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]

[a]Statistics and Data Insights, Bayer Inc., Mississauga, ON, Canada; [b]Medical Affairs Statistics, Bayer plc, Reading, UK; [c]Statistics and Data Insights, Bayer Corporation, Whippany, NJ, USA; [d]Statistics and Data Insights, Bayer AG, Wuppertal, Germany; [e]Clinical Development and Operations, Bayer AG, Berlin, Germany

Keywords
Digital health technology · Clinical validation · Reliability · Statistical methods

Abstract

Background: Assessment of reliability is one of the key components of the validation process designed to demonstrate that a novel clinical measure assessed by a digital health technology tool is fit-for-purpose in clinical research, care, and decision-making. Reliability assessment contributes to characterization of the signal-to-noise ratio and measurement error and is the first indicator of potential usefulness of the proposed clinical measure. ***Summary:*** Methodologies for reliability analyses are scattered across literature on validation of PROs, wet biomarkers, etc., yet are equally useful for digital clinical measures. We review a general modeling framework and statistical metrics typically used for reliability assessments as part of the clinical validation. We also present methods for the assessment of agreement and measurement error, alongside modified approaches for categorical measures. We illustrate the discussed techniques using physical activity data from a wearable device with an accelerometer sensor collected in clinical trial participants. ***Key Messages:*** This paper provides statisticians and data scientists, involved in development and validation of novel digital clinical measures, an overview of the statistical methodologies and analytical tools for reliability assessment.

Introduction

Clinical validation, together with verification and analytical validation, is an integral part of the validation process designed to demonstrate that a novel clinical measure assessed by a digital health technology (DHT) tool is fit-for-purpose in clinical research, care, and decision-making [1, 2]. A previous publication [3] discussed various considerations for planning statistical analyses in support of clinical validation of electronic clinical outcome assessments (eCOAs) and digital biomarkers derived from sensor-based biometric monitoring technology data. In this paper, we focus on statistical methods for evaluation of one of the key elements of clinical validation – reliability.

Assessment of reliability should be performed early in the clinical validation process as it determines potential

karger@karger.com
www.karger.com/dib

Karger
OPEN ACCESS

Correspondence to:
Bohdana Ratitch, Bohdana.ratitch@bayer.com

usefulness of the novel clinical measure and informs interpretations of other analyses. Notably, when interpreting differences or changes in measurement values, it is necessary to ensure that such differences are larger than measurement error before they can be considered meaningful. Reliability assessment contributes to characterization of the signal-to-noise ratio and measurement error.

The objective of this paper was to provide statisticians and data scientists, involved in development and validation of novel digital clinical measures, an overview of the methodologies and analytical tools for reliability assessment. Such methodologies are scattered across literature on validation of PROs, wet biomarkers, etc., yet are equally useful for eCOAs and digital biomarkers measured by sensor-based DHTs.

Reliability of digital clinical measures may be affected by multiple sources of error and variability, which we discuss in the next section. We review a general modeling framework and statistical metrics typically used for reliability assessments of continuous and categorical clinical measures. We illustrate the discussed techniques using data from a wearable device with an accelerometer sensor collected in clinical trial participants with heart failure (HF) disease (the analysis results reported in this paper are intended for illustration purposes only and are not intended to draw practical conclusions about reliability of a specific DHT tool or clinical measure).

Background

Reliability, in the most general terms, refers to the degree to which the results obtained by a measurement procedure can be replicated [4]. Formally, it is defined as the proportion of variance in a measure attributable to true variance in the underlying concept being measured [5]. Other terms, used interchangeably, include repeatability and reproducibility.

Borrowing general terminology regarding reliability from non-digital clinical measures [6], reliability can be split into three types:

- *Intra-rater reliability* generally refers to reproducibility of measurements produced by the same tool or human rater. In the context of sensor-based DHTs, it would refer to reproducibility of measurements produced by the same piece of equipment when applied on the same individual under identical conditions on different occasions.
- *Inter-rater reliability* generally refers to the reproducibility of measurements produced by different pieces of equipment of the same kind or by different, but equally qualified, human raters when assessed on the same individual under identical conditions.
- *Internal consistency reliability* generally refers to the reproducibility of measurements produced by different items of an eCOA, where the responses to these items are combined to yield a score.

Key factors typically impacting reliability are:

- *Analytical variability* refers to variation that may be introduced by the algorithm component of the measurement process, i.e., the ensemble of data transformation steps and clinical measure derivation process. This source of variability in the context of DHTs is mainly applicable if the algorithm involves any stochastic components.
- *Intra-subject variability* reflects sources of random variation in the individual’s physiology, behavior, or environmental factors while the individual remains in a stable state with respect to the target concept of interest (COI). For example, physical activity of a patient in a stable disease state may vary depending on the day of the week, season, and weather.
- *Inter-subject variability* reflects factors that may vary among individuals with the same state of disease and may affect the target COI. These factors may include genetics, demographics, comorbidities, lifestyle, environmental factors, etc. For example, physical activity patterns of patients with the same disease state depend on age and socioeconomic conditions.

When validating DHT tools according to the V3 framework [2], the intra-rater and inter-rater reliability are likely to be evaluated as part of the verification stage at the level of raw signal or after some basic data preprocessing transformations. Analytical variability is expected to be evaluated during the analytical validation step and can additionally be evaluated as part of the clinical validation. During clinical validation, intra-rater and inter-rater reliability is primarily assessed in terms of the intra-subject and inter-subject variability, as well as the total variance of the digital measurements comprising all sources outlined above. Reliability analyses at this stage are applied at the level of the clinical measure [7] which may be an aggregate value derived from multiple granular parameters [3]. For example, a digital clinical measure of physical function may be defined as the average daily number of minutes of sedentary behavior calculated based on data from a wearable device with an accelerometer collected over a 7-day interval.

Reliability can be assessed based on data from a repeated-measure design, where measurements are collected from each participant multiple times under conditions that reflect both natural variability of the target outcome within and between individuals and an intrinsic measurement error. Multiple measurements should be taken over a period of time where patient's disease status is stable with respect to the target COI. Stability can either be assumed (e.g., a short time interval in a relatively stable disease) or confirmed through selecting a subset of individuals with the absence of change on a concurrent measure (e.g., subjects with no change in clinician-rated performance status). Conditions under which measurements are obtained may need to span several aspects. For example, in the case of a sedentary behavior measure, measurements should be obtained over two or more weeks and include both work and weekend days to cover day-to-day variability (although importance of a weekend effect may depend on the target population, e.g., employment likelihood based on age and health status). At the same time, participants with different disease severities should be included in the study. We will further explain this point in the context of a statistical model from which reliability metrics are estimated.

