BBS /// August 27, 2026

Clinical Validation of Novel Digital Measures: Statistical Methods for Reliability Evaluation

Bohdana Ratitch & Andrew Trigg

Statistical Innovation, Bayer

Outline

- Introduction and V3+ Framework
- Case study: Sleepiz One+ device
- Reliability concepts & study design
- Reliability metrics & results illustrations
- Recommendations

Digital Biomarkers

Viewpoint

Digit Biomark 2022;6:83–97
DOI: 10.1159/000525897

Received: February 8, 2022
Accepted: May 31, 2022
Published online: August 29, 2022

Considerations for Analyzing and Interpreting Data from Biometric Monitoring Technologies in Clinical Trials

Bohdana Ratitch[a] Isaac R. Rodriguez-Chavez[b] Abhishek Dabral[c] Adriano Fontanari[d] Julio Vega[e] Francesco Onorati[f] Benjamin Vandendriessche[g] Stuart Morton[h] Yasaman Damestani[i]

Digital Biomarkers

Review Article

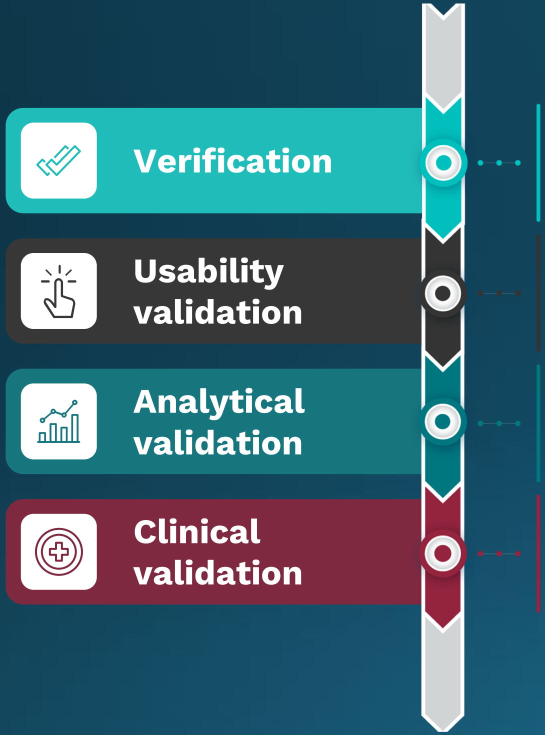
Digit Biomark 2023;7:74–91
DOI: 10.1159/000531054

Received: February 11, 2023
Accepted: May 2, 2023
Published online: August 9, 2023

Clinical Validation of Novel Digital Measures: Statistical Methods for Reliability Evaluation

Bohdana Ratitch[a] Andrew Trigg[b] Madhurima Majumder[c] Vanja Vlajnic[c] Nicole Rethemeier[d] Richard Nkulikiyinka[e]

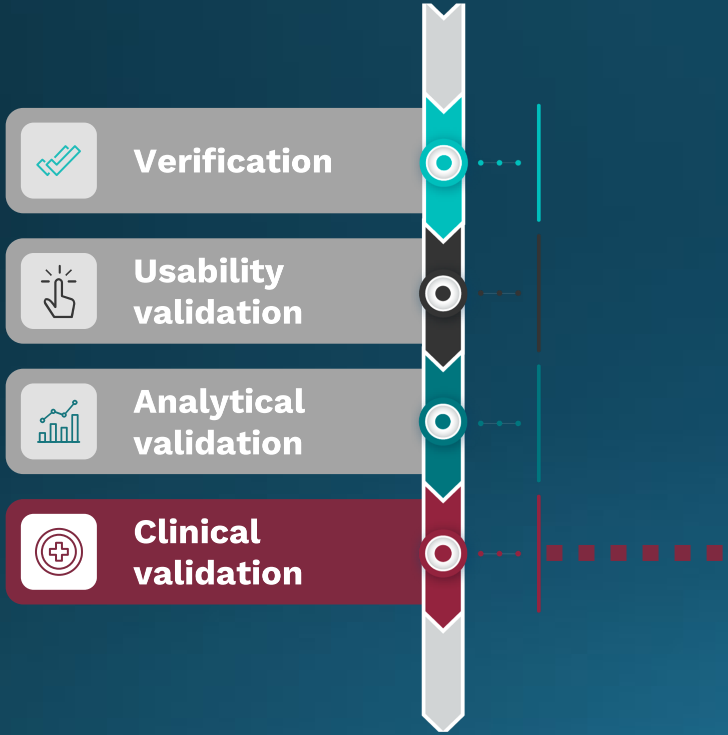
V3+ Framework



A framework designed to demonstrate that a novel digital clinical measure is fit-for-purpose in clinical research, care, and decision-making

