

Review Article

Cailbhe Doherty*, Maximus Baldwin, Rory Lambe, David Burke and Marco Altini

Readiness, recovery, and strain: an evaluation of composite health scores in consumer wearables

<https://doi.org/10.1515/teb-2025-0001>

Received January 30, 2025; accepted March 28, 2025;
published online April 10, 2025

Abstract

Introduction: Consumer wearables increasingly provide users with Composite Health Scores (CHS) – integrated biometric indices that claim to quantify readiness, recovery, stress, or overall well-being. Despite their growing adoption, the validity, transparency, and physiological relevance of these scores remain unclear. This study systematically evaluates CHS from leading wearable manufacturers to assess their underlying methodologies, contributors, and scientific basis.

Content: Information was synthesised from publicly available company documentation, including technical white papers, user manuals, app interfaces, and research literature where available. We identified 14 CHS across 10 major wearable manufacturers, including Fitbit (Daily Readiness), Garmin (Body Battery™ and Training Readiness), Oura (Readiness and Resilience), WHOOP (Strain, Recovery, and Stress Monitor), Polar (Nightly Recharge™), Samsung (Energy Score), Suunto (Body Resources), Ultrahuman (Dynamic Recovery), Coros (Daily Stress), and Withings (Health Improvement Score). The most frequently incorporated biometric contributors in this

catalogue of CHS were heart rate variability (86 %), resting heart rate (79 %), physical activity (71 %), and sleep duration (71 %). However, significant discrepancies were identified in data collection timeframes, metric weighting, and proprietary scoring methodologies. None of the manufacturers disclosed their exact algorithmic formulas, and few provided empirical validation or peer-reviewed evidence supporting the accuracy or clinical relevance of their scores.

Summary and outlook: While the concept of CHS represent a promising innovation in digital health, their scientific validity, transparency, and clinical applicability remain uncertain. Future research should focus on establishing standardized sensor fusion frameworks, improving algorithmic transparency, and evaluating CHS across diverse populations. Greater collaboration between industry, researchers, and clinicians is essential to ensure these indices serve as meaningful health metrics rather than opaque consumer tools.

Keywords: wearable technology; digital health; heart rate variability; consumer wearables

Introduction

The release of the Fitbit Classic in 2009 was a watershed moment in the adoption of self-tracking technologies, catalysing a shift from niche interest to widespread mainstream acceptance [1–4]. Initially embraced by a small community of enthusiasts [5], wearable devices have since become a central part of many people's lives, enabling users to monitor their physical activity, sleep, and overall health [4, 6]. This shift reflects broader societal trends toward personalization and data-driven insights in health and wellness management. In 2024, over half the population in many countries owned a wearable device, with projections suggesting global smartwatch users will exceed 740 million by 2029 [7]. This growth underscores the increasing demand for accessible, real-time health monitoring and the growing role wearables play in bridging the gap between individual health behaviours and data-driven insights [8, 9].

*Corresponding author: Cailbhe Doherty, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland; and Insight Research Ireland Centre for Data Analytics, University College Dublin, Dublin, Ireland, E-mail: cailbhe.doherty@ucd.ie. <https://orcid.org/0000-0002-5284-856X>

Maximus Baldwin, Insight Research Ireland Centre for Data Analytics, University College Dublin, Dublin, Ireland; and Institute for Sport and Health, University College Dublin, Dublin, Ireland. <https://orcid.org/0000-0002-0390-3956>

Rory Lambe, School of Public Health, Physiotherapy and Sports Science, University College Dublin, Dublin, Ireland; and Insight Research Ireland Centre for Data Analytics, University College Dublin, Dublin, Ireland. <https://orcid.org/0009-0009-5866-361X>

David Burke, Cardiology, Beacon Hospital, Dublin, Ireland; and School of Medicine, University College Dublin, Dublin, Ireland

Marco Altini, Department of Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, Netherlands. <https://orcid.org/0000-0001-5981-4066>

The early iterations of consumer wearables relied on relatively simple sensing technologies, such as the single triaxial accelerometer embedded in the Fitbit Classic, which was used to estimate physical activity (steps taken), energy expenditure, and sleep patterns [10]. These early devices provided basic insights that encouraged users to engage with their health in novel ways [11]. However, in the ensuing decade, the capabilities of modern wearables advanced dramatically, integrating multi-sensor fusion techniques to capture an extensive array of physiological data with greater accuracy and resolution [12]. For instance, international measurement units (IMUs) containing triaxial accelerometers and gyroscopes allow for the detailed measurement of spatiotemporal and biomechanical outcomes, facilitating insights into gait and movement patterns [13, 14]. Bioimpedance sensors provide estimations of body composition, including metrics like body fat percentage and hydration levels [15], while photoplethysmography (PPG) sensors enable continuous tracking of heart rate, respiratory rate, and blood oxygen saturation [16]. Electrocardiogram (ECG) sensors are now a standard feature in many devices, enabling real-time monitoring of cardiac rhythm [17].

These technological advancements have broadened the applications of wearables across research, clinical, personal health, and sporting contexts. In research, wearables are increasingly used to collect longitudinal health data at scale, enabling the study of patterns and trends in diverse populations [18–21]. Clinically, they provide opportunities for remote patient monitoring [22–24], predicting patient outcomes [22, 25–27] and depression [28], and personalising treatment strategies [29, 30]. For individual users, modern wearables offer insights into relevant health metrics, facilitating changes in behaviour that can lead to increased physical activity [31], improved fitness [31, 32] and wellness [33, 34]. In sporting contexts, wearables have been used to monitor player movements [35], workloads [36], and biometric markers [37, 38].

The implementation of consumer wearables in research contexts is increasingly being founded on well-established theoretical principles, with best-practice methodologies to ensure the robustness of findings [39–42]. However, an emerging category of wearable derived biometric outcomes has received comparatively less scientific scrutiny: Composite Health Scores (CHS) [43]. Unlike traditional biometric measures that directly quantify physiological parameters such as heart rate or respiratory rate, CHS integrate multiple data streams using signal fusion techniques to generate composite indices [43]. These indices are designed to provide users with a simplified yet comprehensive assessment of their overall health, recovery, or readiness for activity. By distilling complex physiological data into an intuitive

format, CHS aim to enhance user engagement and facilitate health-related decision-making.

CHS have become a prominent feature in several consumer wearables. For example, Fitbit's Daily Readiness score integrates sleep patterns, resting heart rate and heart rate variability to assess physical preparedness [44]. Garmin's Body Battery™ combines stress, recovery, and physical activity data to estimate energy reserves [45], while Oura's Readiness Score leverages sleep quality, heart rate variability, and activity metrics to provide insights into recovery [46]. Similarly, WHOOP's Strain metric focuses on cardiovascular exertion, ostensibly helping users optimize performance and avoid overtraining [47]. While these indices are marketed as user-friendly tools for understanding health and performance, important gaps remain in our understanding of their contributors, calculations, and scientific validation. The algorithms underpinning these indices are often proprietary, limiting transparency regarding which metrics are prioritized, how they are weighted, and whether the resulting scores align with physiological realities [43]. Furthermore, it is unclear whether these indices have undergone rigorous validation in diverse populations or under varying physiological conditions, raising concerns about their generalisability and reliability [43]. Despite these limitations, CHS are now a common feature in consumer wearables, and many users rely on them to guide their training and lifestyle decisions [48].

This review seeks to address these knowledge gaps by systematically evaluating CHS in consumer wearables. Specifically, it aims to (1) identify and categorize the biometric and algorithmic components underpinning these scores, detailing the physiological signals they incorporate; (2) assess transparency regarding their methodologies, including disclosure of calculation processes, metric weightings, and underlying assumptions; and (3) analyse their intended purpose and utility, evaluating how manufacturers present these scores to users and whether their implied applications align with established physiological principles. The summary of this article is presented in Figure 1.

Content

Study design & integrated health index selection

This study employed a secondary analysis of data extracted and synthesized as part of a previously published living

Readiness, recovery, and strain: an evaluation of composite health scores in consumer wearables

INTRODUCTION

Consumer wearables are increasingly used to monitor **health, recovery, and exercise readiness**.



However, the **proprietary nature and lack of validation** of Composite Health Scores (CHS) limit their **scientific and clinical utility**.



Objective: to systematically evaluate CHS from leading wearable manufacturers to assess their underlying methodologies, contributors, and scientific basis.



METHODS

1



This review systematically evaluated 14 CHS across 10 leading wearable manufacturers using technical documentation, app data, and scientific literature.



10 wearables

14 CHS

2



Each CHS such as readiness, body battery, resilience, recovery, and stress, was analyzed for its **biometric contributors, calculation methodology, transparency, and evidence of validation**.

