spread but different mean rates can still differ in coefficient of variation

These dispersion measures are interpretively ambiguous on their own. A narrow range can reflect rigidity or stable regulation, and a wide range can reflect flexible responsiveness or dysregulation. The features described in the families below help disambiguate which.

3 Permutation entropy was computed with the antropy Python package (v. 0.1.9); DFA with nolds (v. 0.6.2); RQA with PyRQA (v. 8.1.0). Remaining features were computed with numpy and scipy.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

2.7.2 Cardiac Complexity

This family describes the temporal organization and predictability of HR patterns across multiple scales. These features locate each segment along a continuum from highly regular sequences, where successive values follow repetitive patterns, to less patterned sequences, where successive values show little sequential structure, i.e., from organized to unorganized variability.

Permutation entropy (Bandt & Pompe, 2002) quantifies unpredictability in HR sequences by analyzing the shape of every triplet of three consecutive values – the order in which they rise and fall – rather than their actual magnitudes. Taking each set of three readings in turn, the algorithm records only which value is smallest, which is middle, and which is largest, giving one of six possible shapes (for example, a steadily rising triplet, or one that dips and then rebounds). It then computes the Shannon entropy of how often each of the six shapes occurs across the segment. Values near 1.0 indicate that all shapes occur about equally often (maximum disorder), while lower values indicate that a few shapes dominate (more predictable sequential structure). Because it depends on rank ordering rather than absolute values, this measure is robust to measurement noise and well suited to coarsely sampled data.

Detrended fluctuation analysis (DFA) (Hardstone et al., 2012) produces a scaling exponent that summarizes how much the past constrains the future of the series at a given timescale. In plain terms, DFA asks whether the roughness of the heart-rate series looks similar when examined over shorter versus longer windows, and reports the answer as a single scaling exponent at each scale. After integrating and detrending the HR series, we computed scaling exponents for two scale ranges: *DFA short range* (windows of 10–60 data points, approximately 50–300 seconds) and *DFA long range* (windows of 60–240 data points, approximately 5–20 minutes). Values near 0.5 indicate that successive fluctuations are essentially independent — the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

recent past carries little information about what comes next. Values near 1.0 indicate balanced scaling, such that variation at short timescales has similar statistical character to variation at longer ones. Values near 1.5 indicate that the series is dominated by slow drifts in which it persistently moves in one direction.

2.7.3 Cardiac Dynamics

This family characterizes moment-to-moment changes in heart rate through the direction, frequency, and consistency of successive differences. Where the Cardiac Baseline features describe how much the series varies overall, these features describe how that variation unfolds from one reading to the next.

Directional asymmetry compares the typical magnitude of upward versus downward heart-rate changes (roughly analogous to deceleration/acceleration capacity measures). We computed the mean of all positive successive differences (upward steps) and the mean of all absolute negative successive differences (downward steps), and report their ratio. A ratio near 1 reflects symmetric step magnitudes; a ratio greater than 1 indicates that upward steps tend to be larger than downward steps; a ratio less than 1 indicates the reverse.

Successive change frequency counts the percentage of absolute successive HR differences exceeding 1 bpm. This captures moment-to-moment volatility that may persist despite overall range reduction.

The *interquartile range (IQR) of successive differences* captures the consistency of change magnitudes: it is the IQR of the absolute successive HR differences. Larger values indicate more variable adjustment sizes; smaller values indicate that successive changes are more uniform in magnitude.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

2.7.4 Pattern Recurrence

This family uses Recurrence Quantification Analysis (RQA) (Marwan et al., 2007) to characterize how often, and for how long, a heart-rate series revisits states it has occupied before. In plain terms, RQA asks whether short stretches of the heart-rate series recur over time, and whether those recurrences take the form of repeated trajectories (the series following the same path twice) or of the series lingering in similar states for a stretch. RQA embeds the series in a reconstructed state space and identifies recurrences. These are time points at which the trajectory returns close to a previously visited state (within a threshold of 5% of the data range; embedding parameters and computational details are reported in Supplementary A). The recurrence matrix can be visualized as a plot in which both axes represent time: a point at coordinate (i,j) marks moments when the state of the heart-rate series at time i closely resembles its state at time j. Patterns within this dot-cloud — diagonal stripes, vertical bars, isolated points — encode different aspects of how the series organizes itself over time. From these patterns we computed five measures.

Three describe diagonal-line patterns in the recurrence matrix, which arise when the heart-rate series follows similar trajectories through state space on different occasions. *Determinism* is the fraction of recurrent points that fall on such diagonal lines, with higher values indicating recurrence organized into sequential trajectories rather than isolated coincidences. *Entropy*, the Shannon entropy of the distribution of diagonal-line lengths, captures the diversity of trajectory durations. *Divergence*, the inverse of the longest diagonal line, indexes how quickly initially similar trajectories separate.

The remaining two measures describe vertical-line structures in the recurrence matrix, which arise when the heart rate dwells in approximately the same state for multiple consecutive

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

time points. *Laminarity* is the fraction of recurrent points that fall on these vertical lines and it indexes how prevalent dwelling episodes are. *Trapping time* is the average length of those vertical lines (the typical duration of a dwell), such that higher values indicate that the heart rate persists in a given configuration before transitioning; lower values indicate more frequent movement between states. Computational details and parameter specifications for all five measures are provided in Supplementary A.

2.8 Analytic Strategy

As previewed in Section 1, we adopted a two-stage analytic approach. Stage 1 used Monte Carlo cross-validation to identify which features predict stress when models are evaluated on participants whose data they have not seen; we then carried the most robust features forward to Stage 2, which fit a single explanatory generalized linear mixed model (GLMM) on the full dataset to estimate effect directions, magnitudes, and confidence intervals.

2.8.1 Stage 1: Prediction and Feature Selection

In Stage 1 we evaluated out-of-sample predictive validity through Monte Carlo cross-validation with participant-level splits. Across 50 iterations, we randomly partitioned participants into 80% training and 20% testing sets (27 and 7 participants, respectively). Critically, splits occurred at the participant level, not the observation level: all segments from a given participant appeared exclusively in either the training or testing set within each iteration. This ensured that the model was evaluated on its ability to generalize to new individuals, not merely new observations from familiar participants.

Because the four feature families represent theoretically distinct constructs, we applied variance inflation factor (VIF) filtering within rather than across families, ensuring that at least one feature from each family entered every model. Within each training set, VIF was calculated

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

separately for each family, and features were iteratively removed in order of highest VIF until all remaining features within a family fell below VIF = 3 or only one feature in that family remained.

After VIF filtering, we fit a GLMM predicting stress occurrence (binary outcome: stress report present vs. absent) from the retained features, controlling for mean METs (activity level) and time of day. The model included random intercepts for participants; nested (1|cohort/user) and crossed (1|user + 1|cohort) cohort/user specifications were also evaluated but produced singular fits at the cohort level, so the simpler participant-only structure was retained (see Supplementary C). All models were fit with glmer from the lme4 package in R (Bates et al., 2015).

$$\text{stress} \sim \text{cardiac features} + \text{mean METs} + \text{time of day} + (1|\text{participant}),$$

where "cardiac features" denotes the post-VIF retained subset for the given iteration; the random intercept "(1|participant)" captures stable between-person differences in baseline stress reporting. We evaluated model performance on the held-out test set using area under the ROC curve (AUC) as the primary metric, with 95% confidence intervals from a participant-stratified percentile bootstrap.