Table 1 summarizes the above discussion on the types of reliability and its evaluation in the context of sensor-based DHTs. Data for reliability assessment should represent the intended context of use. This includes the target population as well as the measurement modality and environment. For example, reliability of a measure of sedentary behavior may be quite different in young adults versus elderly, even when measured with the same DHT, due to different patterns of mobility. Similarly, measures taken in a controlled laboratory or clinic environment may have different properties from those taken in free-living conditions. Even sensor placement location on the body may significantly affect measurement properties (see, e.g., [8]).

Example Dataset

To illustrate the application of statistical methods for reliability analysis that will be discussed in this paper, we use data from two clinical trials in chronic HF: the study NCT02992288 enrolled HF patients with reduced ejection fraction (HFrEF) [9, 10] and the study NCT03098979 enrolled HF patients with preserved ejection fraction (HFpEF) [9, 11]. A wireless cardiac monitoring chest patch device (AVIVO Mobile Patient Management System) was used primarily for ECG monitoring in both trials but also continuously recorded activity data using a triaxial accelerometer sensor measuring acceleration of body movement every 4 s. These high-resolution measurements can be used to derive various measures of physical activity, which are of interest for the chronic HF population because chronic HF patients often present clinically with important physical activity capacity limitations that affect their quality of life. Trial participants were wearing the device during four 7-day periods throughout the study: during the run-in period prior to visit 1 (randomization), 7 days after randomization prior to visit 2, and at weeks 8 and 19 of the trial.

The purpose of this example was not to formally validate or to advocate for use of any specific measure of physical activity in the HF population. This example is purely illustrative with respect to statistical analyses of reliability, and for that purpose we focus on a measure of total activity over 10 most active consecutive minutes in 24 h (10MACM). The most active 10 min in a day captures peak activity levels and is assumed to be related to the COI of physical activity capacity. It is conceptually similar to the six-minute walking distance (6MWD) [12], which is often seen as a gold standard for in-clinic assessments of physical capacity in HF trials. The 6MWD was the primary endpoint in the study of HFpEF patients.

"Total Activity," reflecting the activity volume or magnitude over a time period can be calculated using various so-called activity indices (see, e.g., [13]). An activity index summarizes the high-resolution acceleration measurements over short time intervals (epochs). We use one-minute epochs. We calculate the 10MACM in two ways based on two different activity indices to illustrate that the reliability of an aggregate measure can vary depending on the method used to process the granular data. One 10MACM variant (abbreviated as 10MACM_AAI) is calculated using a proprietary Avivo Activity Index (AAI) provided by the device vendor. The second variant (abbreviated as 10MACM_ASV) is calculated using an activity index defined as the square root of the sum of accelerometer signal variances (ASV), $\sqrt{var(\boldsymbol{x}) + var(\boldsymbol{y}) + var(\boldsymbol{z})}$, where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are vectors of acceleration values along the X, Y, and Z axes, respectively, captured every 4 s within a one-minute time epoch. This index is similar to the activity index proposed in [13].

We first calculate the 10MACM for each assessment day by calculating a sum of the one-minute activity index values over 10-min sliding windows across 24 h and selecting a 10-min window with the largest total activity value for the day. The daily 10MACM values are then averaged over each 7-day measurement period to obtain the week-level aggregate measures representing an average 10MACM. Daily values were calculated if the

Table 1. Types of reliability in the sensor-based DHT context

Type of reliability in the sensor-based DHT context	Assessment
Intra-rater reliability: Reproducibility of measurements produced by the same piece of DHT equipment when applied on the same individual under identical conditions on different occasions	Assessed using repeated measurements from each participant of a reliability study using the same piece of DHT equipment under identical conditions • During verification stage of V3: assessed at the level of sensor measurements • During analytical validation stage of V3: assessed at the level of features and parameters produced after data processing; relevant mostly when data processing involves stochastic components • During clinical validation stage of V3: assessed at the level of a digital clinical measure implicitly as part of evaluation of all sources of intra-subject variability. Intra-rater reliability over time (i.e., test-retest reliability) of particular relevance for DHT tools used to form "change from baseline" endpoints
Inter-rater reliability: Reproducibility of measurements produced by different pieces of equipment of the same kind when assessed on the same individual under identical conditions	Assessed using repeated measurements from each participant of a reliability study using different pieces of equipment of the same kind under identical conditions • During verification stage of V3: assessed at the level of sensor measurements • During analytical validation stage of V3: assessed at the level of features and parameters produced after data processing; relevant mostly when data processing involves stochastic components or when data processing algorithms may be sensitive to differences in sensor measurements by different pieces of equipment; relevant for comparing different DHT tools used for the same purpose • During clinical validation stage of V3: assessed at the level of a digital clinical measure implicitly as part of evaluation of all sources of intra-subject variability; most relevant if individuals may use different pieces of equipment over time in the intended context of use
Internal consistency reliability: Reproducibility of measurements produced by different items of a composite measure	Assessed using a measurement of the composite score from each participant of a reliability study, where the repeated measurements are the different items and identical conditions are imposed by completing each item at the same visit • During clinical validation stage of V3: assessed at the level of a composite digital clinical measure (which may combine ePRO and sensor-based measurements); evaluates whether all items contributing to a composite score are related to each other, varying in a consistent manner

participant had at least 19 h of data in a day, and week-level aggregate values were calculated if at least four out of seven daily values were available, including at least one weekend day. Such criteria resulted in discarding only a small amount of data in our dataset given an excellent adherence of participants with the device wear (a patch was attached to the participant's chest for the entirety of each 7-day measurement period, and mainly the first and last days could have less than the required amount of data). In general, different thresholds for the required wear time and data availability can be evaluated as part of reliability analyses.

As stated above, for reliability estimation multiple measurements under stable disease conditions are required. Each study participant contributed two average 10MACM measurements based on each type of activity index: one from the pre-randomization 7-day period (visit 1) and one from the 7 days following randomization (visit 2). During these 2 weeks, study participants can be considered stable. The 2 weeks were either consecutive or very close in time. The study treatment was deemed not to be effective based on the overall study conclusions and neither the experimental treatment nor placebo was

Table 2. Baseline characteristics of study participants included in the example dataset

Variable	HFpEF (*N* = 191)	HFrEF (*N* = 359)
Sex, *n* (%)		
Female	103 (53.9)	58 (16.2)
Male	88 (46.1)	301 (83.8)
NYHA class, *n* (%)		
I/II	141 (73.8)	213 (59.3)
III/IV	50 (26.2)	146 (40.7)
Age		
Mean (SD)	73.9 (8.43)	67.3 (9.91)
Median [min, max]	75.0 [46.0, 93.0]	68.0 [31.0, 88.0]
Baseline BMI		
Mean (SD)	29.3 (5.22)	28.1 (4.64)
Median [min, max]	29.2 [17.5, 44.1]	27.8 [16.6, 39.9]
KCCQ PLS		
Mean (SD)	65.4 (22.2)	64.4 (25.7)
Median [min, max]	66.7 [8.33, 100]	66.7 [0, 100]

NYHA, New York Heart Association; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; KCCQ PLS, Kansas City Cardiomyopathy Questionnaire Physical Limitations Score; BMI, body mass index.

expected to induce any change between the two assessment periods. The stability of the HF disease in each participant included in the analysis dataset was verified by absence of any cardiovascular events over the duration of measurements. For the HFpEF study, participants were further classified as stable if they did not have a meaningful decline in 6MWD (defined as a reduction greater than 30 m) from visit 1 to visit 2 (increases in 6MWD were allowed as some "learning effect" between assessments close in time can be expected). Participants were included in the analysis dataset if they were considered to have stable disease and had two week-level 10MACM values available as described above. There are 191 and 359 participants included from the HFpEF and HFrEF studies, which represent 63% and 84% of all randomized participants in each study, respectively.