RESULTS

CHS consistently relied on metrics like **heart rate variability** and **resting heart rate**, but **calculation methods varied widely**.



None of the **CHS disclosed** their full **algorithms** hindering **reproducibility** and **independent scrutiny**.



The **absence of rigorous validation** across **diverse populations** raises critical concerns about **generalizability and equity**.



CONCLUSION

CHS have potential in digital health and exercise recovery but require standardization and scientific validation to ensure reliability and clinical relevance.



TRANSLATIONAL EXERCISE BIOMEDICINE - Doherty et al., 2025



Figure 1: Graphical representation of this study. Key points: (1) composite health scores (CHS) are widely used in consumer wearables but lack transparency and empirical validation; (2) this review systematically evaluated 14 CHS across 10 manufacturers based on contributors, methodology, and validation evidence; and (3) findings highlight major inconsistencies in metric integration, limited reproducibility, and the need for standardization. *Figure created with BioRender.*

umbrella review evaluating the accuracy of consumer wearable devices [43] – primary research studies identified as part of the living umbrella review were parsed for any reference to a CHS. This was supplemented by an industry report identifying major wearable device manufacturers [7], which ranked leading manufacturers of wearable devices

based on global market share. From this list, the top 10 consumer wearable device manufacturers were identified, representing key stakeholders in the wearable technology sector with substantial user bases, and their product documentation was screened for any reference to a CHS as defined below.

Inclusion and exclusion criteria

Having first identified the most widely used consumer wearable devices in the research literature [43] and the companies with the largest market share [7], next, we collated a list of the different indices. For the purposes of this analysis, a CHS was defined as a composite measure derived from the integration of multiple biometric signals, such as heart rate variability, sleep metrics, physical activity, and body temperature, into a single score that reflects general health, recovery, or readiness for daily functioning.

Our inclusion criteria were as follows: 1) eligible CHS were required to combine multiple biometric signals, such as sleep metrics, heart rate variability, physical activity, and stress, into a composite score reflecting general health or readiness; 2) devices or platforms included in this study were required to be commercially available, with sufficient documentation detailing the components of the CHS; this documentation could include technical white papers, app screenshots, user manuals, or research publications; and 3) the CHS were designed for general health monitoring rather than sport-specific or performance-focused applications, ensuring relevance to a broader population.

Exclusion criteria were as follows: 1) single-construct metrics that only incorporated one biometric contributor (e.g., heart rate variability) or indices which focused solely on sleep or physical activity in isolation; 2) metrics explicitly tied to training load or other sport-specific performance measures; 3) devices or platforms lacking any publicly available information relating to CHS calculation or its contributing metrics; and 4) research-grade devices, prototypes or algorithms exclusively intended for clinical use and unavailable to the general public.

Procedures for evaluation

The evaluation of CHS was conducted through a systematic and structured approach designed to assess their contributors, methodologies, validation, and applicability. Data for the evaluation were collected from publicly available sources, including technical white papers, patents, user manuals, app screenshots and device interfaces, as well as research studies and validation reports published by manufacturers or independent investigators. Each index was analysed for its intended purpose, the metrics it integrates, its calculation methodology, and any available evidence supporting its validity. Missing or proprietary information was explicitly documented, and the implications of such gaps for scientific and clinical use were synthesized.

Each index was first contextualized with an overview, detailing its official name, the device or platform offering the index, and its primary application, such as readiness assessment, recovery monitoring, or general health evaluation. The contributors to the index were then assessed, focusing on the biometric signals integrated into the calculation, such as heart rate variability, sleep quality, physical activity, and stress levels. For each contributor, the evaluation considered its physiological relevance, the methods used for data acquisition and processing, and any thresholds or ranges associated with health outcomes, where available.

The calculation methodology of each index was then examined to understand how contributors were integrated into a composite score. This included an assessment of whether specific weightings or algorithms were disclosed and whether the indices incorporated short-term, long-term, or combined metrics. The scoring range and any interpretive categories, such as readiness levels or energy states, were documented and evaluated for their clarity and practical utility.

Whether each index had undergone any internal or independent validation was then assessed, provided this information was publicly available. Available evidence supporting the validity of the indices was synthesized, including any independent validation studies and their applicability across diverse populations or physiological conditions. Practical applications were considered, with an emphasis on the index's utility for remote health monitoring, personalised health interventions, and general wellness management.

Finally, the transparency of the index – the extent to which it could be understood and replicated – was evaluated by reviewing the availability of technical documentation and the accessibility of detailed explanations regarding the construction and interpretation of each index. The degree to which manufacturers provided guidance to users, including recommendations based on the scores, was also examined. For indices with incomplete or proprietary data, the absence of key details, such as algorithmic weightings or validation data, was ascertained. The evaluation framework is summarised in Table 1.

Analysis and narrative synthesis

We used the evaluation framework to categorize the indices based on their primary purpose, the biometric contributors they integrated, and the methodologies employed for score calculation. A comparative approach was adopted to assess the consistency and variation in how indices synthesized physiological data into guidance or recommendations for

Table 1: The framework used to evaluate the composite health scores.

Domain	Description
Index overview	General information about the index being evaluated.
Name of index	Official name of the index.
Device/platform	Specific device or platform offering the index.
Purpose	Intended application of the index (e.g., readiness, recovery, health evaluation).
Contributors and metrics	Key biometric contributors to the index and their roles in the final score.
Calculation methodology	Details how biometric signals are combined to generate the index.
Integration of contributors	Describes the integration process, including weighting or algorithms, if available.
Time scales used	Indicates the type of metrics (short-term, long-term, or both) used in the calculation.
Score scale	Documents the score range and its interpretive categories.
Validation and relevance	Evidence supporting the validity and reliability of the index.
Transparency	Assesses transparency in the construction and documentation of the index.
Public documentation	Reviews the availability of technical documents like white papers or patents.
User guidance	Evaluates how the manufacturer communicates the index's purpose and use to users.

the user. Particular focus was placed on understanding the weighting and integration of individual contributors, such as heart rate variability, sleep metrics, and physical activity, within the composite scores. The scoring scales and their interpretive categories were examined to evaluate their clarity and practical implications for users.

Where information was missing or proprietary, its absence was explicitly documented, and the potential consequences for usability and scientific credibility were analysed. For example, indices with undisclosed algorithmic methodologies were flagged as presenting challenges for independent validation and reproducibility. Similarly, gaps in validation data, particularly in diverse populations or under varying physiological conditions, were noted as potential limitations in the generalizability of the indices.

Evaluation of composite health scores in consumer wearables

Overview

Fourteen CHS were identified across 10 major consumer wearable device manufacturers, including Coros (Daily

Stress) [49], Fitbit/Google (Daily Readiness) [44], Garmin (Body Battery™ and Training Readiness) [45, 50], Oura (Oura Readiness Score and Resilience) [46, 51], Polar (Nightly Recharge™) [52], Samsung (Energy Score) [53], Suunto (Body Resources) [54], Ultrahuman (Dynamic Recovery) [55], WHOOP (WHOOP Recovery, WHOOP Strain and the Stress Monitor) [47, 56, 57] and Withings (Health Improvement Score) [58].

These indices are broadly designed to aggregate wearable-derived biometric data streams into guidance or recommendations aimed at optimizing health, physical activity and mental performance. They variously provide guidance on recovery, exercise, and rest, supporting users in making lifestyle choices informed by biometric data related to the cardiovascular system, sleep, and physical activity. Personalised recommendations are a common feature of the indices, and are intended to foster healthier habits and improved long-term outcomes.

While each index integrates a variety of physiological and behavioural metrics, none of the companies disclose specific details about the algorithms or formulae used – nor the relative weightings of individual metrics within the overall score. The following sections expand upon the contributors, calculation methodologies, and reproducibility of these indices.

Contributors to composite health scores

The primary biometric signals incorporated in the CHS were heart rate (86 % of CHS), heart rate variability (86 % of CHS), resting heart rate (79 % of CHS), activity (including, but not limited to, accelerometry derived motion, step counts and energy expenditure estimates; 71 % of CHS), and sleep quantity (71 % of CHS). These were followed by sleep quality/architecture (57 % of CHS), body temperature (29 % of CHS) and respiratory rate (14 % of CHS). Blood oxygen saturation, body morphology, blood pressure and heart rhythm (ECG) each contributed once to the various CHS.

Thus, the two sensing modalities underpinning the greatest number of biometric signals are PPG – which is used to capture heart rate, heart rate variability, respiratory rate and blood oxygen saturation – and accelerometry, which is used to capture movement, enabling assessment of activity and sleep.

The sensing modalities and biometric signals incorporated in the identified CHS are summarised in Table 2.

Table 2: Comparison of metrics incorporated in composite health scores across consumer wearables.