To assess individual feature contributions, we computed *permutation importance* within each test set. For each feature in turn, we randomly shuffled its values across observations (breaking its relationship with the outcome while preserving distributions and correlations among other features), re-predicted outcomes using the trained model, and calculated the drop in AUC:

$$\Delta\text{AUC} = \text{AUC}_{\text{original}} - \text{AUC}_{\text{shuffled}}$$

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Positive values indicate that a feature contributed to prediction; values near zero indicate negligible contribution. Repeating this across the 50 iterations yielded a distribution of ΔAUC for each feature, from which we calculated the mean and a 95% percentile bootstrap confidence interval. We classified a feature as a robust predictor when its mean ΔAUC exceeded zero and its bootstrap confidence interval excluded zero.

In parallel, we performed family-level permutation importance by shuffling all features within a theoretical family simultaneously and calculating the resulting drop in AUC. This tests whether each construct as a whole (e.g., Pattern Recurrence, Complexity) contributes meaningfully to prediction, validating our theoretical organization. If a family shows significant collective importance, it confirms that the construct it represents is relevant to stress detection, even when individual features within it show variable importance due to multicollinearity or person-specific patterns.

To confirm that the Stage 1 results were not specific to the primary modeling choices, we ran two sensitivity analyses in parallel: a GLMM using global rather than family-preserving VIF filtering, and an elastic-net logistic regression ($\alpha = 0.5$) with participant-grouped internal cross-validation. Both were evaluated on the same 50 participant-level splits as the primary model (see Supplementary C).

2.8.2 Stage 2: Explanatory Modeling

Although Stage 1 establishes which features predict stress and which contribute most robustly, it does not directly quantify effect sizes or test the statistical significance of individual coefficients. We addressed this in Stage 2 by fitting a single explanatory GLMM on the full 90-minute dataset (the best-performing window from Stage 1). We carried forward only the features that proved stable across the 50 Stage 1 splits on four counts: they were retained by VIF filtering

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

in at least 90% of splits, held the same direction in at least 80%, showed a mean ΔAUC above zero and had a bootstrap confidence interval that excluded zero. Supplementary B evaluates every feature against all four.

The Stage 2 model followed the same structure as Stage 1, fit by maximum likelihood with glmer from lme4. All continuous predictors were z-scored prior to modeling so that standardized coefficients are directly comparable across features measured on different scales. We report standardized coefficients, odds ratios with 95% Wald confidence intervals, and Wald z-test p-values from the model summary. Because the retained cardiac features constitute a single hypothesis family, we applied a Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) to those feature p-values; covariate p-values were not corrected.

We do not claim that Stage 2 establishes independent out-of-sample validity, since it uses the same data for interpretation that informed feature selection in Stage 1. The role of Stage 2 is to quantify the size and direction of the relationships whose generalization has already been assessed in Stage 1.

Our sample size of 33 cadets was determined by access to five training cohorts at the Phoenix Police Regional Training Academy to test feasibility of our novel approach rather than by an *a priori* power calculation (Lin et al., 2026). Because the study evaluates 14 features organized into four theoretically distinct families rather than testing a single hypothesized effect, sample-size justification instead emerges from the analytic strategy itself, such that features whose effects are too small or inconsistent at the present sample size do not survive Stage 1's conjoint criteria across 50 participant-stratified resamples.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

3 Results

3.1 Stage 1: Stress Prediction Performance

We evaluated whether heart rate dynamics measured 30, 60, and 90 minutes before self-reported stress could predict the occurrence of heightened stress states. Prediction performance improved systematically as the temporal window approached the stress event (Table 3). At 30 minutes before stress, the model achieved modest but consistent above-chance discrimination (AUC mean = 0.54, 95% CI [0.53, 0.55]). Performance strengthened at 60 minutes (AUC mean = 0.55, 95% CI [0.54, 0.56]) and improved further at 90 minutes before stress reports (AUC mean = 0.60, 95% CI [0.58, 0.62]). Median AUC values closely tracked mean performance (30 min: median = 0.54; 90 min: median = 0.61). Together, these patterns suggest that physiological precursors of subjective stress become increasingly detectable as they unfold over longer timescales, though effect sizes remain modest throughout.

Additionally, performance variability across folds, captured by the interquartile range (30 min: IQR [0.52, 0.57]; 90 min: IQR [0.54, 0.64]), indicates individual differences. These are expected given how physiological dynamics map onto subjective stress experience, as well as variation in the contexts and stressors encountered during police training.

Table 3

Performance metrics across different window sizes

Window	Splits	AUC (mean)	95% CI	AUC (median)	IQR
30 min	50	0.541	[0.531, 0.551]	0.539	[0.516, 0.567]
60 min	50	0.546	[0.536, 0.557]	0.545	[0.520, 0.568]
90 min	50	0.597	[0.577, 0.616]	0.608	[0.538, 0.636]

Note. Cross-validated prediction performance for stress occurrence at 30, 60, and 90 minutes before self-report. AUC = area under the ROC curve; CI = 95% bootstrap confidence interval;

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

IQR = interquartile range across 50 user-level cross-validation folds. All analyses used generalized linear mixed-effects models with VIF-based multicollinearity control (primary model).

We also conducted two sensitivity analyses that confirmed that the reported pattern of results was not model-specific. We reran Stage 1 analyses with global VIF filtering (versus within each feature family) and an elastic net with user-grouped cross-validation. Both models yielded consistent results across all windows (Supplementary C).