As previously mentioned and will be further explained later, it is important to ensure that the reliability analysis dataset includes participants with characteristics that represent well the diversity of the target population in terms of demographic and disease severity factors. Our example dataset has good representation across age, sex, baseline body mass index, New York Heart Association (NYHA) functional class, and Kansas City Cardiomyopathy Questionnaire Physical Limitations Score (KCCQ PLS), which are summarized in Table 2. For both studies, the cohorts of participants included in our example dataset are very similar to the overall cohorts in the original studies, except for a somewhat larger proportion of participants with NYHA class III/IV in our dataset for the HFpEF study. The analyses of reliability will be performed separately for the two studies because the HFpEF and HFrEF populations are considered to be quite different in terms of the target COI. The distributions of 10MACM_AAI and 10MACM_ASV values from visits 1 and 2 included in the analysis dataset are summarized for each study in Figure 1.

Statistical Methods for Reliability Evaluation

Core Reliability Metrics for Continuous Measures

Intraclass Correlation Coefficient

As discussed above, data for reliability analysis should include measurements from n individuals taken on k measurement occasions for each individual, i.e., a $(n \times k)$ data matrix. If the measurement process is expected to have analytical variability, additionally, m replicates of the clinical measure value calculated from the same input data would be required, i.e., a $(m \times n \times k)$ dataset.

A general statistical approach for modeling intra- and inter-subject sources of variability, as well as total variability of continuous clinical measures, relies on a two-way random-effects analysis of variance (ANOVA) model [4, 14]. In the most general case, where analytical variability is present, the following model can be used:

$$y_{ijs} = \mu + r_i + c_j + (rc)_{ij} + v_{ijs}; \quad i = 1, \ldots, n, j = 1, \ldots, k, s = 1, \ldots, m \qquad (1)$$

where $y_{ijs} \sim N(\mu,\sigma^2)$ is a digital clinical measure value for subject i on occasion j and determination s; μ is a population mean; $r_i \sim N\left(0, \sigma_r^2\right)$ is a random effect (offset in mean) for subject i (inter-subject variability); $c_j \sim N\left(0, \sigma_c^2\right)$ is a random effect (offset in mean) for measurement j (intra-subject variability); $(rc)_{ij} \sim N\left(0, \sigma_{rc}^2\right)$ is an offset in mean response for the interaction between subject i and measurement j; $v_{ijs} \sim N\left(0, \sigma_v^2\right)$ is a measurement error.

Inclusion of the interaction term, $(rc)_{ij}$, in the model reflects the assumption that differences among subjects may vary from measurement to measurement. The total variance is, therefore, modeled as follows:

$$\sigma^2 = \sigma_r^2 + \sigma_c^2 + \sigma_{rc}^2 + \sigma_v^2$$

The key results for interpretation from this model are the estimates of variance components $\widehat{\sigma}_r^2$, $\widehat{\sigma}_c^2$, $\widehat{\sigma}_{rc}^2$, and $\widehat{\sigma}_v^2$ and significance of each effect. Ideally, in the case of a

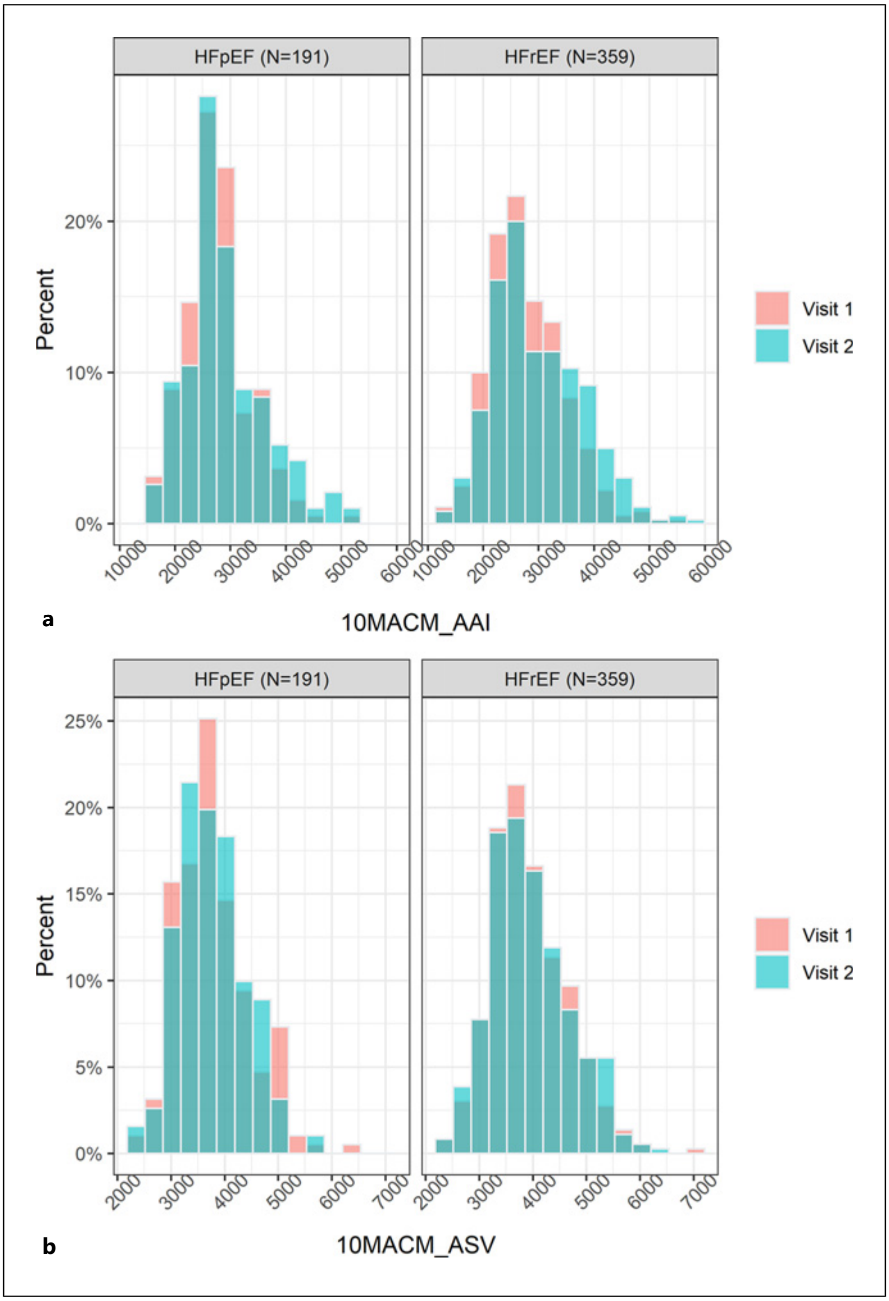

Fig. 1. Histogram plots of the 10MACM_AAI (**a**) and 10MACM_ASV (**b**) values in the analysis dataset.