Metric	Activity	Body morphology (e.g., weight, body fat %)	Blood oxygen	Blood pressure	ECG	HR	HRV	RHR	RR	Sleep quantity/ duration	Sleep quality or architecture, REM/NREM	Temperature
Coros Daily Stress					✓	✓	✓	✓	✓	✓	✓	17 %
Fitbit Daily Readiness					✓	✓	✓	✓	✓	✓	✓	33 %
Garmin Body	✓				✓	✓	✓	✓	✓	✓	✓	42 %
Battery™					✓	✓	✓	✓	✓	✓	✓	50 %
Garmin Training Readiness	✓				✓	✓	✓	✓	✓	✓	✓	50 %
Oura Readiness Score	✓				✓	✓	✓	✓	✓	✓	✓	58 %
Oura Resilience	✓				✓	✓	✓	✓	✓	✓	✓	58 %
Polar Nightly Recharge™					✓	✓	✓	✓	✓	✓	✓	50 %
Samsung Energy Score	✓				✓	✓	✓	✓	✓	✓	✓	42 %
Suunto Body Resources	✓				✓	✓	✓	✓	✓	✓	✓	50 %
Ultrahuman Dynamic Recovery	✓				✓	✓	✓	✓	✓	✓	✓	50 %
WHOOP Recovery					✓	✓	✓	✓	✓	✓	✓	17 %
WHOOP Strain	✓				✓	✓	✓	✓	✓	✓	✓	25 %
WHOOP Stress Monitor	✓				✓	✓	✓	✓	✓	✓	✓	67 %
Withings Health Improvement Score	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	29 %

ECG, electrocardiogram, a measure of heart rhythm; HRV, heart rate variability, a measure of autonomic nervous system balance; HR, heart rate, the number of heartbeats per minute; RHR, resting heart rate, the heart rate when the body is at complete rest; Activity, physical activity levels assessed via movement (accelerometry) and/or heart rate; Sleep, general sleep duration and quality; Sleep architecture (REM/NREM), analysis of sleep stages, including REM (rapid eye movement) and NREM (non-rapid eye movement); Temperature, body temperature variations monitored during rest or activity; RR, respiratory rate, the number of breaths per minute; Blood Oxygen, levels of oxygen saturation in the blood (SpO_2). Bottom row counts: Summarises the number of metrics incorporated in each CHS. Side row counts: summarizes the number of metrics incorporated into a given CHS, expressed as a percentage of the total number of metrics (12).

Comparative analysis

Calculation methodologies, timescales, and scoring

The calculation methodologies, timescales, and scoring approaches of the CHS vary across platforms but exhibit several common features. First, the integration of physiological and behavioural metrics is achieved through proprietary algorithms which dynamically combine inputs to generate composite scores. Metrics related to heart rate – primarily resting heart rate and heart rate variability (which is typically calculated using the root mean square of successive differences between normal heartbeats [RMSSD]) – are central to most calculations, seemingly carrying more weight than other inputs. These PPG-derived outcomes related to the cardiovascular system are variously combined with different measures of accelerometry-derived physical activity or ‘load’, to provide users with an overall sense of recovery or readiness to perform or train.

For instance, Garmin’s Training Readiness [50] and Oura’s Readiness Score [46] incorporate both short-term metrics (e.g., previous day’s sleep and activity) and long-term trends (e.g., 7-day heart rate variability averages or 14-day activity data). Similarly, Polar’s Nightly Recharge™ [52] integrates heart rate variability and sleep metrics calculated against a 28-day baseline, focusing on recovery insights over a medium-term horizon. Fitbit’s Daily Readiness [44] combines heart rate variability and resting heart rate with sleep data over a 7-day baseline to inform users about how much physical activity they should undertake that day, while WHOOP’s Strain [47] integrates cardiovascular (from PPG-derived heart rate) and muscular (from accelerometry-derived activity) loads, which are adjusted based on individual fitness levels and baseline data. WHOOP’s Recovery index [57] also integrates heart rate variability, resting heart rate, sleep quality, and respiratory rate to nominally provide a daily score that quantifies how prepared an individual’s body is to adapt to physical and mental stressors (later devices by WHOOP, including the WHOOP 4.0, also incorporates blood oxygen levels and skin temperature for enhanced sensitivity to illness). Withings’ Health Improvement Score combines heart rate, physical activity, body morphology and sleep data sub-scores to give users a “better understanding what you can focus on to improve your overall health score [58].”

Most indices do not disclose the precise weightings of their contributors. For example, Polar’s Nightly Recharge™ prioritizes “autonomic nervous system recovery metrics” (PPG derived heart rate and heart rate variability) over sleep metrics [52], while Fitbit’s Daily Readiness emphasizes heart rate variability and sleep more than resting heart rate [44].

Similarly, Garmin’s Body Battery™ [45] and Samsung’s Energy Score [53] highlight heart rate variability and stress (typically extrapolated from heart rate and activity) as primary contributors but provide no specific details regarding the exact interplay of these metrics.

The timescales used to calculate each score reflects both short-term data and long-term trends. For example, short-term data such as daily heart rate variability and the previous night’s sleep hold greater weighting in Ultrahuman’s Dynamic Recovery [55] and Suunto’s Body Resources [54] indices. Long-term trends, including rolling averages (e.g., 7-day heart rate variability in Garmin’s Training Readiness [50] or 28-day baselines in Polar’s Nightly Recharge™) [52], provide context for variations from an established ‘baseline’. WHOOP reports that their strain [47] and recovery [57] algorithms prioritize immediate physiological responses, reset scores with sleep cycles, and offer dynamic updates throughout the day.

All of the scores are normalized to interpretable scales, typically ranging from 0 to 100 (e.g., Fitbit Daily Readiness, Garmin Training Readiness, Ultrahuman Dynamic Recovery). WHOOP’s Strain Score uses a logarithmic scale of 0–21 [47], where higher strain values are progressively harder to achieve, whereas Polar’s Nightly Recharge™ [52] Score categorizes recovery into six levels on a scale from -10 to +10 based on “autonomic nervous system charge” – calculated through heart rate, heart rate variability (using RMSSD), and breathing rate during the early sleep period – and “sleep charge”. Interpretive categories such as “Optimal,” “Fair,” and “Pay Attention” (Oura Readiness Score) or “Prime,” “Moderate,” and “Poor” (Garmin Training Readiness) provide insights to help users adjust activity levels or prioritize recovery. These interpretive categories are often color-coded or accompanied by textual explanations, to aid usability and understanding.

Transparency and validation

The proprietary nature of these algorithms significantly limits transparency in how metrics are integrated and weighted, raising challenges for scientific evaluation and user trust. For example, Fitbit’s Daily Readiness Score [44] does not disclose the relative contributions of HRV, sleep, and resting heart rate, making it difficult for users or researchers to understand its rationale. Garmin’s Body Battery™ [45] provides user-facing guidance on metrics like stress and sleep but offers no detailed insights into its algorithmic interactions, restricting independent validation. Polar’s Nightly Recharge™ [52] and Oura’s Readiness Score [46] include some white papers and/or validation frameworks (theoretical, not empirical) for individual components

such as heart rate variability or sleep detection algorithms but lack comprehensive, independent assessments of their composite readiness scores. For example, Polar's Nightly Recharge™ white paper states "there is no gold standard method for assessing recovery [...] however, Nightly Recharge™ is based on up-to-date scientific knowledge of stress and recovery, and it utilizes generally accepted tools

Table 3: Approaches to heart rate variability (HRV) and resting heart rate (RHR) calculation across consumer wearables.

Company	HRV calculation	RHR calculation
Coros	rMSSD; manual – coros recommends "taking your HRV measurement between 4:00 and 10:00 in the morning" and provides a series of steps to do so.	Manual input or lowest value recorded during sleep
Fitbit	rMSSD; longest sleep period >3 h in past 24 h	During sleep or periods of wakeful inactivity with no steps detected
Garmin	rMSSD; continuous during sleep, averaged across entire sleep period with 5-min windows	Lowest 30-min average in a 24-h period
Oura	rMSSD; mean of all 5-min samples throughout sleep, also reports maximum HRV	Average and lowest values during sleep, sampled every 10 min
Polar	rMSSD; 4-h window starting 30 min post-sleep onset	Average over 4-h period starting 30 min post-sleep onset
Samsung	Unspecified; continuous monitoring available but unclear methodology	Measured throughout the day during inactive periods; methodology unspecified
Suunto	rMSSD; continuous monitoring during sleep	Manual input or lowest value recorded during sleep
Ultrahuman	rMSSD; filters out motion periods during sleep, trends emphasized over absolute values	Measured during sleep; no further details provided
WHOOP	rMSSD; weighted towards last slow-wave sleep period each night	Weighted towards last slow-wave sleep period each night
Withings	As of January 2025, Withings has not yet implemented HRV measurement in their devices.	Manual reading or average values during sleep.