3.2 Stage 1: Cardiac Feature Importance

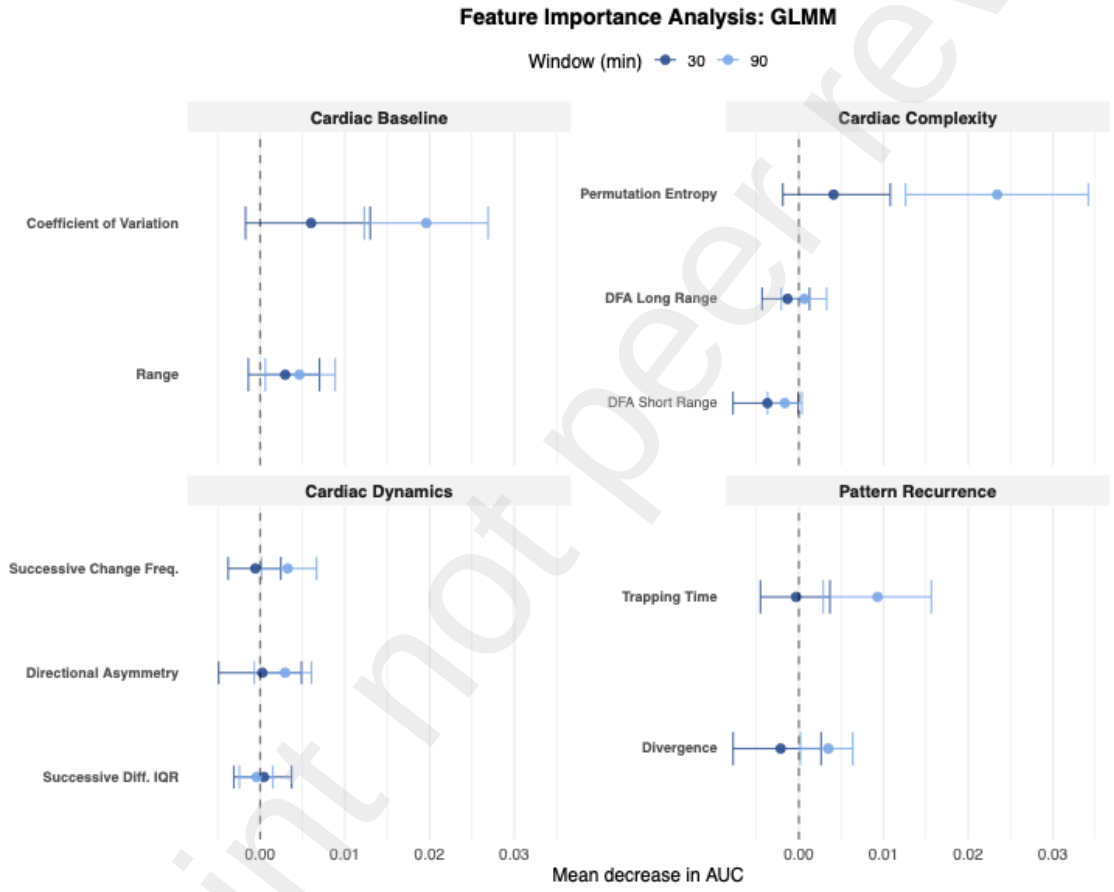
We examined feature importance at 30- and 90-minute windows to capture how predictive dynamics evolve across timescales (Figure 2).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Figure 2

Feature Importance Analysis. Each panel represents a feature family. Points indicate mean change in AUC when features are permuted across 50 data splits; error bars show 95% confidence intervals. Dashed line marks the 0.01 threshold. Features are ordered by maximum importance across windows. Bold labels indicate features meeting the significance threshold in at least one window.



At 30 minutes before stress reports, no features met all four selection criteria. While several features showed high inclusion and directional consistency, none demonstrated permutation importance with confidence intervals reliably excluding zero. This pattern indicates

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

that while features contribute to prediction at 30 minutes, individual features lack robust, statistically certain predictive power at this short timescale.

At 90 minutes, three features met all four criteria, spanning three of the four feature families (the complete evaluation of all features against each criterion is reported in Supplementary B). From the Cardiac Baseline family, *coefficient of variation* (drop = 0.020, 95% CI [0.013, 0.028]) showed consistent negative associations with stress (100% directional consistency), indicating that reduced relative heart rate variability preceded stress reports. From the Cardiac Complexity family, *permutation entropy* (drop = 0.023, 95% CI [0.012, 0.034]) showed consistent positive associations (100% consistency), indicating that increased local unpredictability in successive heart-rate patterns preceded stress. From the Pattern Recurrence family, *trapping time* (duration in similar states; drop = 0.009, 95% CI [0.003, 0.016]) qualified with positive associations (92% consistency), suggesting that prolonged persistence in particular cardiac states preceded stress at this longer temporal horizon.

These three features exhibited variable effect sizes. *Coefficient of variation* and *permutation entropy* demonstrated the strongest individual contributions (drops of 0.020 and 0.023 respectively), while *trapping time* showed a more modest but statistically certain effect (drop of 0.009).

Several additional features approached but did not meet all selection criteria. *Divergence* (rate of trajectory separation) from the Pattern Recurrence family showed 100% inclusion, 94% directional consistency, and a positive mean drop (0.003), but its confidence interval included zero (CI [− 0.001, 0.006]), failing the statistical certainty requirement. Similarly, *detrended fluctuation analysis (long range)* from the Cardiac Complexity family showed 100% inclusion and 100% consistency but a confidence interval spanning zero (CI [− 0.001, 0.004]). No

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

features from the Cardiac Dynamics family met all criteria, despite some showing high inclusion and consistency. These near-misses indicate features with consistent directional associations that nonetheless lack statistically certain individual predictive contributions, potentially due to redundancy with the selected features or weaker signal strength.

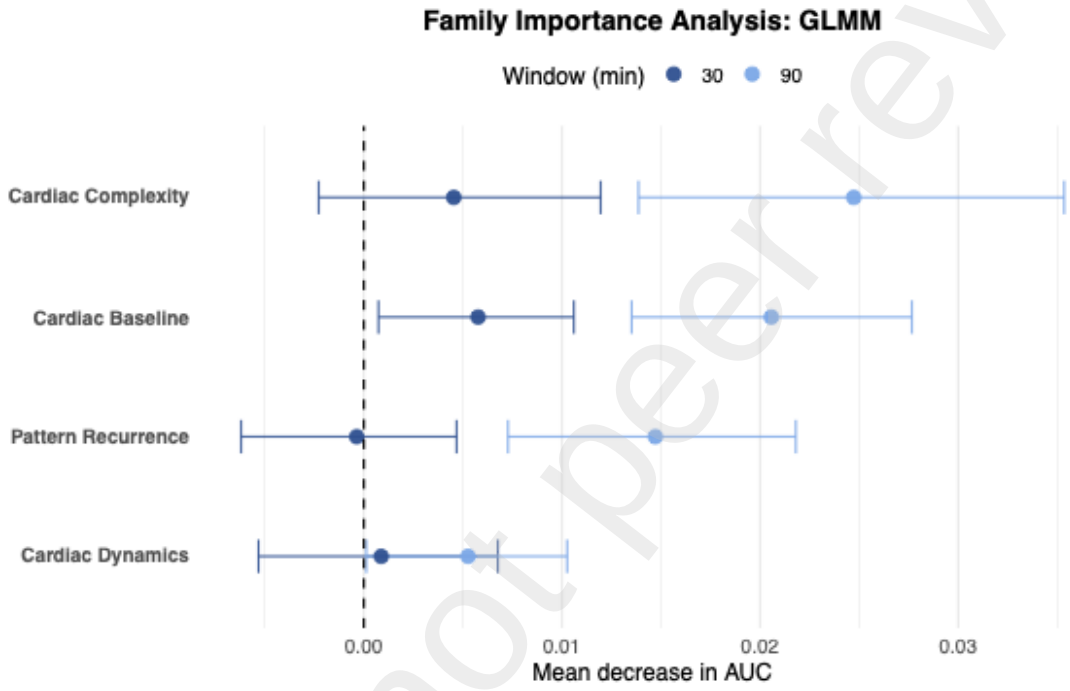
Family-level permutation analyses, which simultaneously shuffle all features within each conceptual domain, revealed coordinated patterns that substantially exceeded individual feature contributions (see Figure 3). At 30 minutes, only the Cardiac Baseline family showed a statistically certain effect (mean drop = 0.006, 95% CI [0.002, 0.011]), while the other three families exhibited confidence intervals overlapping zero. At 90 minutes, three of the four families demonstrated statistically certain aggregate importance, ranked by effect size: Cardiac Complexity (mean drop = 0.025, 95% CI [0.014, 0.036]), the largest family-level effect, followed by Cardiac Baseline (mean drop = 0.021, 95% CI [0.014, 0.028]), Pattern Recurrence (mean drop = 0.014, 95% CI [0.007, 0.022]). The fourth family, Cardiac Dynamics, did not reach statistical certainty (mean drop = 0.005, 95% CI [─ 0.0002, 0.010]).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Figure 3

Feature Importance Analysis. Points indicate mean change in AUC when all features within a family are permuted across 50 data splits; error bars show 95% confidence intervals. Dashed line marks the 0.01 threshold. Families are ordered by maximum importance. Bold labels indicate families meeting the significance threshold in at least one window.