reliable measure, all effects in the model would not be statistically significant. In addition to characterizing reliability of a proposed clinical measure, understanding sources of variation may be helpful for attempting to modify the definition of the measure, if possible, to improve reliability. For example, if $\hat{\sigma}_c^2$ (intra-subject variability) is relatively large, a different aggregation level may be considered.

For continuous clinical measures, the most widely used reliability metric is the intraclass correlation coefficient (ICC) [15, 16]. Very generally, the ICC represents a ratio

$$\frac{[variance\ of\ interest]}{[total\ variance]} = \frac{[variance\ of\ interest]}{[variance\ of\ interest + unwanted\ variance]}$$

The variance of interest can be thought of as the "true" variance in the COI. The ICC, therefore, reflects a formal definition of reliability as the proportion of variance in a measure attributable to true variance in the COI [5]. When the unwanted variance is equal to or larger than the variance of interest, the reliability of the measurement method is poor (ICC ≤0.5). ICC values >0.75 are generally considered good [17], but an acceptable ICC value depends on the context of use.

As previously mentioned, it is important that the between-subject variability in a reliability study reflects the expected variability in the population of interest (i.e., a full range of disease severity is represented). This is because the ICC is dependent on variance, and an artificially low variability in a sample will generally result in lower ICCs [18, 19]. This may occur, e.g., if data come from a screening phase of a clinical study with eligibility criteria that limit the severity range, e.g., ≥7 on a 0–10 pain scale.

Different types of ICC exist, with various recommendations in the literature on how they should be categorized (see, e.g., [15, 20]). However, all proposed types of ICC are based on modifications to the general ANOVA model described in Eq. (1). In the rest of this paper, we assume that the analytical variability, as defined above, is not present in the evaluated clinical measure and, without loss of generality, drop the index s from the measurement error term ν_{ijs} in model (1).

We summarize here the categorization of ICC by Liljequist et al. [16], as it provides a clear guidance on prespecification and interpretation of analysis. There are three types of ICC.

ICC(1) (also denoted as ICC(1,1)) is applicable when it is assumed there is no systematic error (bias) associated with measurements j. It is derived from a simplification of model (1) to the following form (referred to as *Model 1*, one-way random-effects model):

$$y_{ij} = \mu + r_i + \nu_{ij} \tag{2}$$

where the random effect of measurement is omitted and $\mu + r_i$ represents a "true score" for individual i if there was no measurement error. From this model, the population ICC is defined as follows:

$$ICC(1) = \rho_1 = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\nu^2} \tag{3}$$

ICC(A,1) is a coefficient of absolute agreement in presence of measurement bias. This type of ICC is relevant when an absolute agreement between measurements is desired. It can be defined based on one of two models. *Model 2*, referred to as two-way random-effects model, has the following form:

$$y_{ij} = \mu + r_i + c_j + \nu_{ij} \tag{4}$$

where contribution of $(rc)_{ij}$ and ε_{ij} are combined so that $\sigma_\nu^2 = \sigma_{rc}^2 + \sigma_\varepsilon^2$. Here, c_j represents a systematic but random measurement bias that follows a normal distribution $c_j \sim N\left(0, \sigma_c^2\right)$. In other words, all measurements at occasion j are affected by a common bias c_j, e.g., weather conditions affecting physical activity of all study participants at time j, but these biases are random, i.e., if a new study is conducted, biases c_j will generally not be the same. Based on this model, the population ICC is defined as

$$ICC(A,1) = \rho_{2A} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_c^2 + \sigma_\nu^2} \tag{5}$$

Alternatively, *Model 3*, referred to as two-way mixed effects model, can be used. It has a similar form as (4), but the effects c_j are considered to be fixed instead of random (i.e., bias of the j th measurement occasion remains the same when applied to a new set of individuals). For example, this may occur if there is a consistent practice or motivational effect associated with the measurement process, where individuals may behave somewhat differently on repeated measurement occasions. From this model, the population ICC is defined as

$$ICC(A,1) = \rho_{3A} = \frac{\sigma_r^2}{\sigma_r^2 + \theta_c^2 + \sigma_\nu^2} \tag{6}$$

where $\theta_c^2 = \frac{\sum_j (c_j - \bar{c})^2}{k-1}$, where $\bar{c}$ is the mean of c_j, $j = 1,\ldots,k$. For the remainder of this article, we calculate ICC(A,1) using Model 2.

ICC(C,1) is a coefficient of consistency in presence of measurement bias. It is relevant when measurement bias, represented by c_j, is present but considered acceptable, in a sense that the measurement method produces a consistent ranking order of the individuals, and the main focus is on the between-subject differences. This type of ICC is defined based on the Model 2 or 3 in an identical way as it ignores the bias in measurements. The population ICC is defined as

$$ICC(C,1) = \rho_{2C} = \rho_{3C} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_\nu^2} \tag{7}$$

All three types of ICC can be estimated from the estimates of the mixed model variance components or using mean squares from the ANOVA applied to a $(n \times k)$ data matrix [16].

It has been recommended in the past that the Model 1, 2, or 3 and the type of ICC be prespecified in advance to reflect initial assumptions [17]. Model 1 would be prespecified based on an assumption of no systematic bias, Model 2 would be prespecified based on an assumption of systematic bias that is random, and Model 3 would be prespecified based on an assumption of systematic bias that is fixed. However, Liljequist et al. [16] showed that this is not necessary, and may, in fact, be undesirable. Even if a priori, it is assumed that there is no systematic measurement bias (and therefore, ICC(1) from Model 1 is of primary interest), the existence of such bias should be verified from data. The existence of systematic biases c_j can be tested using an *F* test based on Model 3. Additionally, it is argued [16] that there is little practical difference between the use of Model 2 versus Model 3.

Liljequist et al. [16] also showed the relationship between ICC(1), ICC(A,1), and ICC(C,1) in presence and absence of the measurement bias and recommended the following analysis and reporting procedure.