HRV, heart rate variability; RHR, resting heart rate; rMSSD, root mean square of successive differences; PPG, photoplethysmography. Footnote: heart rate (HR) refers to the number of times the heart beats per minute (bpm) and can fluctuate widely based on physical activity, stress, and other factors throughout the day. In contrast, resting heart rate (RHR) represents the heart rate during periods of complete rest, typically measured during sleep or prolonged inactivity. Unlike HR, RHR is used as a baseline indicator of cardiovascular fitness and overall health, with lower values generally indicating better aerobic fitness and heart efficiency. For most composite health scores (CHSs), RHR is calculated using data from sleep or wakeful inactivity, filtered to exclude periods of movement or physiological stress.

for measuring autonomic nervous system functioning and nightly sleep in real-life settings" [59] – however, no citations to any primary research are provided. Similarly, Ultrahuman's Dynamic Recovery [55] leverages multiple inputs, including heart rate variability, resting heart rate, and "stress", but its proprietary algorithm also lacks published validation studies, limiting scientific scrutiny. This reflects a broad trend among CHS, where theoretical underpinnings are emphasized over empirical validation.

Finally, a number of CHS can only be accessed by paying an app subscription fee, including Oura (for their Resilience), WHOOP (WHOOP Recovery, WHOOP Strain and the Stress Monitor) and Withings (for their Health Improvement Score).

The different approaches employed by the companies to calculate resting heart rate and heart rate variability are displayed in Table 3.

The full comparison between the CHS is presented in Table 4.

Discussion

Overview of composite health scores and key contributors

This study sought to evaluate the design, contributors, and validity of consumer wearable CHS, aiming to clarify their construction and scientific basis. Our search of the scientific literature and of leading device manufacturers, including the available company documentation, identified 14 individual CHS: Coros' Daily Stress [49], Fitbit's Daily Readiness [44], Garmin's Body Battery™ [45] and Training Readiness [50], Oura's Readiness [46] and Resilience [51] Scores, Polar's Nightly Recharge™ [52], Samsung's Energy Score [53], Suunto's Body Resources [54], Ultrahuman's Dynamic Recovery [55], WHOOP's Recovery [57], Strain [47] and Stress Monitor [56] and Withings's Health Improvement Score [58]. Through a systematic analysis of the methodologies employed by these 10 manufacturers across their 14 indices, we identified consistent trends in how these tools integrate temporal biometric data streams – such as physical activity, cardiovascular function, and sleep – into a composite score.

Our findings reveal that CHS incorporate diverse physiological signals, including physical activity, body morphology, blood oxygen saturation, blood pressure, heart rhythm (ECG), heart rate, heart rate variability, resting heart rate, respiratory rate, sleep duration and quality, and body temperature. However, while all of the CHS ostensibly transform different variations of these complex physiological data into a discrete

Table 4: Comparison of composite health scores in consumer wearables.

Company	Name of index	Device/platform	Contributors	Calculation	Timescales used	Score scale
COROS	Daily Stress	Coros wearable devices	HR, HRV	Combines HR data (daytime) and HRV data (nighttime) into a 1–100 scale. Details on integration and weighting are proprietary.	Short-term (5 min); long-term (weekly, monthly averages)	1–100: Calm (1–25), low stress (26–50), medium stress (51–75), high stress (76–100)
Fitbit	Daily readiness score	Fitbit devices (e.g., charge, Sense, Versa) and Google Pixel watch series	Sleep (quantity, quality, and debt), HRV, RHR	Proprietary algorithm comparing contributors to personalised baselines. Uses trends over short-term (7 days) and long-term (1 month) periods.	Short-term (7 days), long-term (1 month)	1–100: Low (1–29), moderate (30–64), high (65–100)
Garmin	Body Battery™ energy monitoring	Garmin wearable devices (e.g., Fore-runner, Fenix)	Stress (HR, RHR HRV), activity levels, sleep	Proprietary algorithm (powered by First-beat Analytics™) dynamically integrates HR, HRV, stress, activity, and sleep. Weights and specific interactions not disclosed. Recent data weighted more heavily.	Short-term (daily activity); long-term (sleep and recovery trends)	0–100: Higher scores indicate optimal energy levels and readiness for activity.
Garmin	Training readiness	Garmin wearable devices (e.g., Fore-runner, Fenix)	Sleep score, recovery time, acute training load, HRV status, sleep history, stress history (HR, RHR, HRV)	Proprietary algorithm dynamically integrates contributors, prioritizing recent sleep and recovery data. Personalised baselines adjust over time.	Short-term (last night, recent activity); long-term (7-day HRV, 3-day stress/sleep, 10-day training load)	1–100: Prime (95–100), high (75–94), moderate (50–74), low (25–49), poor (1–24)
Oura	Oura Readiness Score	Oura ring	RHR, HRV, body temperature, recovery index (time spent sleeping after the heart rate lowers to its lowest point during the night), sleep, sleep balance, sleep regularity, previous day activity, activity balance	Integrates contributors against personal baselines using proprietary weighting. Combines short- and long-term trends for a composite readiness score.	Short-term (24 h, 14 days); long-term (2 months)	0–100: Optimal (85–100), good (70–84), Fair (60–69), pay attention (0–59)
Oura	Resilience	Oura ring	Daytime stress load (HR, HRV), motion, body temperature, daytime recovery (restorative time), nighttime recovery (sleep score, RHR, HRV balance, recovery index)	Weighted combination of “stress and recovery contributors” over a 14-day average. Proprietary algorithm; exact weightings not disclosed.	Short-term (5 days of complete data), long-term (14-day average)	Exceptional, strong, solid, adequate, limited
Polar	Nightly Recharge™	Polar wearable devices (e.g., Polar Ignite, vantage series)	HR, HRV, RR, sleep amount, sleep solidity (interruptions, continuity, sleep efficiency), sleep regeneration (REM, deep sleep)	Combines autonomic nervous system charge (HR, HRV and RR) and sleep charge against a 28-day dynamic baseline using proprietary algorithms. Recovery status categorized into six levels: Very good, good, OK, compromised, poor, very poor.	Short-term (overnight, 4-h window), long-term (28-day baseline)	Recovery status categories: Exceptional, strong, solid, adequate, limited
Samsung	Energy score	Samsung Galaxy Watch4 series and later (Samsung Health app)	Sleep (7-day average, timing, consistency), activity (Acute:Chronic workload ratio), sleeping HR, HRV	Proprietary algorithm combining contributors dynamically using optimization and generative AI. Incorporates user demographics like age and gender.	Short-term (24 h); long-term (7 days, workload patterns)	0–100: Accompanied by personalized insights and recommendations for health optimization.
Suunto	Body resources	Suunto wearable devices	Stress (HRV, HR), activity, sleep	Integrates HRV, HR, and activity data with sleep recovery using proprietary algorithms. Updates dynamically to reflect short-term and long-term changes.	Short-term (16 h); nightly recovery	Not provided

Table 4: (continued)

Company	Name of index	Device/platform	Contributors	Calculation	Timescales used	Score scale
Ultrahuman	Dynamic recovery	Ultrahuman ring	AIR Sleep quotient, stress rhythm, body temperature, RHR, HRV	Integrates contributors dynamically against personal baselines with recent data weighted more heavily. Proprietary algorithms limit detailed disclosure.	Short-term (daily); long-term (weekly trends)	0–100: Optimal (85–100), needs focus (<70)
WHOOP	WHOOP recovery	WHOOP wearable devices (e.g., WHOOP 3.0, 4.0)	HRV, RHR, sleep performance, RR, blood oxygen levels, skin temperature (WHOOP 4.0 only)	Proprietary algorithm integrating HRV (~85 %), RHR, and sleep performance with short-term trends; additional metrics for illness detection in WHOOP 4.0.	Short-term (24 h); long-term (cumulative trends)	0–100: Green (67–99 %), yellow (34–66 %), red (1–33 %)
WHOOP	WHOOP strain	WHOOP wearable devices (e.g., WHOOP 3.0, 4.0)	“Cardiovascular load” (HR), muscular load (movement physics, accelerometer, gyroscope)	Integrates “cardiovascular and muscular loads” using a personalized, dynamic proprietary algorithm. Logarithmic scale adjustments based on individual fitness baselines.	Short-term (daily cycles, resets with sleep cycle)	0–21: Light (0–9), moderate (10–13), high (14–17), all out (18–21)
WHOOP	Stress monitor	WHOOP wearable devices (e.g., WHOOP 3.0, 4.0)	HR, HRV, motion	Integrates HR and HRV against 14-day personal baseline, accounts for motion to differentiate stress from physical exertion.	Short-term (real-time), long-term (14-day baseline)	0–3: Low (0), moderate (1, 2), high (3)
Withings	Health improvement score	Withings wearable devices and scales	Activity (steps, active minutes), body weight, body fat %, muscle mass), heart (RHR, blood pressure, vascular age, ECGs), sleep (duration, quality)	Proprietary algorithm not disclosed. Combines sub-scores from tracked metrics to provide a single wellness score. Sub-scores are based on individual measurement categories.	Short-term (daily/tracking), long-term (ongoing improvements)	1–100: Higher scores indicate better overall health and wellness