The stark contrast between windows, where zero features meeting criteria at 30 minutes versus three at 90 minutes, and one family showed a reliable effect at 30 minutes versus three at 90 minutes, reveals a temporal accumulation process in stress precursors. At 30 minutes, physiological signals remain diffuse. While the Cardiac Baseline family collectively contributes to prediction, individual features lack statistical certainty, and other families show minimal aggregate effects. At 90 minutes, organized patterns emerge across multiple domains: reduced relative variability (*coefficient of variation*) and increased local unpredictability (*permutation*

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

entropy) provide the strongest individual signals, while *trapping time* contributes additional information about persistence in particular cardiac states.

3.3 Stage 2: Explanatory Modeling

We fitted an explanatory generalized linear mixed-effects model (GLMM) to the complete 90-minute dataset using the three features meeting all selection criteria in the previous step: *coefficient of variation*, *permutation entropy*, and *trapping time* (Table 4 shows the full results). Feature p-values were corrected for multiple comparisons using the Benjamini–Hochberg procedure.

Unlike the cross-validation approach used for feature selection, this explanatory model leverages the full dataset's statistical power to estimate precise effect sizes and test coefficient significance while controlling for time of day and physical activity. The model included random intercepts for participants to account for individual differences in baseline stress propensity. All predictors were standardized (z-scored) prior to modeling, allowing direct comparison of effect magnitudes across features measured on different scales. Variance inflation factors for all predictors were below 3.0, indicating acceptable multicollinearity (Supplementary C).

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Table 4

Mixed-Effects Logistic Regression Results (90-Minute Window). Model includes all three features meeting Stage 1 selection criteria, controlling for time of day and physical activity. All continuous predictors are standardized (z-scored). Feature p-values are Benjamini–Hochberg corrected. OR = odds ratio; Std. β = standardized coefficient on the logit scale. $^{*}p < .05$, *** $p < .001$.

Predictor	OR [95% CI]	Std. β	p-value
Covariates			
Time of Day (ref: Evening)			
Morning	1.21 [0.80, 1.82]	0.190	0.361
Afternoon	1.13 [0.76, 1.69]	0.126	0.535
Physical Activity (METs)	1.12 [0.93, 1.36]	0.115	0.239
Cardiac Features			
Coefficient of Variation	0.70 [0.53, 0.93]	− 0.354	0.021*
Permutation Entropy	1.58 [1.25, 2.01]	0.460	<0.001***
Trapping Time	1.27 [0.96, 1.67]	0.239	0.090

n = 1,097 (34 users, 5 cohorts); Marginal R^2 = .046; Conditional R^2 = .432

Two of the three selected features demonstrated statistically significant associations with stress occurrence after correction for multiple comparisons. *Permutation entropy* showed the strongest association: a one SD increase in local unpredictability was associated with 58% increased odds of stress (OR = 1.58, 95% CI [1.25, 2.01], $p < .001$). This effect remained highly significant even after accounting for individual differences and covariates, indicating that erratic, less predictable moment-to-moment heart rate patterns robustly preceded stress reports by 90 minutes. *Coefficient of variation* showed a significant negative association (OR = 0.70, 95% CI [0.53, 0.93], $p = .021$): a one standard deviation increase in relative heart rate variability was associated with 30% decreased odds of stress. In other words, reduced relative variability—more constrained heart rate dynamics—preceded stress. These two features, which also showed the

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

strongest permutation importance in Stage 1, emerge as the most reliable individual markers of impending stress at the 90-minute horizon.

Trapping time, the sole Pattern Recurrence feature meeting all Stage 1 selection criteria, showed a positive trend consistent with its cross-validated direction (OR = 1.27, 95% CI [0.96, 1.67], p = .090) but did not reach conventional significance thresholds after correction. This pattern—meeting cross-validated selection criteria without achieving significance in the full-dataset model—can occur when a feature contributes to prediction through complex interactions or non-linear patterns that cross-validation captures but traditional regression does not fully represent, or when it provides incremental predictive value that does not translate to a strong individual association after accounting for other features.

None of the control variables showed significant associations: time of day (morning vs. off-hours: OR = 1.21, p = .361; afternoon vs. off-hours: OR = 1.13, p = .535) and physical activity (mean METs: OR = 1.12, p = .239) did not independently predict stress after accounting for heart rate features. This indicates that the cardiac features capture stress-related dynamics beyond what circadian rhythms or physical exertion alone explain.

The model's variance decomposition revealed that most of the explainable variance in stress occurrence resides in stable individual differences rather than in the cardiac features themselves. Marginal R^2 (variance explained by fixed effects alone) was .046, indicating that the cardiac features and covariates account for approximately 5% of variance in stress occurrence—a modest but meaningful contribution given the naturalistic data and the 90-minute anticipatory horizon. Conditional R^2 (variance explained by both fixed and random effects) was .432, indicating that nearly half of the total variance in stress reports is captured once participant-level differences in baseline stress propensity are included. The large gap between these values reflects

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

the substantial between-person heterogeneity confirmed by the random intercept variance (SD = 1.50). Individuals differed markedly in how often they reported stress, independent of their cardiac patterns. In other words, who a person is matters considerably for stress reporting, but within that individual context, the cardiac features provide reliable anticipatory information about when stress will be reported.

4 Discussion

This research was conducted as part of an interdisciplinary initiative on stress monitoring in real-world police training environments. In the present study, we asked whether the temporal organization of consumer-wearable heart-rate data carries information about anticipatory stress before subjective self-report. Across 50 user-stratified cross-validations at three time windows (30, 60, and 90 minutes prior to a stress event), we detected modest but reliable predictive signals, with prediction stronger over longer windows. At 90 minutes, a parsimonious feature set distinguished stress from non-stress reports at roughly ten percentage points above chance (AUC = 0.60). For the feature-importance analysis, we focused on the 30- and 90-minute contrast. At 90 minutes, three individual features met all selection criteria and three of the four conceptual feature families contributed reliably at the family level. At 30 minutes, no individual features met criteria and only one family did. These temporal differences suggest the signal accumulates over time and is easier to detect across a broader window, where changes in cardiac organization have more opportunity to appear and stabilize.