Source: playbook.dimesociety.org ,Verification, analytical validation, and clinical validation (V3), Extending V3 with Usability Validation

V3+ Framework

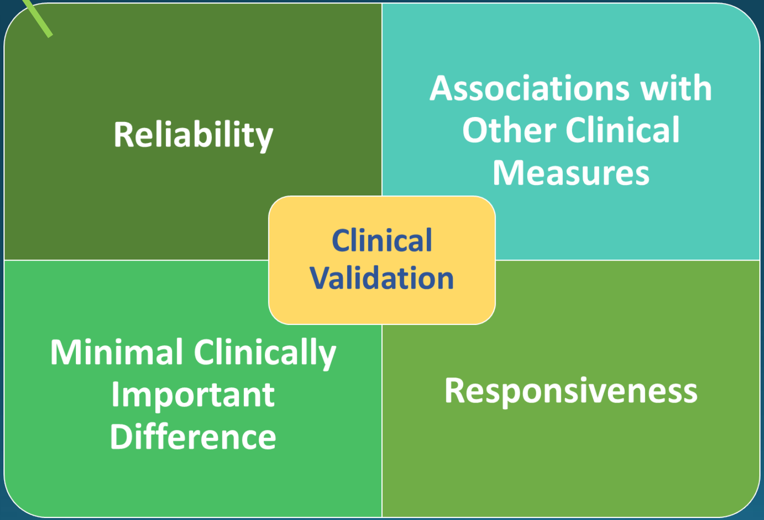


"Evaluates whether a sensor-based digital health technology (sDHT) acceptably identifies, measures, or predicts a meaningful clinical, biological, physical, functional state, or experience, in the stated context of use (which includes a specified population)"

Source: playbook.dimesociety.org ,Verification, analytical validation, and clinical validation (V3), Extending V3 with Usability Validation

Clinical Validation – Key Elements

Prerequisite for subsequent analyses — it characterizes the signal-to-noise ratio and measurement error necessary for interpreting meaningful differences and changes



Source: Based on Fig 2 from DOI: 10.1159/000525897

Case Study: Sleepiz One+

- A contactless sleep monitor developed for longitudinal use at home
- Emits electromagnetic waves that are reflected from the human body; changes in the frequency and phase of these reflections detect body movements
- Body movement signals are converted into sleep parameters

- Administered (and clinical validation performed) in a study of elinzanetant in women with sleep disturbances associated with menopause (NCT06112756)

https://us.sleepiz.com/clinical-trials

Reliability in the sDHT Context

- Intra-rater reliability: Reproducibility of measurements on different occasions
  from the same piece of DHT equipment
  on the same individual
  under the same conditions

- Inter-rater reliability: Reproducibility across different pieces of equipment (of the same kind)
  on the same individual
  under the same conditions

- Internal consistency reliability: Reproducibility across different items of a composite measure (e.g., combining ePRO and sensor-based data)

What Is Reliability?

Reliability is the proportion of variance in a measure attributable to true variance in the underlying concept of interest (COI)

Sources of variability affecting reliability:

- Analytical variability: Variation from stochastic components in the data transformation and clinical measure derivation algorithm
- Intra-subject variability: Random variation in an individual's physiology, behavior, or environment while disease state is stable (e.g., day-of-week effects, weather)
- Inter-subject variability: Factors varying among individuals with the same disease state (e.g., genetics, demographics, comorbidities, lifestyle)

We focus on intra-subject and inter-subject variability at the level of the clinical measure (outcome)

Digital Measure: WASO as Measured by Sleepiz One+

Raw signal	Intermediate-frequency beat signal captured by a radar sensor encoding individual's sub-millimeter body surface displacements
Features	Respiratory rate and its variability, heart rate and its variability, and body movement intensity in 30-second epochs
Classification	Sleep versus wake classification for every 30-second epoch
Outcome	Wakefulness after sleep onset (WASO): Average (weekly) number of minutes awake between the sleep onset and the last awakening
Endpoint	Change in WASO from baseline week to Week 4

Reliability Study Design Considerations

- A repeated-measure design: measurements from n individuals on k occasions - an ($n \times k$) data matrix
- Stability requirement: Multiple measurements must be taken during a period when the participant's disease status is stable with respect to the target COI
- Population representativeness and sufficient between-subject variability: The study sample must cover the full range of disease severity and relevant demographic characteristics
- Measurement conditions should reflect the intended context of use: target population, measurement modality, environment (e.g., free-living vs. lab), sensor placement, and seasonal considerations

Sleepiz One+ Study Design

- A repeated-measure design: Nightly sleep parameters over screening, baseline, and 12 week intervention period. Test-retest reliability assessed between Weeks 6 and 7, and Weeks 8 and 9.
- Stability requirement: For WASO, 3 subsets of stable subjects based on patient-reported outcomes (no/minimal change in questions reflecting time spent awake, perceived WASO, trouble staying asleep). Weeks 6-9 pre-specified based on assumption of higher stability in this period.
- Population and study sample: Women with sleep disturbance associated with the menopause, all randomized subjects agreeing to use of the Sleepiz One+ device.
- Measurement conditions: In the home environment. Weeks with a lab-based sleep assessment were avoided to maintain consistency.