- All three ICC versions should be estimated from data.
- A close agreement between the three estimated values would be an indicator of the absence of systematic measurement biases. If ICC(1) ≈ ICC(A,1) ≈ ICC(C,1) and if the *F* test indicates that the systematic biases are negligible, the ICC(1) and its confidence interval (CI) may be reported as the primary measure of reliability.
- Substantial differences between ICC(1), ICC(A,1), and ICC(C,1) indicate presence of non-negligible bias, i.e., systematic differences between different measurement occasions. In this case, ICC(1) is not an appropriate measure, and ICC(A,1) and ICC(C,1) should both be reported together with their CIs. They provide complementary information regarding absolute agreement and consistency reliability. If ICC(A,1) is good and randomly varying biases are not too large, the measurement method may still be useable depending on the context of use. If ICC(A,1) is poor or moderate but ICC(C,1) is good or excellent, the measurement method may still be used to rank subjects or compare groups with respect to their measurements on the same measurement occasion, but the measurement method may not be suitable for within-subject differences of values taken at different occasions. The latter corresponds to a scenario where the clinical measure is intended to be used in a clinical trial as an endpoint defined as the change from baseline to some posttreatment time point.

We illustrate the estimation of the ICCs using data from the two studies described above. Code excerpts using the psych and lme4 R packages are provided as online supplementary Material (for all online suppl. material, see https://doi.org/10.1159/000531054). Table 3 presents ICC(1), ICC(A,1), and ICC(C,1), together with their 95% CIs, for each study and each of the considered digital clinical measures (see Example Dataset section for description of the measures). As a point of reference, we also present the estimated ICCs for the 6MWD measure in the HFpEF study, noting that absence of meaningful deterioration in 6MWD was used to define the stable sample. For all measures, the three ICCs are very similar and even appear identical when rounded to the second decimal digit. This indicates negligible systematic differences between measurements during the screening and baseline weeks. The *p* values from the *F* test support this observation for the 10MACM_ASV measure in both studies. The *p* values from the *F* test are statistically significant for the 10MACM_AAI measure and 6MWD; however, the estimated variance components $\widehat{\sigma}_c^2$ are still very small in magnitude (comprising 1.3–3.2% of the total variance). Therefore, we can conclude the ICC(1) coefficient would be appropriate as the primary ICC metric in this case to enable comparison between the different indices. We can see that the two digital clinical measures, targeting the same concept but based on different granular indices of accelerometer signals, AAI and ASV, exhibit quite different ICCs: the former has an ICC below the level typically considered good (the point estimates and upper confidence limits <0.75), whereas the latter has an ICC that can be interpreted as good (the point estimates and the lower confidence limits >0.75). The ASV-based measure's ICC is slightly smaller but close to that of 6MWD in the HFpEF study (ICC(1) of 0.88 versus. 0.91, respectively). The difference between the estimated ICCs for both digital measures between the two studies is small and the estimates are consistent in the two respective populations. The ICC estimates and their 95% CIs are also summarized graphically in Figure 2.

Measures of Agreement

Reliability and agreement are sometimes used interchangeably, and, as we will discuss in the next section, some agreement metrics are often used to assess reliability of categorical clinical measures. However, reliability and agreement generally target different questions. As discussed in [6, 21], agreement assesses how close the repeated measurements are to each other, whereas reliability assesses how well the individuals can be distinguished from each other, despite measurement errors. Both questions are important, and agreement should be assessed alongside reliability.

Bland-Altman plots [22] are a widely used tool for assessing the extent of agreement across differing levels of the measurement and are presented below. It is helpful to examine data in a graphical manner for presence of systematic errors, e.g., systematic patterns of differences between two measurements. The Bland-Altman plot is a two-dimensional scatterplot, where each point corresponds to one subject in the reliability study: the x-axis represents the mean of the two measurements and the y-axis – the difference between the two measurements of

Table 3. ICCs estimated from example datasets

Study and digital clinical measure	ICC(1) (95% CI)	ICC(A,1) (95% CI)	ICC(C,1) (95% CI)	*F* test *p* value
HFrEF				
10MACM_AAI	0.64 (0.58, 0.70)	0.65 (0.56, 0.72)	0.67 (0.61,0.72)	<0.0001
10MACM_ASV	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.89 (0.87, 0.91)	0.3037
HFpEF				
10MACM_AAI	0.64 (0.54, 0.71)	0.64 (0.54, 0.72)	0.65 (0.56, 0.73)	0.0007
10MACM_ASV	0.88 (0.84, 0.91)	0.88 (0.84, 0.91)	0.88 (0.84, 0.91)	0.4090
6MWD	0.91 (0.88, 0.93)	0.91 (0.85, 0.94)	0.92 (0.89, 0.94)	<0.0001

CI, confidence interval; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; ICC, intraclass correlation coefficient; 10MACM_AAI, 10 most active continuous minutes based on AAI index; 10MACM_ASV, 10 most active continuous minutes based on ASV index; 6MWD, six-minute walking distance.

each subject. The dotted horizontal line at $\overline{d}$ in the middle represents the mean of the pairwise difference, and the two dashed lines at the top and bottom of the graph represent the limits of agreement drawn at $\left(\overline{d} \pm 1.96 \times \mathrm{SD}_{\mathrm{diff}}\right)$, where $\mathrm{SD}_{\mathrm{diff}}$ is the standard deviation of pairwise differences. The value of $\overline{d}$ is interpreted as the systematic error and $1.96 \times \mathrm{SD}_{\mathrm{diff}}$ as the random error. Under the assumption of normal distribution for the pairwise differences, this would mean that 95% of pairs agree within the corresponding limits. These limits are expressed in the units of measurement and can be interpreted from the point of view of clinical relevance, provided that there is good clinical understanding of the range of the measurement values.

The limits of agreement are drawn as horizontal lines under the assumption that pairwise differences do not change depending on the magnitude of the measurements (which is also assumed by the one-way random effects Model 1 and the resulting ICC(1) discussed above). However, violations of this assumption can be observed on the Bland-Altman plot as systematic patterns in the distribution of the plotted dots, e.g., a dot cloud that widens with increasing values on the x-axis.

Figure 3 presents Bland-Altman plots for the two digital clinical measures in both studies. We can observe that the mean pairwise differences between the two assessment visits (dotted $\overline{d}$ lines) are somewhat more away from zero for the 10MACM_AAI measure compared to the 10MACM_ASV measure in both studies, which indicates larger systematic differences between assessment weeks for the former measure and adds context to the results of the *F* test discussed above. For the 10MACM_AAI measure, there are more increases in activity from the first to second week (negative pairwise differences), with most of the extreme differences being below the lower limit of agreement (lower dashed line). The pairwise differences appear to be smaller for smaller values of 10MACM_AAI and increase for values 10MACM_AAI >25,000. The pairwise differences for the 10MACM_ASV measure are distributed more uniformly randomly around zero of the y-axis as well as across the range of the measurement values. We can see one extreme pairwise difference for the 10MACM_ASV measure in the HFpEF study (shown in Fig. 3d), but otherwise, the differences that are outside of the limits of agreement (dashed lines) do not fall very far from those limits. In our example, the observations from the Bland-Altman plots would reinforce the conclusion from the ICC analysis suggesting that the ASV-based measure has better reliability properties compared to the AAI-based measure in both studies.