4.1 Constrained unpredictability and the flexibility–structure framework

In our data, cardiac dynamics before a self-reported stress event do not show unstructured arousal. They show a specific reorganization. Three patterns appear in concert, though with asymmetric evidence: the operating band narrows (reduced coefficient of variation), local

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

unpredictability within that narrower band increases (elevated permutation entropy), and the series tends to dwell longer in particular cardiac states (trapping time, reliable in cross-validation and directionally consistent in the explanatory model, where it reaches trend rather than conventional significance). We refer to this signature as a *constrained unpredictability* pattern. Concretely, the heart-rate series in the 90 minutes before a stress report fluctuates within a smaller envelope than during non-stress windows, but movement within that envelope becomes more erratic, and the series shows some tendency to settle into recurring configurations for longer stretches. This is a coherent signature in flexibility–structure terms. We interpret it as consistent with — though not a canonical operationalization of — the loss-of-complexity account, in which adaptive regulation depends on multiscale organization that breaks down when systems are under sustained demand (Goldberger et al., 2002; Lipsitz & Goldberger, 1992). Similar losses of structured variability appear across clinical populations and aging trajectories, where reduced complexity tracks impaired regulatory function rather than reduced activity (Lipsitz & Goldberger, 1992). The pattern is also broadly consistent with allostatic accounts that frame regulation as anticipatory (Sterling, 2012), though Sterling's model does not itself predict the specific entropy and recurrence signatures observed here.

Beyond individual feature prediction, the family-level pattern speaks to the feature set's organization. At 90 minutes, three of the four conceptual families (Cardiac Baseline, Cardiac Complexity, and Pattern Recurrence) contributed reliably to prediction. In each, the aggregate effect was as large as or larger than that of its strongest individual feature, which is what we would expect if each feature captures only one facet of the underlying construct: no single feature carries the full signal, but pooling features within a family recovers more of it. Pattern Recurrence is the clearest case, where the family-level effect exceeded the contribution of

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

trapping time, its only feature to meet all selection criteria. This reading is tentative, as which individual features cross conventional thresholds may shift with a cleaner signal or larger samples. Even so, the pattern lends support to the four-family scheme as capturing distinguishable rather than redundant aspects of regulation.

4.2 Cardiac signals from consumer wearables

Our marginal R^2 in the explanatory models of approximately 5% is small in absolute terms but is also commensurate with the broader literature on ambulatory cardiac stress. Martinez et al. (2022) computed conventional HRV from millisecond-resolution beat-to-beat intervals across 14,695 momentary stress assessments in 657 information workers, yet recovered marginal R^2 values of only 1–2% in predicting those stress reports. Smets et al. (2018) reported F1-scores (a composite classification metric) of 0.43 in classifying self-reported stress levels across 1,002 office workers with research-grade multimodal sensing, only a modest improvement (roughly 20%) over a baseline of 0.36. Tutunji et al. (2023) reported group-level classification error rates of 42–48% (only modestly above chance) when distinguishing high-stress examination weeks from control weeks in 84 students wearing an Empatica E4. These designs differ from ours in signal fidelity, sample size, occupational context, and EMA timing. Yet each recovers only a small fraction of variance in self-reported stress, suggesting that population-level effects are not tied to any particular design dimension. Whatever physiology alone can recover from momentary stress reports in daily life appears to be limited. This was not entirely unexpected. Self-reported stress is shaped by cognitive appraisal, social context, and recall framing alongside physiological state (Hoemann et al., 2020), so any single channel will recover a limited share. The physiological signal itself is also distributed across people whose

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

baseline levels and reactivity vary substantially, a pattern documented across both research-grade and wearable studies (Smets et al., 2018; Tutunji et al., 2023).

Despite prediction and explained variance not showing substantial improvements over existing research, the current study does contribute a novel conceptual perspective. As argued, the cardiac series in the period before a stress report shows a coherent signature of constrained unpredictability — a coordinated pattern of narrowed operating range, elevated within-band entropy, and longer dwelling in particular states. That this signature is detectable in 5–15 second bpm data, without recourse to inter-beat intervals or the HRV constructs they support, lends support to the central commitment in Section 1 that consumer-wearable cardiac data, analyzed on its own terms, could carry information about regulatory state.

4.3 Implications for stress in policing

Policing arguably sits near the top of any list of high-stress occupations. Persistent and unaddressed stress is well-known to be detrimental to health and cognitive functioning (Andersen et al., 2023; McEwen, 2017), so it is unsurprising that law enforcement reports above-average rates of burnout and work dissatisfaction, or that stress plays a pronounced role in the poor decision-making outcomes that occasionally make news headlines (Hope, 2016; Nieuwenhuys & Oudejans, 2011). Although our participants were not yet operational officers, they were trainees in a police academy experiencing a wide range of actual and simulated police work over compressed weeks of training, under continuous instructor scrutiny and in competition with each other. The training environment closely simulated what officers do (office and classroom work, social interaction with peers and the public, and high-intensity scenarios demanding decisions under duress), and was correspondingly fraught with persistent pressure and acute stressors. In a setting where stress reactivity has direct consequences for both

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

performance and long-term well-being, an anticipatory window of even modest predictive accuracy opens a class of intervention designs that downstream detection alone cannot. The academy also provided a feasibility test of whether continuous monitoring and in-situ prompting can be sustained in a demanding occupational training environment, rather than lower-friction everyday settings where most ambulatory work has been done. Our study indicated that trainees are willing and able to integrate stress monitoring tools into their daily routine, especially with support from police leadership and in a cohort model where officers are collectively (i.e., with their peers) following study protocols. Our study also underscored the benefits of adapting a lightweight, durable wrist-worn device that could integrate with a uniform and daily activities, rather than requiring a separate device (i.e., smartphone) for data input.

4.4 Limitations

Several limitations qualify what these findings can and cannot support. The effect sizes are small in absolute terms, even at the most informative window and after controlling for individual differences. This level of prediction is meaningful for a naturalistic design, but it does not support strong individual-level claims without replication, personalization, and prospective validation.

Our sample also limits generalization. We studied 33 trainees across five training cohorts at a single academy. Participation was voluntary, and therefore not all trainees in each cohort participated. While the trainee population faces substantial stressors, it is not interchangeable with operational officers, who face field conditions, repeated trauma exposure, and longer-term occupational strain. Whether these patterns extend to working law enforcement is an empirical question this study cannot answer.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

A second set of limitations concerns measurement and design. Stress itself is a complex construct; subjective awareness and affective experience vary widely, as do their associations with physiology. Simply asking participants to provide a single rating of perceived stress is an important starting point but does not capture the wealth of individual and contextual information that shapes the experience. As with other experience sampling studies reporting predominantly low negative affective in daily life (von Klipstein et al., 2023), the bulk of responses – even among police trainees – indicated "no stress", and secondarily "a little". We dichotomized the stress outcome from a 1–4 Likert response, a choice motivated by substantial class imbalance in the data and by the analytic complications of ordinal mixed-effects modeling, but one that collapses informative gradations within the rated experience. Beyond measurement, our two-stage analytic strategy reuses the same data for feature selection (Stage 1) and effect estimation (Stage 2). We addressed this with strict cross-validated selection criteria in Stage 1 and by separating the two stages transparently in our reporting, but a fully held-out replication would provide a stronger basis for the effect estimates.