Company: the manufacturer of the wearable device offering the health index. Name of index: The proprietary name of the health or readiness index. Device/platform: the specific wearable devices or platforms where the index is available. Contributors: The primary physiological and behavioral metrics integrated into the index. Calculation: a brief description of how the contributors are processed to generate the score. Timescales used: the temporal range of data considered (e.g., short-term, long-term). Score scale: The range of values for the index and any interpretive categories or thresholds.

numerical value, the validity and robustness of their approaches is unclear. These tools are marketed as supporting health monitoring, optimising performance, and promoting recovery [44–47, 49–58] – and although specific metrics, such as heart rate variability for stress monitoring [60], or actigraphy for sleep onset detection [61, 62], are underpinned by robust empirical evidence, the composite indices themselves have undergone little to no validation in the peer-reviewed scientific literature [63, 64]. This is concerning given their popularity among users [48], and their intended applicability across diverse populations and physiological conditions. The reliance on opaque algorithms also presents significant barriers to independent scrutiny, hindering reproducibility and raising concerns about the robustness of the outputs [65, 66].

Physiological rationale for resting heart rate and heart rate variability

Findings from this analysis revealed that heart rate variability and resting heart rate were two of the most common contributors to the different CHS; 86 % and 79 % of indices incorporated resting heart rate and heart rate variability in their calculations, respectively. This prominence reflects their physiological significance; resting heart rate reflects the intrinsic pace-making activity of the sinoatrial node, which is modulated by the balance between sympathetic and parasympathetic inputs [67]. Under resting conditions, parasympathetic dominance reduces the sinoatrial node's intrinsic firing rate (~100 beats per minute) to typical values of 50–70 beats per minute in the general population [67]. Higher resting heart rate is associated with lower cardiorespiratory fitness and less parasympathetic tone, while lowered values often indicate reduced sympathetic activity or increased cardiovascular efficiency [68]. The inclusion of resting heart rate in CHS is therefore well-justified, as it provides a measure of baseline cardiovascular function [68, 69].

However, resting heart rate alone is less sensitive to acute physiological changes, such as an acute training 'stress', limiting its ability to capture dynamic autonomic fluctuations [70, 71]. In contrast, heart rate variability provides a more nuanced view of autonomic balance by quantifying fluctuations in the time intervals between successive heartbeats (RR intervals) [72]. Heart rate variability directly reflects rapid, millisecond-scale adjustments in parasympathetic tone, particularly vagal modulation during the respiratory cycle [71]. High heart rate variability at rest indicates robust parasympathetic activity and physiological adaptability, whereas low heart rate variability is associated with stress, fatigue, and poor

health outcomes [70, 73] – although it should be noted that what represents "high" and "low" are unique to the individual [73]. Among the indices analysed, heart rate variability – more accurately termed pulse rate variability because it is derived from PPG in wearables [74] – was predominantly calculated using the root mean square of successive differences (rMSSD) between heartbeats, a standard method for capturing parasympathetic-driven variability [71]. This approach is widely recognized for its ecological validity over short- and ultra-short measurement intervals (1–5 min), although not all popular consumer wearables have adopted this approach [74].

The complementarity of resting heart rate and heart rate variability underpins their integration into composite indices. Resting heart rate captures slower, cumulative changes in autonomic tone, such as those resulting from chronic stress or shifts in fitness [68–70], while heart rate variability provides a dynamic indicator of acute autonomic responses [67, 71–73]. For example, heart rate variability typically recovers more rapidly than resting heart rate following intense exercise, reflecting transient physiological states [71]. This temporal distinction supports the use of heart rate variability as an indicator of readiness or recovery, with resting heart rate providing context for longer-term trends.

Variability in measurement protocols across manufacturers

The measurement protocols for these metrics, however, varied significantly across different companies and indices. Some companies sample their heart rate data during specific sleep periods to minimise confounding factors, but differences in timing, duration, and algorithmic approaches were identified. For example, Fitbit's Daily Readiness Score calculates heart rate variability using the longest period of sleep that exceeds 3 h [44], while Garmin's Training Readiness metric averages 5-minute heart rate variability windows throughout a full night of sleep [50]. Oura's Readiness Score computes heart rate variability as the mean of all 5-minute samples across the entire sleep period [46], whereas Polar's Nightly Recharge™ focuses on a 4-hour window beginning 30 min after sleep onset [52]. WHOOP's Recovery and Stress Scores measure heart rate variability during the final slow-wave sleep period [47, 56, 57], however the classification accuracy of WHOOP's actigraphy-derived sleep stage identification is unclear – it is widely acknowledged that the accuracy of consumer wearables for measuring sleep is poor [43]. In contrast, Samsung's Energy Score is based on continuous heart rate

variability monitoring [53], but details on its integration and calculation remain opaque; similarly, Ultrahuman's Dynamic Recovery Score detects deviations from user-specific 'baselines' rather than absolute values [55] – but the specific formulae and sampling periods were unclear.

Challenges with continuous monitoring and signal validity

Compounding this issue is the fact that heart rate variability's physiological link to parasympathetic activity is only valid under specific conditions, typically at rest, and artifacts introduced by daily activities such as swallowing, talking, or drinking water can "break" this relationship when heart rate variability is measured continuously [75, 76]. This makes composite metrics derived from 24/7 heart rate variability data, such as stress and recovery scores, scientifically questionable without proper contextualisation or validation.

The approach to calculating resting heart rate among the different manufacturers was similarly inconsistent. Fitbit derives resting heart rate from data collected during sleep or periods of inactivity throughout the day [44]. Garmin calculates resting heart rate as the lowest 30-min average within a 24-h period [50], while Oura combines nightly averages with the lowest resting heart rate sampled every 10 min [46]. Polar's Nightly Recharge™ incorporates resting heart rate from a 4-hour window after sleep onset which is similar to their approach to measuring heart rate variability [52]. WHOOP, again calculates resting heart rate during the same actigraphy-derived slow-wave sleep stage as their heart rate variability calculation [47, 56, 57]. Samsung measures resting heart rate continuously during periods of inactivity but does not disclose thresholds or protocols [53], while Suunto allows users to input resting heart rate manually or rely on the lowest value recorded during sleep [54]. Taken together, the variability in algorithms, for measuring heart rate variability and resting heart rate bely any meaningful comparison between their outputs; differences in sampling protocols, timing, and integration strategies across platforms limit direct comparability between the composite indices.

Concerns regarding scientific robustness and real-world alignment

These discrepancies underscore broader uncertainties regarding the validity, interpretability, and real-world impact of each CHS. While marketed as objective indicators of

readiness, recovery, or stress, the inconsistencies in their calculation methodologies – particularly in how and when physiological signals such as heart rate variability, resting heart rate, and activity levels are integrated – raise concerns about their scientific robustness. The research in this field is sparse, but the available studies highlight potential disconnects between CHS outputs and real-world physiological or psychological states.

For example, a 2023 study of elite swimmers investigated the relationship between wearable-derived physiological metrics and metabolic and psychological stress markers [63]. The study found a moderate positive correlation between heart rate variability and relative resting metabolic rate ($r=0.448$, $p=0.032$), suggesting that increased heart rate variability may reflect some aspects of metabolic function. However, no correlation was found between heart rate variability and thyroid hormone levels or the resting metabolic rate ratio, indicating that this wearable-derived measure does not consistently align with broader markers of energy availability. Furthermore, WHOOP's recovery score showed no correlation with metabolic suppression when analyzing all participants together, and its relationship with the resting metabolic rate ratio was only significant in male swimmers ($r=0.653$, $p=0.041$), but not in females. The same study also explored the relationship between wearable-derived CHS metrics and self-reported stress levels. Heart rate variability was negatively correlated with sport-specific stress ($r= -0.462$, $p=0.026$) and total stress ($r= -0.459$, $p=0.028$), supporting its role as a sensitive marker of autonomic stress. However, WHOOP's recovery score showed no significant correlation with self-reported stress or recovery measures in either sex, raising concerns about whether such CHS outputs provide meaningful insights into psychological stress or autonomic recovery.