If the values of the digital clinical measure were directly interpretable, the tightness of the limits of agreement as well as the spread of values beyond them could be interpreted in terms of their clinical relevance. Also, in addition to the limits of agreement as defined above, one can prespecify clinically relevant ranges of agreement and report the proportion of individuals in the study with agreement of their pairwise measurements within the prespecified ranges. For example, for the 10MACM_AAI measure, 82% and 85% of participants' pairs of measurements agree within the [−5,000, 5,000] range for the HFrEF and HFpEF studies, respectively, and for the 10MACM_ASV measure, 84% and 93% of repeated measurements agree within the [−500, 500] range for the HFrEF and HFpEF studies, respectively.

Core Reliability Metrics for Categorical Measures

Kappa and Agreement Statistics

For reliability assessment of categorical clinical measures, a variety of Kappa statistics are recommended [6], such as Cohen's Kappa, prevalence-adjusted and bias-adjusted

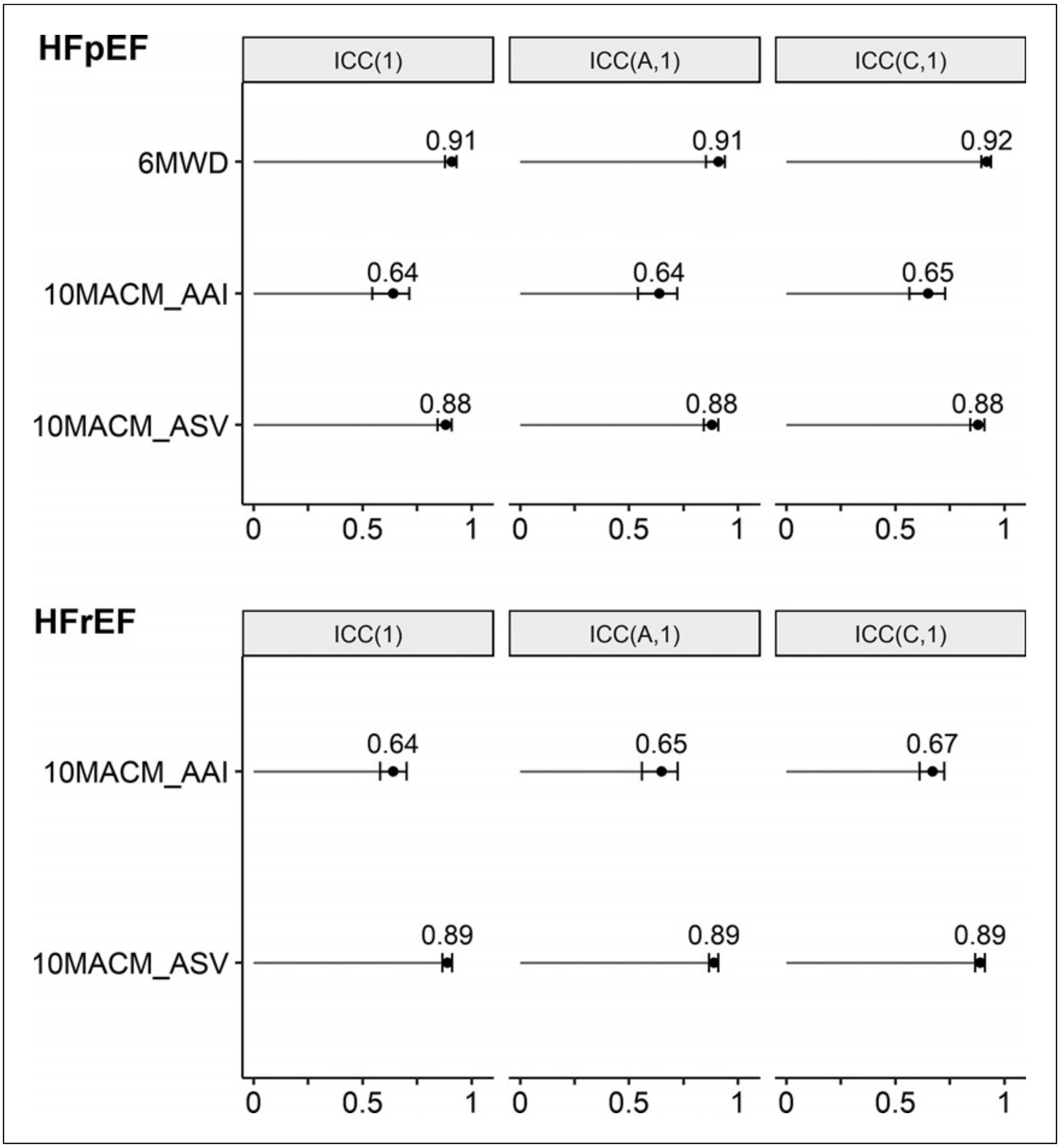

Fig. 2. Plots of ICCs estimated from example datasets.

Kappa (PABAK) or Kappa with Cicchetti weighting [18, 23] for dichotomous and ordinal scales (see, e.g., [4, 18]). While Kappa statistics are often interchangeably referred to as measures of agreement or reliability [24], we choose to align with [6] and present Kappa as a reliability measure.

The quantities used in the definition of Kappa statistics can be represented in a 2 × 2 table with the number of pairs (categorical values at visit 1 and visit 2 for the same participant) that agree or disagree (see Table 4). Cohen's Kappa is defined as follows:

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{8}$$

where $p_0=(a+d)/n$ and $n = a+b+c+d$ is the total number of paired observations from n subjects. The proportion p_0 is commonly known as the "Index of Crude Agreement" and may be interpreted as the observed proportionate agreement. On the other hand, p_e is the proportionate agreement between the two visits that can be attributed to chance, that is, sum of expected probabilities that both visits would be either high or low at random. It is calculated as:

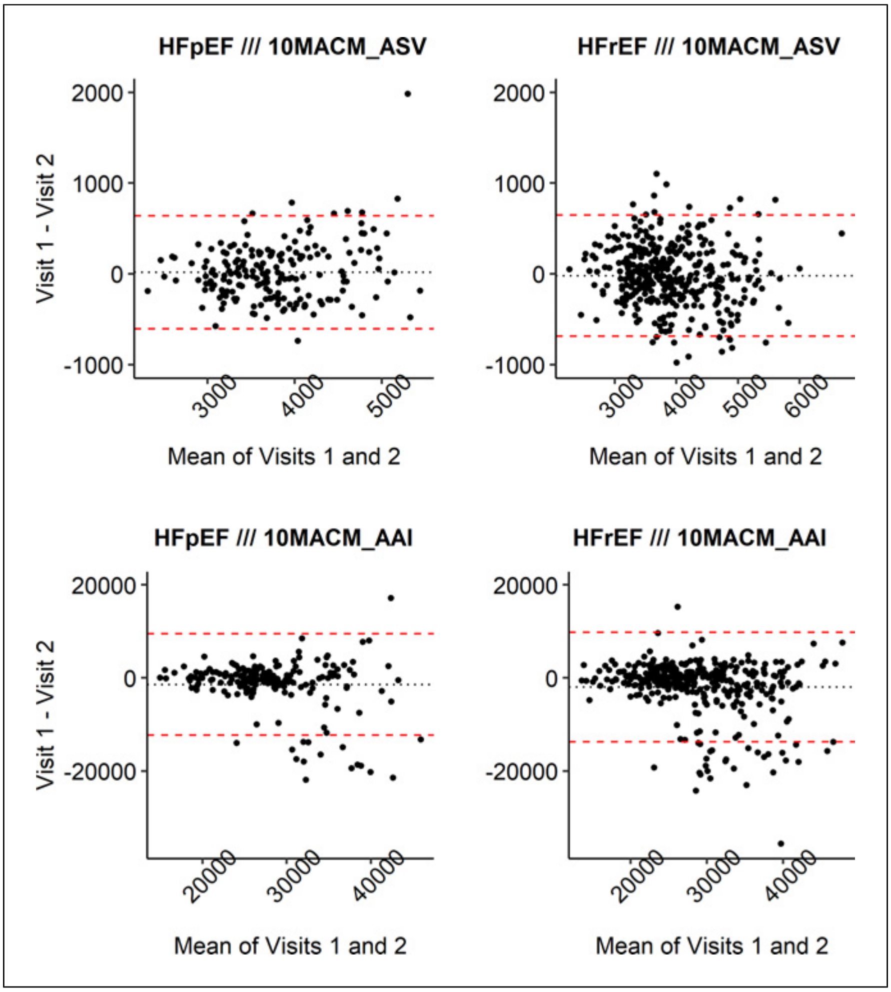

Fig. 3. Bland-Altman plots for digital clinical measures from example datasets.

Table 4. Number of pairs with agreement and disagreement

Visit 1:	High	Low
Visit 2		
High	a	b
Low	c	d

$$p_e = p_{high} + p_{low} = \frac{a+b}{a+b+c+d} \times \frac{a+c}{a+b+c+d} + \frac{c+d}{a+b+c+d} \times \frac{b+d}{a+b+c+d} \quad (9)$$

The standard error of the estimated κ is calculated using a correction to Cohen's original formula [25] and CIs are then obtained based on a normal approximation (i.e., by multiplying the standard error by 1.96 and adding/subtracting from κ to derive a 95% CI). Guidelines for the interpretation of κ are: >0.75 is considered excellent, 0.40–0.75 – fair to good, <0.40 – poor [26].

Although Cohen's Kappa is a generally accepted method for assessing reliability of dichotomous variables, it has its limitations [27] as it is affected by the degree of asymmetry or imbalance in the agreement/disagreement table. Values of κ can be very different across cases where the proportion of agreement is the same because they are influenced by the total percentage of positives. Also, high agreement can occur even if κ is very low. The PABAK statistic mentioned above is designed to overcome this limitation.

The index of crude agreement, p_0, calculated as part of Cohen's Kappa can itself be used as a measure of agreement for dichotomous variables.

Table 5. Reliability (κ and ICC) and agreement (p_0) measures for binary variables estimated from example dataset

Study and digital clinical measure	κ (95% CI)	p_0 (95% CI)	ICC (95% CI)
HFrEF			
10MACM_AAI	0.69 (0.62, 0.77)	0.85 (0.81, 0.88)	0.79 (0.70, 0.86)
10MACM_ASV	0.71 (0.63, 0.78)	0.85 (0.82, 0.89)	0.80 (0.72, 0.87)
HFpEF			
10MACM_AAI	0.72 (0.62, 0.82)	0.86 (0.81, 0.91)	0.81 (0.69, 0.89)
10MACM_ASV	0.64 (0.54, 0.75)	0.82 (0.77, 0.88)	0.74 (0.62, 0.81)

HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; 10MACM_AAI, 10 most active continuous minutes based on AAI index; 10MACM_ASV, 10 most active continuous minutes based on ASV index.

The corresponding standard error can be calculated as follows and CIs can be calculated based on a normal approximation:

$$\mathrm{SE}\ (p_0) = \sqrt{\frac{p_0(1-p_0)}{n}} \qquad (10)$$

For illustration with our example dataset, we dichotomize the measures 10MACM_AAI and 10MACM_ASV at the median value of the first visit for each variable (high = above median; low = equal to or below median) separately in each of the two studies described above. Table 5 presents estimated κ and p_0, along with their 95% CIs, for each study and digital clinical measure. Both κ and p_0 take values that indicate a good level of reliability and agreement, respectively. The values, including CIs, for p_0 are numerically higher than for κ due to chance agreement not being accounted for. When analyzing these digital measures as continuous variables in the previous section, reliability was consistently higher for 10MACM_ASV than 10MACM_AAI (Table 3); however, after dichotomizing this is no longer the case (Table 5). Dichotomization generally leads to loss of information, and it is recommended to use the underlying continuous values where possible to avoid such loss [28]. However, if a categorized version of the clinical measure is of primary interest for the intended context of use, then it would be useful to perform reliability evaluations both at the underlying continuous level and categorical level and assess any ensuing loss of reliability.

Intraclass Correlation Coefficient

Unfortunately, the above statistics do not provide the ability to investigate factors affecting reliability and, therefore, the ICC was suggested as a preferred metric [29]. One approach to derive ICCs for dichotomous or ordinal measurements is through fitting a generalized linear mixed model (GLMM) with a random effect for subject. Dichotomous variables can be modeled using a logit link function as follows:

$$\log\frac{p_{ij}}{1-p_{ij}} = \beta x_{ij} + b_i$$

$$Y_{ij}\big|p_{ij} \sim Bernoulli\big(p_{ij}\big)$$

where p denotes the probability of being in the "High" category, β denotes a fixed effect coefficient acting on a time-varying covariate (in this context, visit), and b denotes a random effect varying between subjects (i.e., a random intercept). The actual observed dichotomous response Y follows a Bernoulli distribution given p. While additional fixed and random effects can be specified, these are omitted here for simplicity. The above model can be alternatively formulated [30–32] in terms of an underlying continuous response, Y*, that is categorized using a threshold:

$$Y = \begin{cases} 0 & if\ Y^* \le \tau_1 \\ 1 & if\ Y^* > \tau_1 \end{cases}$$

A useful property of this formulation is that a linear model can be specified:

$$Y^*_{ij} = \beta x_{ij} + b_i + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim Logistic\,(0,1)$$

As the error is distributed according to the standard logistic function, its variance is not estimated from data but rather fixed at a value of $\pi^2/3$, given that the correlations of the underlying continuous measurements are constant [32]. Therefore, an ICC can be calculated based on the variance of the b parameter as follows for a logit model [33]:

$$\rho_1 = \frac{\sigma_b^2}{\sigma_b^2 + \frac{\pi^2}{3}}$$

Note that this is similar in nature to the ICC presented in Eq. (3). The 95% CIs can be derived by obtaining the upper and lower confidence limits for σ_b^2 through profile likelihood [34] and inputting those into the ICC formula or, alternatively, using bootstrap. This approach can be modified for GLMMs with a probit link function (where error variance is fixed at 1), or extended to ordinal variables, as described in [32]. The above approach is only appropriate if a clinical measure is indeed defined based on a dichotomization of an underlying continuous measurement; otherwise, other methods are available such as model linearization or simulation [33].