Finally, our design cannot disambiguate two compatible readings of the constrained-unpredictability pattern: it may reflect genuinely anticipatory regulation before stress crosses into awareness, or it may reflect sustained strain that the participant has not yet reported because the psychological threshold has not yet been triggered. Distinguishing these accounts may require pairing physiological monitoring with contextual event coding, such as scenario exercises, evaluations, physical training blocks, or other externally identifiable periods of likely demand. This would allow the same backward-looking analysis to be applied not only to subjective stress reports, but also to occupationally meaningful events whose timing is known independently of self-report.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

4.5 Future directions

Several specific extensions follow from these findings. Most directly, the 90-minute anticipatory window opens a design space for just-in-time interventions paired with the physiologically-triggered prompting we used here. A wearable that detects the constrained-unpredictability signature could deliver a brief, low-cost intervention, such as a paced-breathing prompt or short attentional reset, during a pre-report window, and test whether that support reduces later stress reports or improves performance during subsequent demand.

A second direction follows from the gap between the marginal and conditional R^2 values we report. Most of the explainable variance in stress reports lie in stable between-person differences rather than in the population-level cardiac features. Person-specific models, which estimate within-individual feature thresholds and importance weights rather than a single set across the sample, may recover substantially more of this variance and would also align the analytic unit with the practical unit of any deployed monitoring system.

A third direction is cross-device validation. The framework is in principle agnostic to manufacturer-specific processing pipelines, but Fitbit's particular smoothing and sampling decisions shape the signal in ways that may or may not generalize to Apple Watch, Garmin, or Polar devices. Replication across processing pipelines would help establish where the framework's claims are about cardiac dynamics in general and where they are tied to one device type.

Taken together, these results suggest that coarse beats-per-minute data, when analyzed through a framework suited to its temporal resolution, carry detectable anticipatory information about subjective stress. The gap between consumer wearables and traditional psychophysiology need not be a limitation to stall inquiry; it is a space for principled descriptive science, in which

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

we match empirical questions and analytic methods to the signal that is actually available. The present work takes a first step into that by showing that, even in coarse consumer-wearable data, the signal is there to find.

Acknowledgments

This research builds upon the Master's theses of Tara Paranjpe (published in ProQuest, 2023) and Sang-Hun Sim (published in ProQuest, 2022) at Arizona State University, presented and published as a conference proceeding: Sim, S., Paranjpe, T., Roberts, N., & Zhao, M. (2022, January). Exploring edge machine learning-based stress prediction using wearable devices. In 21st *IEEE International Conference on Machine Learning and Applications* (ICMLA). We thank the Phoenix Police Department (PPD) and Phoenix Police Regional Training Academy, including participating officers, trainees, staff, and leadership. This work was supported by the Arizona State University Institute for Social Science Research (ISSR) and the Arizona State University School of Social and Behavioral Sciences.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

The authors used Claude (Anthropic) and ChatGPT (OpenAI) while preparing this manuscript. Claude helped implement and debug analysis code that the lead author designed and specified. ChatGPT was used to check citation accuracy and identify candidate references; every source was independently retrieved, read, and evaluated by the authors before inclusion, and all AI-assisted code was reviewed, tested, and executed by the lead author. AI tools were not used to choose the analytic strategy, design or interpret analyses, generate data, or develop the manuscript's arguments or conclusions. The authors take full responsibility for the content of the manuscript.

References

Aldao, A., Sheppes, G., & Gross, J. J. (2015). Emotion Regulation Flexibility. *Cognitive Therapy and Research*, *39*(3), 263–278. https://doi.org/10.1007/s10608-014-9662-4

Andersen, J. P. (2015). Mental Preparedness as a Pathway to Police Resilience and Optimal Functioning in the Line of Duty. *International Journal of Emergency Mental Health and Human Resilience*, *17*(03). https://doi.org/10.4172/1522-4821.1000243

Andersen, J. P., Di Nota, P. M., Alavi, N., Anderson, G., Bennell, C., McGregor, C., Ricciardelli, R., Scott, S. C., Shipley, P., & Vincent, M. L. (2023). A Biological Approach to Building Resilience and Wellness Capacity Among Police Exposed to Posttraumatic Stress Injuries: Protocol for a Randomized Controlled Trial. *JMIR Research Protocols*, *12*(1), e33492. https://doi.org/10.2196/33492

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Anderson, G. S., Litzenberger, R., & Plecas, D. (2002). Physical evidence of police officer stress. *Policing: An International Journal of Police Strategies & Management, 25*(2), 399-420. https://doi.org/10.1108/13639510210429437

Appelhans, B. M., & Luecken, L. J. (2006). Heart Rate Variability as an Index of Regulated Emotional Responding. *Review of General Psychology*, *10*(3), 229–240. https://doi.org/10.1037/1089-2680.10.3.229

Bandt, C., & Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, *88*(17), 174102. https://doi.org/10.1103/PhysRevLett.88.174102

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beauchaine, T. P., & Thayer, J. F. (2015). Heart rate variability as a transdiagnostic biomarker of psychopathology. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *98*, 338–350. https://doi.org/10.1016/j.ijpsycho.2015.08.004

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berntson, G. G., Thomas Bigger Jr., J., Eckberg, D. L., Grossman, P., Kaufmann, P. G., Malik, M., Nagaraja, H. N., Porges, S. W., Saul, J. P., Stone, P. H., & Van Der Molen, M. W. (1997). Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, *34*(6), 623–648. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Bonanno, G. A., & Burton, C. L. (2013). Regulatory Flexibility: An Individual Differences Perspective on Coping and Emotion Regulation. *Perspectives on Psychological Science*, *8*(6), 591–612. https://doi.org/10.1177/1745691613504116

Brown, S. B. R. E., Brosschot, J. F., Versluis, A., Thayer, J. F., & Verkuil, B. (2018). New methods to optimally detect episodes of non-metabolic heart rate variability reduction as an indicator of psychological stress in everyday life. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *131*, 30–36. https://doi.org/10.1016/J.IJPSYCHO.2017.10.007

Can, Y. S., Chalabianloo, N., Ekiz, D., & Ersoy, C. (2019). Continuous stress detection using wearable sensors in real life: Algorithmic programming contest case study. *Sensors*, *19*(8), 1849. https://doi.org/10.3390/s19081849

Craddock, T. B., & Telesco, G. (2022). Police stress and deleterious outcomes: Efforts towards improving police mental health. *Journal of Police and Criminal Psychology*, *37*, 173–182.

de Vries, H., Kamphuis, W., van der Schans, C., Sanderman, R., & Oldenhuis, H. (2022). Trends in Daily Heart Rate Variability Fluctuations Are Associated with Longitudinal Changes in Stress and Somatisation in Police Officers. *Healthcare*, *10*(1), 144. https://doi.org/10.3390/healthcare10010144