Similarly, a 2025 study on Dutch police officers assessed the relationship between wearable-derived stress scores (including Garmin's "Body Battery" metric) and subjective stress perception [64]. The study found that wearable-based stress metrics did not consistently align with users' self-reported experiences of stress and recovery. While using a CHS-equipped wearable led to small but statistically significant improvements in stress awareness, self-efficacy, and well-being (Hedges' $g=0.25-0.46$, $p<0.05$), many participants struggled to interpret their wearable-derived stress scores. Some perceived their stress levels as high despite their CHS metrics suggesting otherwise, and vice versa. Additionally, while CHS use increased awareness of physiological trends, this did not necessarily translate to significant behavioural changes, reinforcing concerns that CHS alone may be insufficient for guiding meaningful adaptations in stress regulation or recovery behaviours.

Multicollinearity and redundancy in CHS calculations

Taken together, these findings emphasize the need for independent validation of CHS to determine whether they provide physiologically and psychologically relevant insights or whether their outputs may mislead users due to misalignment with subjective experiences and established physiological markers. Indeed, when considered in isolation, a primary limitation of the CHS concept lies in the inherent interdependence of the core metrics that contribute to their overall scores. Many CHS exhibit an inherent bias due to multicollinearity among their input metrics, such as sleep duration and heart rate variability, which are often interdependent. For instance, a poor night's sleep typically results in reduced heart rate variability, yet penalizing 'readiness' twice – once for low sleep quality and again for reduced heart rate variability – introduces redundancy and amplifies the negative score disproportionately. Conversely, if heart rate variability remains within normal ranges despite poor sleep, it suggests the user's body has effectively mitigated the physiological impact of sleep disruption, making an additional penalty for poor sleep unnecessary. Such over-penalization creates a false perception of precision but often obscures the actual physiological response to stressors. This limitation highlights a deeper issue in the design of CHS: the tendency to conflate behavioural data with physiological outcomes, resulting in outputs that are not wholly reflective of the body's actual state. Seventy-one percent of the indices integrated physical activity and 'stress' (inferred based on fluctuations in baseline heart rate taken throughout the day), and 71 % integrated some construct of sleep – often without adequately accounting for how an individual's physiology responds to these factors. For instance, if a user undertakes a high-intensity workout, their physiological readiness for subsequent activity should ideally be assessed through a single measure like heart rate variability, which reflects how well the body has assimilated the stressor [71, 73]. Penalizing readiness solely on the basis of high activity data from the previous day, regardless of physiological recovery, undermines the purpose of measuring these signals. By relying on behavioural data to infer physiological states, CHS may fail to capture the nuanced variability in individual responses, limiting their utility for personalised health recommendations. Granted, several of the indices integrate 'baseline' measurements to contextualize short-term fluctuations – Fitbit, Garmin, Oura, Polar, Ultrahuman and WHOOP all incorporate user-specific baselines into their calculations – but the lack of consistency in defining and calculating baselines across platforms undermines our

ability to compare indices or draw meaningful conclusions about their utility.

Lack of standardization and terminological ambiguity

Ultimately, our synthesis cannot shed light on many of these unknowns – and this reflects the primary limitation of this work, and the field in general: the lack of transparency and standardization among the CHS we evaluated. Despite their widespread adoption and growing role in consumer health monitoring [48, 65], the manufacturers provided limited details about the algorithms underpinning these tools, how often they change these algorithms, the rationale behind their contributors, or the validity of their outputs in diverse populations. Much of the data available for analysis was derived from publicly accessible resources, such as user manuals, white papers, and app interfaces, none of which offered the level of detail necessary to fully evaluate the scientific integrity of these indices.

A symptom of this lack of standardization is the ambiguity surrounding the terminology used in the different indices. For example, while multiple manufacturers – including Garmin, Fitbit, and Oura – use variations of the term "readiness", their calculation methodologies and contributing metrics differ significantly. For instance, Fitbit's Daily Readiness Score integrates heart rate variability, resting heart rate, sleep quantity, and sleep architecture, whereas Garmin's Training Readiness includes activity levels and recent training load, and Oura's Readiness Score extends further to incorporate body temperature. Despite their terminological similarity, these indices are not methodologically equivalent, making direct comparisons difficult. Users may reasonably assume that readiness scores from different devices reflect comparable physiological constructs, yet differences in data sources, weightings, and interpretation frameworks introduce inconsistencies that challenge their scientific and practical applicability.

The net result of the lack of transparency is that key questions about how the different contributors are weighted, how composite scores are calculated, whether any given score can be considered superior in terms of validity, transparency, or scientific rigor, and whether these scores align with physiological realities remain unanswered. This lack of transparency has significant implications for the scientific reproducibility and clinical applicability of CHS. Without access to proprietary algorithms or validation datasets, researchers and clinicians cannot independently assess whether the scores produced by CHS are meaningful, reproducible, or generalizable.

This is particularly concerning given the inherent variability in physiological responses among individuals and across populations. For example, metrics like heart rate variability are highly sensitive to factors such as age, sex, fitness level, and circadian rhythms [73], yet no evidence was found to suggest that these factors are consistently accounted for in the indices we reviewed. Similarly, the reliance on fixed baselines for assessing deviations in metrics like resting heart rate or heart rate variability may not adequately capture the dynamic, context-dependent nature of physiological responses, further limiting the utility of these tools.

The absence of rigorous validation in diverse populations also raises concerns about the equity of CHS. While some indices may perform well in the demographic groups most commonly represented in validation studies – typically young, healthy, and predominantly male participants [77, 78] – there is little evidence to suggest that these tools are reliable or accurate in populations with different characteristics. For instance, older adults, individuals with chronic conditions, and those from diverse racial or ethnic backgrounds may exhibit physiological patterns that differ from the normative baselines assumed by these algorithms. Without targeted validation efforts, CHS risk perpetuating biases that could undermine their utility for underserved populations, further exacerbating health inequities in digital health.

Implications for clinical, personal, and public health use

These limitations have significant implications for the clinical, telehealth, and personal utility of CHS. In remote health monitoring and telemedicine, where CHS could serve as potential digital biomarkers for assessing patient status, their lack of validation across diverse populations introduces risks of misclassification and inappropriate decision-making. Clinicians may be reluctant to rely on these scores for decision-making if the underlying methodologies are unclear or if the indices have not been validated across a sufficiently broad population. Moreover, without robust evidence to support the accuracy and reliability of CHS, their use in telehealth-based personalized health interventions or disease prevention programs may lead to inconsistent or inaccurate insights, potentially undermining trust among users and healthcare providers alike. From a personal health perspective, CHS have the potential to empower individuals by providing insights into their recovery, readiness, and overall health status. However, the utility of these tools is compromised when the outputs are overly complex,

insufficiently validated, or fail to align with the user's unique physiological profile. For example, indices that rely heavily on behavioural data may penalize individuals who deviate from normative patterns (e.g., athletes or shift workers), despite maintaining adequate physiological recovery. To maximize their utility, CHS must evolve to incorporate individualized baselines and adaptive algorithms that account for user-specific responses to stress, recovery, and activity. Future advancements in CHS may benefit from independent validation frameworks, industry-wide certification standards, or structured regulatory oversight to ensure their reliability and applicability without compromising proprietary algorithms.

Summary and outlook

The core concept of a CHS – offering wearable device users simplified interpretations of complex physiological data to inform health-related decisions – is a promising innovation. However, this analysis underscores significant gaps in the transparency, standardization, and validation of consumer wearable CHS. While metrics such as resting heart rate and heart rate variability are well-supported by empirical evidence, the proprietary algorithms underpinning CHS and the lack of methodological disclosure hinder reproducibility and independent scrutiny. Furthermore, the absence of rigorous validation across diverse populations raises critical concerns about generalizability and equity. The variability in measurement protocols and integration strategies further complicates comparability and reliability, ultimately limiting both scientific and clinical confidence in these tools. This lack of standardization undermines the scientific basis of CHS and limits their long-term potential for clinical adoption, personalized health management, and public health applications. If CHS are to evolve beyond opaque consumer-facing features into scientifically validated health metrics, the field must shift from merely evaluating industry-led innovations to defining best practices for their development and deployment.

Research ethics: Not applicable. The study did not involve human participants or animals and was exempt from ethical review.

Informed consent: Not applicable.

Author contributions: CD: Conceptualized the study, led the manuscript preparation, and led the analysis of composite health scores. MB: assisted in synthesizing the results, and drafted sections of the manuscript. RL: Provided input into the methodological framework, data extraction, and manuscript

revisions. DB: Reviewed and validated the findings, providing domain-specific expertise on wearable technologies. MA: Contributed to the interpretation of findings, particularly regarding algorithmic transparency, and critically revised the manuscript for important intellectual content.

Use of Large Language Models, AI and Machine Learning Tools: None declared.

Conflict of interest: The authors state no conflict of interest.