General Statistical Model – Continuous Measures

General two-way random-effects ANOVA model:

$$y_{ij} = \mu + r_i + c_j + v_{ij}$$

$r_i \sim N(0, \sigma_r^2)$ (inter-subject variability)

$c_j \sim N(0, \sigma_c^2)$ (intra-subject/occasion variability),

$v_{ij} \sim N(0, \sigma_v^2)$ (residual/measurement error)

Key parameters: variance components

$$\text{Intraclass Correlation Coefficient (ICC)} = \frac{\textit{Variance of interest}}{\textit{Variance of interest} + \textit{Unwanted variance}}$$

The model may be simplified based on assumptions about the measurement process

ICC Definitions

Categorization per Liljequist et al. (2019) (https://doi.org/10.1159/000531054)

- No systematic bias assumed (one-way random-effects model):

$$\mathrm{ICC}(1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2}$$

$$y_{ij} = \mu + r_i + v_{ij}$$

- Absolute agreement with (random) bias (two-way random- or mixed-effects model):

$$\mathrm{ICC}(A, 1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_v^2}$$

$$y_{ij} = \mu + r_i + c_j + v_{ij}$$

- Consistency, ignoring bias:

$$\mathrm{ICC}(C, 1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2}$$

$$y_{ij} = \mu + r_i + c_j + v_{ij}$$

ICC Definitions

when the values are averages of *k* multiple measurements (e.g., weekly average)

Categorization per Liljequist et al. (2019) (https://doi.org/10.1159/000531054)

- No systematic bias assumed (one-way random-effects model):

$$\text{ICC}(k) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2/k}$$

$$y_{ij} = \mu + r_i + v_{ij}$$

- Absolute agreement with (random) bias (two-way random- or mixed-effects model):

$$\text{ICC}(A, k) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_v^2/k}$$

$$y_{ij} = \mu + r_i + c_j + v_{ij}$$

- Consistency, ignoring bias:

$$\text{ICC}(C, k) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_v^2/k}$$

$$y_{ij} = \mu + r_i + c_j + v_{ij}$$

Recommended ICC Analysis

- Estimate all three ICCs with 95% CIs
- Test for systematic bias using an F test from the two-way mixed-effects model (fixed bias c_j)
- If ICC(1) ≈ ICC(A,1) ≈ ICC(C,1) and F test is non-significant → report ICC(1) as primary
- If substantial differences exist → report both ICC(A,1) and ICC(C,1) with interpretation of implications for absolute agreement vs. consistency
- Interpretation thresholds (acceptable values depend on context of use):
  - ICC > 0.75 - good
  - ICC ≤ 0.50 - poor

Sleepiz One+ WASO - ICC

- Average-score ICCs calculated for weekly-average sleep parameters
- F-test for systematic bias was non-significant in all but one case. As such, estimates of ICC(k) are focused on

Stability definition	Weeks 6 and 7		Weeks 8 and 9	
	N	ICC(k) (95% CI)	N	ICC(k) (95% CI)
PGI-S time spent awake	38	0.911 (0.836, 0.953)	30	0.941 (0.880, 0.971)
Sleep Diary sWASO	44	0.959 (0.926, 0.977)	48	0.956 (0.923, 0.975)
PROMIS SD SF 8b trouble staying asleep	34	0.950 (0.902, 0.975)	41	0.932 (0.876, 0.963)

PGI-S, patient global impression of severity; sWASO, subject-reported wakefulness after sleep onset; PROMIS SD SF 8b, Patient Reported Outcomes Measurement Information System Sleep Disturbance Short Form 8b

Measurement Error

- Standard Error of Measurement (SEM) — expressed in measurement units, inversely related to reliability
  - $SEM_{agreement} = \sqrt{\sigma_c^2 + \sigma_v^2}$
  - $SEM_{consistency} = \sqrt{\sigma_v^2}$
  - Alternative formulation: $SEM_{consistency} = SD_{total} \times \sqrt{1 - ICC}$
- SEM represents the SD of measurement error around an individual score
- Useful to determine minimal detectable change (MDC), representing measurement error around an individual's change in score
  - $MDC_{80} = 1.282 * \sqrt{2} * SEM$