Dell'Aversana, G., Herold, M., Simbula, S., Ruina, I., Pillan, M., & Miglioretti, M. (2026). Wearable devices and ecological momentary assessment EMA in the workplace: A study protocol on work stress assessment. *PLOS One*, *21*(1), e0339288. https://doi.org/10.1371/journal.pone.0339288

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, *49*, 146–169. https://doi.org/10.1016/j.yfrne.2018.03.001

Ferreira, S., Rodrigues, F., Kallio, J., Coelho, F., Kyllonen, V., Rocha, N., Rodrigues, M. A., & Vildjiounaite, E. (2024). From Controlled to Chaotic: Disparities in Laboratory vs Real-World Stress Detection. *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, 1–7. https://doi.org/10.1109/CBMI62980.2024.10859254

Fitabase. (n.d., n.d.). *Fitbit Research Library*. Fitabase. https://www.fitabase.com/research-library/

Fitbit, LLC. (2022). *Get heart rate intraday by date range*. Fitbit Web API documentation. https://dev.fitbit.com/build/reference/web-api/intraday/get-heartrate-intraday-by-date-range/

Folkman, S. (2013). Stress: Appraisal and Coping. In *Encyclopedia of Behavioral Medicine* (pp. 1913–1915). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-1005-9_215

Furberg, R. D., Taniguchi, T., Aagaard, B., Ortiz, A. M., Hegarty-Craver, M., Gilchrist, K. H., & Ridenour, T. A. (2017). Biometrics and policing: A protocol for multichannel sensor data collection and exploratory analysis of contextualized psychophysiological response during law enforcement operations. *JMIR Research*, *6*(3). https://doi.org/10.2196/resprot.7499

Goldberger, A. L., Amaral, L. A. N., Hausdorff, J. M., Ivanov, P. C., Peng, C., & Stanley, H. E. (2002). Fractal dynamics in physiology : Alterations with disease and aging. *PNAS*, *99*, 2466–2472. https://doi.org/10.1073/pnas.012579499

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Goodison, S. E., Barnum, J. D., Vermeer, M. J. D., Woods, D., Sitar, S. I., Shelton, S. R., & Jackson, B. A. (2020). *Wearable Sensor Technology and Potential Uses Within Law Enforcement: Identifying High-Priority Needs to Improve Officer Safety, Health, and Wellness Using Wearable Sensor Technology*. https://www.rand.org/pubs/research_reports/RRA108-7.html

Guy, M. E., Newman, M. A., & Mastracci, S. H. (2014). *Emotional Labor: Putting the Service in Public Service*. Routledge. https://doi.org/10.4324/9781315704852

Hardstone, R., Poil, S.-S., Schiavone, G., Jansen, R., Nikulin, V. V., Mansvelder, H. D., & Linkenkaer-Hansen, K. (2012). Detrended fluctuation analysis: A scale-free view on neuronal oscillations. *Frontiers in Physiology*, *3*, 450. https://doi.org/10.3389/fphys.2012.00450

Hernando, D., Roca, S., Sancho, J., Alesanco, Á., & Bailón, R. (2018). Validation of the Apple Watch for Heart Rate Variability Measurements during Relax and Mental Stress in Healthy Subjects. *Sensors (Basel, Switzerland)*, *18*(8), 2619. https://doi.org/10.3390/s18082619

Hoemann, K., Khan, Z., Feldman, M. J., Nielson, C., Devlin, M., Dy, J., Barrett, L. F., Wormwood, J. B., & Quigley, K. S. (2020). Context-aware experience sampling reveals the scale of variation in affective experience. *Scientific Reports*, *10*(1), 1–16. https://doi.org/10.1038/s41598-020-69180-y

Hope, L. (2016). Evaluating the effects of stress and fatigue on police officer response and recall: A challenge for research, training, practice and policy. *Journal of Applied Research in Memory and Cognition*, *5*(3), 239–245. https://doi.org/10.1016/j.jarmac.2016.07.008

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Ivanov, P. C., Amaral, L. A. N., Goldberger, A. L., Havlin, S., Rosenblum, M. G., Struzik, Z., & Stanley, H. E. (1999). Multifractality in Human Heartbeat Dynamics. *Nature*, *399*(6735), 461–465. https://doi.org/10.1038/20924

Jetelina, K. K., Molsberry, R., Malthaner, L., Beauchamp, A., Cannell, M. B., Hall, T., Fowler, E., & Anderson, L. (2022). Acceptability of a real-time notification of stress and access to self-help therapies among law enforcement officers. *BMC Public Health*, *22*(32). https://doi.org/10.1186/s12889-021-12423-y

Kim, K. B., & Baek, H. J. (2023). Photoplethysmography in Wearable Devices: A Comprehensive Review of Technological Advances, Current Challenges, and Future Directions. *Electronics*, *12*(13), 2923. https://doi.org/10.3390/ELECTRONICS12132923

Kinoshita, S., Hanashiro, S., Tsutsumi, S., Shiga, K., Kitazawa, M., Wada, Y., Inaishi, J., Kashiwagi, K., Fukami, T., Mashimo, Y., Minato, K., & Kishimoto, T. (2024). Assessment of Stress and Well-Being of Japanese Employees Using Wearable Devices for Sleep Monitoring Combined With Ecological Momentary Assessment: Pilot Observational Study. *JMIR Formative Research*, *8*, e49396. https://doi.org/10.2196/49396

von Klipstein, L., Servaas, M. N., Lamers, F., Schoevers, R. A., Wardenaar, K. J., & Riese, H. (2023). Increased affective reactivity among depressed individuals can be explained by floor effects: An experience sampling study. *Journal of Affective Disorders*, *334*, 370–381. https://doi.org/10.1016/j.jad.2023.04.118

Laborde, S., Mosley, E., & Thayer, J. F. (2017). Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Analysis, and Data Reporting. *Frontiers in Psychology*, *08*(213), 8. https://doi.org/10.3389/fpsyg.2017.00213

Lin, R., Bartels, S. L., Su, J. J., Cho, Y., Mace, R. A., & Heffner, K. L. (2026). Smartphone-based ecological momentary assessment for mild cognitive impairment: A feasibility study of measuring mood and stress responses in everyday setting. *Digital Health*, *12*, Article 20552076261444579. https://doi.org/10.1177/20552076261444579

Lipsitz, L., & Goldberger, A. (1992). Loss of 'Complexity' and Aging Potential Applications of Fractals and Chaos Theory to Senescence. *JAMA*. https://doi.org/10.1001/jama.1992.03480130122036

Lumley, M. A., Shi, W., Wiholm, C., Slatcher, R. B., Sandmark, H., Wang, S., Hytter, A., & Arnetz, B. B. (2014). The relationship of chronic and momentary work stress to cardiac reactivity in female managers: Feasibility of a smart phone-assisted assessment system. *Psychosomatic Medicine*, *76*(7), 512–518. https://doi.org/10.1097/PSY.0000000000000085