Research funding: This project was funded by the Health Research Board in Ireland (Grant ID: HRB ILP-PHR-2024-005) and Research Ireland (Grant ID: 12/RC/2289_P2).

Data availability: Not applicable. This study did not generate or analyse datasets.

References

- Day S. Self-tracking over time: the FITBIT® phenomenon [internet]. Place of Publication; 2016. Available from: https://citrenz.org.nz/citrenz/conferences/2016/pdf/2016CITRENZ_1_Day_Fitbit_13-3.pdf.
- St FRG, St George SM, Leite R, Kobayashi M, Agosto Y, Jake-Schoffman DE. Use of fitbit devices in physical activity intervention studies across the life course: narrative review. *JMIR Mhealth Uhealth* 2021;9:e23411.
- Esmonde K, Jette S. Assembling the 'Fitbit subject': a Foucauldian-sociomaterialist examination of social class, gender and self-surveillance on Fitbit community message boards. *Health (London)* 2020;24:299–314.
- Doherty C. How accurate are wearable fitness trackers? Less than you might think [internet]. Place of Publication; 2024. Available from: <https://theconversation.com/how-accurate-are-wearable-fitness-trackers-less-than-you-might-think-236462>.
- Wolf G. Quantifiedself.com2007. Available from: https://web.archive.org/web/20140222150044/http://www.ted.com/talks/gary_wolf_the_quantified_self.html.
- Patel MS, Asch DA, Volpp KG. Wearable devices as facilitators, not drivers, of health behavior change. *JAMA* 2015;313:459–60.
- Statista. Wearables statista dossier [internet]. Place of Publication; 2024. Available from: <https://www.statista.com/statistics/437871/wearables-worldwide-shipments>.
- Kang HS, Exworthy M. Wearing the future-wearables to empower users to take greater responsibility for their health and care: scoping review. *JMIR Mhealth Uhealth* 2022;10:e35684.
- Canali S, Schiaffonati V, Aliverti A. Challenges and recommendations for wearable devices in digital health: data quality, interoperability, health equity, fairness. *PLOS Digit Health* 2022;1:e0000104.
- Sasaki JE, Hickey A, Mavilia M, Tedesco J, John D, Kozey KS, et al. Validation of the Fitbit wireless activity tracker for prediction of energy expenditure. *J Phys Activ Health* 2015;12:149–54.
- Swan M. Sensor mania! The internet of things, wearable computing, objective metrics, and the quantified self 2.0. *J Sens Actuator Netw* 2012;1:217–53.
- Rodgers MM, Pai VM, Conroy RS. Recent advances in wearable sensors for health monitoring. *IEEE Sens J* 2015;15:3119–26.
- Prisco G, Pirozzi MA, Santone A, Esposito F, Cesarelli M, Amato F, et al. Validity of wearable inertial sensors for gait analysis: a systematic review. *Diagnostics* 2024;15:36.
- Benson LC, Clermont CA, Bošnjak E, Ferber R. The use of wearable devices for walking and running gait analysis outside of the lab: a systematic review. *Gait Posture* 2018;63:124–38.
- Bennett JP, Liu YE, Kelly NN, Quon BK, Wong MC, McCarthy C, et al. Next-generation smart watches to estimate whole-body composition using bioimpedance analysis: accuracy and precision in a diverse, multiethnic sample. *Am J Clin Nutr* 2022;116:1418–29.
- Castaneda D, Esparza A, Ghamari M, Soltanpur C, Nazeran H. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron* 2018;4:195–202.
- Hilbel T, Frey N. Review of current ECG consumer electronics (pros and cons). *J Electrocardiol* 2023;77:23–8.
- Radin JM, Quer G, Pandit JA, Gadaleta M, Baca-Motes K, Ramos E, et al. Sensor-based surveillance for digitising real-time COVID-19 tracking in the USA (DETECT): a multivariable, population-based, modelling study. *Lancet Digit Health* 2022;4:e777–86.
- Radin JM, Wineinger NE, Topol EJ, Steinhubl SR. Harnessing wearable device data to improve state-level real-time surveillance of influenza-like illness in the USA: a population-based study. *Lancet Digit Health* 2020;2:e85–93.
- González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. *Nature* 2008;453:779–82.
- Blumenstock J, Cadamuro G, On R. Predicting poverty and wealth from mobile phone metadata. *Science* 2015;350:1073–6.
- Low CA, Bovbjerg DH, Ahrendt S, Choudry MH, Holtzman M, Jones HL, et al. Fitbit step counts during inpatient recovery from cancer surgery as a predictor of readmission. *Ann Behav Med* 2018;52:88–92.
- Abbadessa G, Lavorgna L, Miele G, Mignone A, Signoriello E, Lus G, et al. Assessment of multiple sclerosis disability progression using a wearable biosensor: a pilot study. *J Clin Med* 2021;10:1160.
- Finley DJ, Fay KA, Batsis JA, Stevens CJ, Sacks OA, Darabos C, et al. A feasibility study of an unsupervised, pre-operative exercise program for adults with lung cancer. *Eur J Cancer Care* 2020;29:e13254.
- Pyrkov TV, Slipensky K, Barg M, Kondrashin A, Zhurov B, Zenin A, et al. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Sci Rep* 2018;8:5210.
- Low CA, Bovbjerg DH, Jenkins FJ, Ahrendt S, Choudry MH, Holtzman M, et al. Patient-reported and fitbit-assessed physical activity: associations with inflammation and risk of readmission after metastatic cancer surgery. *Psychosom Med* 2016;78:A127.
- Takahashi T, Kumamaru M, Jenkins S, Saitoh M, Morisawa T, Matsuda H. In-patient step count predicts re-hospitalization after cardiac surgery. *J Cardiol* 2015;66:286–91.
- Moshe I, Terhorst Y, Opoku Asare K, Sander LB, Ferreira D, Baumeister H, et al. Predicting symptoms of depression and anxiety using smartphone and wearable data. *Front Psychiatr* 2021;12:625247.
- Cox SM, Lane A, Volchenboum SL. Use of wearable, mobile, and sensor technology in cancer clinical trials. *JCO Clin Cancer Inform* 2018;2:11.
- Keogh A, Brennan C, Johnston W, Dickson J, Leslie SJ, Burke D, et al. Six-month pilot testing of a digital health tool to support effective self-care in people with heart failure: mixed methods study. *JMIR Form Res* 2024;8:e52442.
- Ferguson T, Olds T, Curtis R, Blake H, Crozier AJ, Dankiw K, et al. Effectiveness of wearable activity trackers to increase physical activity and improve health: a systematic review of systematic reviews and meta-analyses. *Lancet Digit Health* 2022;4:e615–26.
- Stiglbauer B, Weber S, Batinic B. Does your health really benefit from using a self-tracking device? Evidence from a longitudinal randomized control trial. *Comput Hum Behav* 2019;94:131–9.