Intersection of reliability and MCID

Sleepiz One+ WASO – SEM and MDC

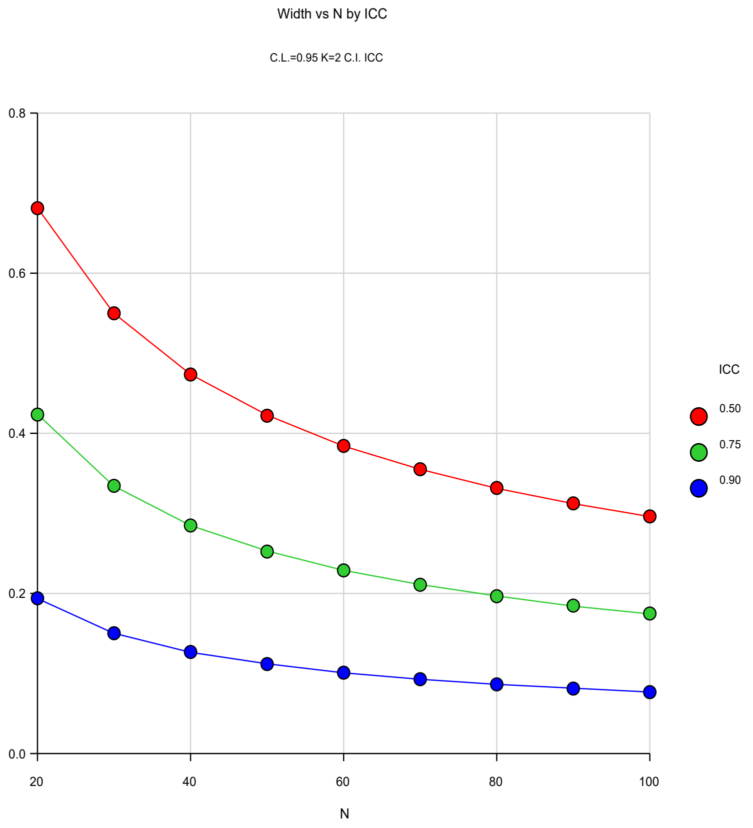
Stability definition	Visits	SEM type	N	SEM value	MDC80	MDC90
PGI-S time spent awake	Week 6 to 7	Agreement	38	6.8	12.4	15.9
	Week 8 to 9	Agreement	30	9.1	16.4	21.1
Sleep Diary sWASO	Week 6 to 7	Agreement	44	6.4	11.7	15.0
	Week 8 to 9	Agreement	48	8.0	14.6	18.7
PROMIS SD SF 8b trouble staying asleep	Week 6 to 7	Agreement	34	6.7	12.1	15.5
	Week 8 to 9	Agreement	41	8.3	15.1	19.4

MDC80 has a critical value of 1.282, MDC90 has a critical value of 1.645. MDC = critical_value * sqrt(2) * SEM

- In a separate qualitative patient interview study, most considered a 30-minute reduction as meaningful
- Highly likely that a change of 30 mins reflects true change rather than measurement error (as MDC values <30)
  - Smaller changes (e.g., 5 or 10 minutes) could be due to measurement error rather than real change – difficult to accurately define responders based on such small thresholds

Sample Size Planning

- Commonly reliability is assessed as a secondary analysis of trial data
  - e.g. Sleepiz One+ NIRVANA trial was powered for primary endpoint
- Even in a standalone validation study, other analyses (responsiveness, MCID) exist with no single 'primary' parameter
- Therefore, rather than statistical power to reject a null hypothesis related to ICC, we aim to estimate parameters of interest with sufficient precision
- Useful to see confidence interval width across a range of ICC magnitudes and sample sizes



Summary of Recommendations

- Follow the V3+ framework: Ensure adequate verification and analytical validation before embarking on clinical validation
- Reporting for continuous measures:
  - Report ICC(1), ICC(A,1), ICC(C,1) with 95% CIs and the F test for systematic bias
  - Bland-Altman plots can help visualize agreement and detect systematic patterns
- Assess sample size: Available closed-form formulas, `ICC.Sample.Size`, PASS
- If reliability is insufficient: Consider averaging over multiple values to reduce error variance, improve measurement protocols

For categorical measures and composite/hybrid measures – see Ratitch 2023

For repeated measures ICC (same device at multiple visits) – see Parker 2020

https://doi.org/10.1159/000531054
https://doi.org/10.1186/s12874-020-01022-x

Q&A

Thank you