Martinez, G. J., Grover, T., Mattingly, S. M., Mark, G., D'Mello, S., Aledavood, T., Akbar, F., Robles-Granda, P., & Striegel, A. (2022). Alignment Between Heart Rate Variability From Fitness Trackers and Perceived Stress: Perspectives From a Large-Scale In Situ Longitudinal Study of Information Workers. *JMIR Human Factors*, *9*(3), e33754. https://doi.org/10.2196/33754

Marwan, N., Carmen Romano, M., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, *438*(5–6), 237–329. https://doi.org/10.1016/J.PHYSREP.2006.11.001

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

McEwen, B. S. (2017). Neurobiological and Systemic Effects of Chronic Stress. *Chronic Stress (Thousand Oaks, Calif.)*, *1*, 2470547017692328. https://doi.org/10.1177/2470547017692328

Melillo, P., Bracale, M., & Pecchia, L. (2011). Nonlinear Heart Rate Variability features for real-life stress detection. Case study: Students under stress due to university examination. *BioMedical Engineering OnLine*, *10*(1), 96. https://doi.org/10.1186/1475-925X-10-96

Gonzalez, J. M. R., Jetelina, K. K., Bishopp, S. A., Livingston, M. D., Perez, R. A. & Gabriel, K. P. (2019). The feasibility of using real-time, objective measurements of physiological stress among law enforcement officers in Dallas, Texas. *Policing: An International Journal, 42*(4), 701-710. https://doi.org/10.1108/PIJPSM-12-2018-0184.

Nelson, B. W., Low, C. A., Jacobson, N., Areán, P., Torous, J., & Allen, N. B. (2020). Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *Npj Digital Medicine*, *3*(1), 90. https://doi.org/10.1038/s41746-020-0297-4

Nieuwenhuys, A., & Oudejans, R. R. D. (2011). Training with anxiety: Short- and long-term effects on police officers' shooting behavior under pressure. *Cognitive Processing*, *12*(3), 277–288. https://doi.org/10.1007/s10339-011-0396-x

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Quality & Quantity*, *41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Pinge, A., Gad, V., Jaisighani, D., Ghosh, S., & Sen, S. (2024). Detection and monitoring of stress using wearables: A systematic review. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1478851

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Porges, S. W. (2011). *The polyvagal theory: Neurophysiological foundations of emotions, attachment, communication, and self-regulation* (pp. xvii, 347). W. W. Norton & Company.

Radin, J. M., Quer, G., Ramos, E., Baca-Motes, K., Gadaleta, M., Topol, E. J., & Steinhubl, S. R. (2021). Assessment of Prolonged Physiological and Behavioral Changes Associated With COVID-19 Infection. *JAMA network open*, *4*(7), e2115959. https://doi.org/10.1001/jamanetworkopen.2021.15959

Roos, L. G., & Slavich, G. M. (2023). Wearable technologies for health research: Opportunities, limitations, and practical and conceptual considerations. *Brain, Behavior, and Immunity*, *113*, 444–452. https://doi.org/10.1016/j.bbi.2023.08.008

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability? *International Journal of Cardiology*, *166*(1), 15–29. https://doi.org/10.1016/j.ijcard.2012.03.119

Schwerdtfeger, A. R., & Rominger, C. (2021). Feelings from the heart: Developing HRV decrease-trigger algorithms via multilevel hyperplane simulation to detect psychosocially meaningful episodes in everyday life. *Psychophysiology*, *58*(11), e13914. https://doi.org/10.1111/PSYP.13914

Sheikh, M., Qassem, M., & Kyriacou, P. A. (2021). Wearable, Environmental, and Smartphone-Based Passive Sensing for Mental Health Monitoring. *Frontiers in Digital Health*, *3*. https://doi.org/10.3389/fdgth.2021.662811

Singletary, G. (2024). Law enforcement and mental health: The limpid kryptonite—a clear and present danger. *Journal of Police and Criminal Psychology*, *39*, 55–63.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Smets, E., Rios Velazquez, E., Schiavone, G., Chakroun, I., D'Hondt, E., De Raedt, W., Cornelis, J., Janssens, O., Van Hoecke, S., Claes, S., Van Diest, I., & Van Hoof, C. (2018). Large-scale wearable data reveal digital phenotypes for daily-life stress detection. *NPJ Digital Medicine*, *1*, 67. https://doi.org/10.1038/s41746-018-0074-9

Smith, R., Thayer, J. F., Khalsa, S. S., & Lane, R. D. (2017). The hierarchical basis of neurovisceral integration. *Neuroscience & Biobehavioral Reviews*, *75*, 274–296. https://doi.org/10.1016/j.neubiorev.2017.02.003

Specker, P., Sheppes, G., & Nickerson, A. (2024). Does emotion regulation flexibility work? Investigating the effectiveness of regulatory selection flexibility in managing negative affect. *Social Psychological and Personality Science*, *15*(5), 561-569.

Sterling, P. (2012). Allostasis: A model of predictive regulation. *Physiology & Behavior*, *106*(1), 5–15. https://doi.org/10.1016/j.physbeh.2011.06.004

Thayer, J. F., & Lane, R. D. (2000). A model of neurovisceral integration in emotion regulation and dysregulation. *Journal of Affective Disorders*, *61*(3), 201–216. https://doi.org/10.1016/S0165-0327(00)00338-4

Tutunji, R., Kogias, N., Kapteijns, B., Krentz, M., Krause, F., Vassena, E., & Hermans, E. J. (2023). Detecting Prolonged Stress in Real Life Using Wearable Biosensors and Ecological Momentary Assessments: Naturalistic Experimental Study. *Journal of Medical Internet Research*, *25*, e39995. https://doi.org/10.2196/39995

Violanti, J. M. (2007). Homicide-suicide in police families: Aggression full circle. *International Journal of Emergency Mental Health*, *9*(2), 97–104. https://www.ncbi.nlm.nih.gov/pubmed/17725078

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Dynamics of cardiac organization: A flexibility-structure framework for detecting anticipatory stress from consumer-grade heart-rate data in police trainees

Nicholas D. Duran[1,†], Nicole A. Roberts[2,†], Tara Paranjpe[3], Sang-Hun Sim[3], Michael S. Scott[4], and Ming Zhao[3]

[1] Department of Psychology, The University of Texas at Austin, Austin, TX, USA

[2] School of Social and Behavioral Sciences, Arizona State University, Glendale, AZ, USA

[3] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA

[4] School of Criminology and Criminal Justice, Arizona State University, Tempe, AZ, USA

† These authors contributed equally to this work and share first authorship.

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109

Author Note

Nicholas D. Duran and Nicole A. Roberts contributed equally to this work and share first authorship. Nicholas D. Duran is listed first and serves as corresponding author.

We have no known conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Nicholas D. Duran, Department of Psychology, The University of Texas at Austin, 108 East Dean Keeton Street, SEA 4.208, Austin, TX 78712-1043, United States. Email: nicholas.duran@austin.utexas.edu

Preprint not peer reviewed

This preprint research paper has not been peer reviewed. Electronic copy available at: https://ssrn.com/abstract=7478109