33. Torres EN, Zhang T. The impact of wearable devices on employee wellness programs: a study of hotel industry workers. *Int J Hospit Manag* 2021;93:102769.
34. Souza M, Miyagawa T, Melo P, Maciel F, editors. *Wellness programs: wearable technologies supporting healthy habits and corporate costs reduction*. HCI International 2017 – posters' extended abstracts. Cham: Springer International Publishing; 2017.
35. Loader J, Montgomery PG, Williams MD, Lorenzen C, Kemp JG. Classifying training drills based on movement demands in Australian football. *Int J Sports Sci Coach* 2012;7:57–67.
36. Varley MC, Fairweather IH, Aughey RJ. Validity and reliability of GPS for measuring instantaneous velocity during acceleration, deceleration, and constant motion. *J Sports Sci* 2012;30:121–7.
37. Foster CD, Twist C, Lamb KL, Nicholas CW. Heart rate responses to small-sided games among elite junior rugby league players. *J Strength Condit Res* 2010;24:906–11.
38. Noonan B, Bancroft RW, Dines JS, Bedi A. Heat- and cold-induced injuries in athletes: evaluation and management. *J Am Acad Orthop Surg* 2012;20:744–54.
39. Argent R, Hetherington-Rauth M, Stang J, Tarp J, Ortega FB, Molina-Garcia P, et al. Recommendations for determining the validity of consumer wearables and smartphones for the estimation of energy expenditure: expert statement and checklist of the INTERLIVE network. *Sports Med* 2022;52:1817–32.
40. Johnston W, Judice PB, Molina García P, Mühlen JM, Lykke Skovgaard E, Stang J, et al. Recommendations for determining the validity of consumer wearable and smartphone step count: expert statement and checklist of the INTERLIVE network. *Br J Sports Med* 2021;55:780–93.
41. Molina-Garcia P, Notbohm HL, Schumann M, Argent R, Hetherington-Rauth M, Stang J, et al. Validity of estimating the maximal oxygen consumption by consumer wearables: a systematic review with meta-analysis and expert statement of the INTERLIVE network. *Sports Med* 2022;52:1577–97.
42. Mühlen JM, Stang J, Lykke Skovgaard E, Judice PB, Molina-Garcia P, Johnston W, et al. Recommendations for determining the validity of consumer wearable heart rate devices: expert statement and checklist of the INTERLIVE Network. *Br J Sports Med* 2021;55:767–79.
43. Doherty C, Baldwin A, Argent R, Keogh A, Caulfield B. Keeping pace with wearables: a living umbrella review of systematic reviews evaluating the accuracy of commercial wearable technologies in health measurement. *Sports Med* 2024;54:2907–26.
44. Centre FH. What's my daily readiness score in the Fitbit app? [internet]. Place of publication; 2025. Available from: <https://support.google.com/fitbit/answer/14236710?hl=en>.
45. Garmin. Body Battery™ energy monitoring [internet]. Place of Publication; 2025. Available from: <https://www.garmin.com/en-IE/garmin-technology/health-science/body-battery/> {Garmin, 2024 #1277}.
46. Oura. Your Oura readiness score [internet]. Place of Publication; 2025. Available from: <https://ouraring.com/blog/readiness-score/> ?srsltid=AfmBOOr6dc1NQGLqW7h-pmsGQcBj79gF5Gu7X1ji-FqwnGBqz3671_gX.
47. WHOOP. How does WHOOP strain work? [internet]. Place of Publication; 2025. Available from: <https://www.whoop.com/us/en/thelocker/how-does-whoop-strain-work-101/> ?srsltid=AfmBOovr5ge9Npp_1uvTnk4_q96vMbDMUJSBcFQIE8jFPpEXTav0IrIf.
48. Ibrahim AH, Beaumont CT, Strohacker K. Exploring regular exercisers' experiences with readiness/recovery scores produced by wearable devices: a descriptive qualitative study. *Appl Psychophysiol Biofeedback* 2024;49:395–405.
49. Coros. COROS METRICS | introducing daily stress [internet]. Place of Publication; 2025. Available from: <https://coros.com/stories/coros-metrics/c/daily-stress>.
50. Garmin. Understanding Garmin training readiness [internet]. Place of Publication; 2025. Available from: <https://www.garmin.com/en-IE/garmin-technology/running-science/physiological-measurements/training-readiness/>.
51. Oura. Resilience [internet]. Place of Publication; 2025. Available from: <https://support.ouraring.com/hc/en-us/articles/25358829055251-Resilience>.
52. Polar. Nightly Recharge™ build the day on the night [internet]. Place of Publication; 2025. Available from: <https://www.polar.com/en/smart-coaching/nightly-recharge?srsltid=AfmBOooI2jGrDAHPRSojgrMcb-IIczumyWmAKHsE11EDXNwYv-FYwo9q>.
53. Health S. Everyday wellness [internet]. Place of Publication; 2025. Available from: https://www.samsung.com/ie/apps/samsung-health/#energy_score.
54. Suunto. Suunto 7 user guide [internet]. Place of Publication; 2025. Available from: https://www.suunto.com/en-ie/Support/Product-support/suunto_7/suunto_7/activity-tracking/body-resources/.
55. Ultrahuman. Ultrahuman ring AIR: dynamic recovery [internet]. Place of Publication; 2025. Available from: <https://blog.ultrahuman.com/blog/ultrahuman-ring-recovery-score-guide/>.
56. WHOOP. Introducing stress monitor: a new way to monitor & manage stress [internet]. Place of Publication; 2025. Available from: <https://www.whoop.com/ie/en/thelocker/introducing-stress-monitor-a-new-way-to-monitor-manage-stress/?srsltid=AfmBOophXF5jTtqhGagZnCbDmqLrejbOFFVfsUvoUIoelY8zkpqz-IWp>.
57. WHOOP. How does WHOOP recovery work? [internet]. Place of Publication; 2025. Available from: https://www.whoop.com/ie/en/thelocker/how-does-whoop-recovery-work-101/?srsltid=AfmBOoqDhgN_DMypJrxrdlHPH-VuLGC_vbAyVudOEknDa4eTeu6CC.
58. Withings. Withings+ – learn more about your health improvement score [internet]. Place of Publication; 2025. Available from: <https://support.withings.com/hc/en-us/articles/15547200464273-Withings-Learn-more-about-your-Health-Improvement-Score>.
59. Polar. White paper – polar nightly Recharge [internet]. Place of Publication: 2019. Available from: https://www.polar.com/img/static/whitepapers/pdf/polar-nightly-recharge-white-paper.pdf?srsltid=AfmBOoqzaWUVvKZVYn86qN9v8MBCWEvUR1HxxVonxVtgRDjh_uBGgF6L.
60. Kim HG, Cheon Ej, Bai DS, Lee YH, Koo BH. Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig* 2018;15:235–45.
61. Marino M, Li Y, Rueschman MN, Winkelman JW, Ellenbogen JM, Solet JM, et al. Measuring sleep: accuracy, sensitivity, and specificity of wrist actigraphy compared to polysomnography. *Sleep* 2013;36:1747–55.
62. Jean-Louis G, Kripke DF, Cole RJ, Assmus JD, Langer RD. Sleep detection with an accelerometer actigraph: comparisons with polysomnography. *Physiol Behav* 2001;72:21–8.
63. Lundstrom EA, De Souza MJ, Koltun KJ, Strock NCA, Canil HN, Williams NI. Wearable technology metrics are associated with energy deficiency and psychological stress in elite swimmers. *Int J Sports Sci Coach* 2024;19:1578–87.
64. de Vries HJ, Delahij R, van Zwieten M, Verhoef H, Kamphuis W. The effects of self-monitoring using a smartwatch and smartphone app on stress awareness, self-efficacy, and well-being-related outcomes in police officers: longitudinal mixed design study. *JMIR Mhealth Uhealth* 2025;13:e60708.

65. Schumann M, Doherty C. Bridging gaps in wearable technology for exercise and health professionals: a brief review. *Int J Sports Med* 2024; 949–57. <https://doi.org/10.1055/a-2376-6332>.
66. Keogh A, Argent R, Doherty C, Duignan C, Fennelly O, Purcell C, et al. Breaking down the digital fortress: the unseen challenges in healthcare technology—lessons learned from 10 Years of research. *Sensors (Basel)* 2024;24:3780.
67. Gordan R, Gwathmey JK, Xie LH. Autonomic and endocrine control of cardiovascular function. *World J Cardiol* 2015;7:204–14.
68. Fox K, Borer JS, Camm AJ, Danchin N, Ferrari R, Lopez Sendon JL, et al. Resting heart rate in cardiovascular disease. *J Am Coll Cardiol* 2007;50: 823–30.
69. Jensen MT, Suadicani P, Hein HO, Gyntelberg F. Elevated resting heart rate, physical fitness and all-cause mortality: a 16-year follow-up in the Copenhagen Male Study. *Heart* 2013;99:882–7.
70. Bosquet L, Merkari S, Arvisais D, Aubert AE. Is heart rate a convenient tool to monitor over-reaching? A systematic review of the literature. *Br J Sports Med* 2008;42:709–14.
71. Altini M, Plews D. What is behind changes in resting heart rate and heart rate variability? A large-scale analysis of longitudinal measurements acquired in free-living. *Sensors (Basel)* 2021;21:7932.
72. Pichot V, Roche F, Gaspoz JM, Enjolras F, Antoniadis A, Minini P, et al. Relation between heart rate variability and training load in middle-distance runners. *Med Sci Sports Exerc* 2000;32:1729–36.
73. Tiwari R, Kumar R, Malik S, Raj T, Kumar P. Analysis of heart rate variability and implication of different factors on heart rate variability. *Curr Cardiol Rev* 2021;17:e160721189770.
74. O'Grady B, Lambe R, Baldwin M, Acheson T, Doherty C. The validity of apple watch series 9 and ultra 2 for serial measurements of heart rate variability and resting heart rate. *Sensors* 2024;24:6220.
75. Yildiz M, Doma S. Effect of spontaneous saliva swallowing on short-term heart rate variability (HRV) and reliability of HRV analysis. *Clin Physiol Funct Imag* 2018;38:710–7.
76. Kotani K, Tachibana M, Takamasu K. Investigation of the influence of swallowing, coughing and vocalization on heart rate variability with respiratory-phase domain analysis. *Methods Inf Med* 2007;46: 179–85.
77. Zinzuwadia A, Singh JP. Wearable devices – addressing bias and inequity. *Lancet Digit Health* 2022;4:e856–7.
78. Sui A, Sui W, Liu S, Rhodes R. Ethical considerations for the use of consumer wearables in health research. *Digital Health* 2023;9: 20552076231153740.