For dichotomized measures in our example dataset, the ICCs were derived through GLMMs as described above and are provided in Table 5. The profile likelihood method failed to converge in the estimation of the CI for the 10MACM_ASV in the HFpEF study, so bootstrapping was used instead. The point estimates of the ICCs of the binary measures were "good" (i.e., >0.75) for 10MACM_AAI in both studies and for 10MACM_ASV in the HFrEF study, but all lower confidence limits fell below 0.75 (see Table 5). Estimated ICCs for the two binary measures followed a similar pattern as the estimated κ coefficients but would lead to different conclusions regarding reliabilities compared to the ICCs based on continuous variables (see Table 3).

Reliability of Average-Score Measures

The ICCs discussed above are single-score ICCs (also referred to as single measurement ICCs) [20], where each value used in the analysis represents a single value of the clinical measure, not the average of two or more measurements. A clinical measure may also be average-score, e.g., when a clinical severity score is derived as an average of two clinicians' ratings. However, we do not consider all measures that are calculated as averages of some feature values to be average-score. For example, an average of daily number of minutes of sedentary behavior over a 7-day interval does not represent an average-score measure because the aim is to capture an average pattern of a person's physical activity over a week, which is assumed to vary each day, including on workdays versus weekends. The 7-day average is considered a single representative value, and the use of average-score ICCs would not be appropriate. Examples of average-score measures may include those based on multiple features or sensors, e.g., when the physical activity intensity classification is derived from a combination of heart rate and accelerometry sensors [35], or those where multiple measurements are averaged to reduce measurement error, e.g., averaging values of a sleep quality measure over several nights while sleep quality is assumed to be relatively similar across nights [36]. Another common example is an ePRO score obtained through averaging several item responses; however, in practice single-score ICCs are often calculated for these.

When a clinical measure is designed indeed as an average-score, the reliability should be assessed using average-score ICCs [20]. They are labeled as ICC(k), ICC(A,k), and ICC(C,k) and defined similarly to the single-score ICCs [20], but the unwanted variance in the denominator (e.g., σ_v^2 in Eq. 3) is divided by the number of averaged measurements, thus, increasing reliability with increasing number of averaged scores.

Internal Consistency Reliability

A special form of reliability for multi-item measures is common in validation of eCOAs, e.g., where reliability across the different items of a PRO scale is the focus. It may also be useful for composite hybrid measures, e.g., a composite of ePRO and sensor-based physical activity monitoring.

The error variance arising from the different items is assessed to form an internal consistency reliability coefficient. A common example is Cronbach's alpha [37, 38] (equivalent to ICC[C,k]); however, this method makes a strict assumption that all items are equally informative of the COI, measure a single unidimensional concept, and have uncorrelated errors. Therefore, several other composite reliability measures exist that relax these assumptions in various ways [39]. Internal consistency reliability can also be assessed through item response theory models [40] that allow reliability to vary across the range rather than being fixed for all possible scores [41]. Note that intra-rater and inter-rater reliability should still be assessed for multi-item eCOA scores where possible, in addition to internal consistency. It is common practice to estimate the three types of reliability separately; however, it is possible to simultaneously assess the different types through a method called Generalizability Theory [18, 42]. For example, for the KCCQ questionnaire, published estimates of internal consistency (Cronbach's alpha) range from 0.87 to 0.91 in patients with HF [43, 44], where estimates of at least 0.7 are generally considered satisfactory for making comparisons between groups.

Measurement Error

Measurement error is another important measurement property alongside reliability. It is inversely related to reliability and is commonly represented by the standard error of measurement (SEM). In contrast to ICC, which is a unitless metric ranging from 0 to 1, the SEM is expressed in the same units as the measurement values and is therefore useful to help interpret actual scores. The SEM is the square root of the error variance.

Table 6. SEMs estimated from example dataset

Study and digital clinical measure	SEM$_{\text{consistency}}$	SEM$_{\text{agreement}}$
HFrEF		
10MACM_AAI	4243	4449
10MACM_ASV	240	240
HFpEF		
10MACM_AAI	3938	4047
10MACM_ASV	224	224

HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; SEM, standard error of measurement; 10MACM_AAI, 10 most active continuous minutes based on AAI index; 10MACM_ASV, 10 most active continuous minutes based on ASV index.

As previously discussed, ICC(C,1) is often used as a measure of *consistency reliability*, which characterizes the consistency of the position or rank of individuals in the group relative to others. In contrast, *absolute agreement reliability* characterizes agreement of measurement values themselves (and not just relative ranking) and ICC(A,1) characterizes that aspect. Similarly, two versions of the SEM exist [21]:

$$\text{SEM}_{\text{consistency}} = \sqrt{\sigma_{\text{v}}^2}$$

$$\text{SEM}_{\text{agreement}} = \sqrt{\sigma_{\text{c}}^2 + \sigma_{\text{v}}^2}$$

These two versions differ with respect to whether systematic differences between measurements (e.g., due to time) are considered or not. Both SEM versions can be estimated from the same models as used to estimate the ICCs (see equations 2 and 4). The SEM$_{\text{consistency}}$ (but not SEM$_{\text{agreement}}$) can be expressed alternatively in relationship to ICC as follows:

$$\text{SEM}_{\text{consistency}} = \sqrt{\text{error variance}} = \sqrt{\text{total variance}\,(1 - \text{ICC})}$$

A smaller SEM indicates a higher reliability of the measure. Note that the SEM concerns measurement error surrounding an individual score. While it is possible to estimate SEM for categorical measures using the ICC discussed earlier, this SEM will be on the scale of the underlying continuous response and may be of limited value in practice. For ordinal measures with at least 5 categories, it is reasonable to treat the measure as continuous and estimate the SEM accordingly [45].

Estimated SEMs for our example dataset are presented in Table 6. Note that the scales of the two types of measures are not the same and, therefore, the SEMs of the 10MACM_AAI and 10MACM_ASV measures cannot be directly compared. Some differences between the two study populations can be observed, with somewhat larger SEMs in the HFrEF study compared to the HFpEF study for both measures. The minimal differences between SEM$_{\text{consistency}}$ and SEM$_{\text{agreement}}$ are consistent with the earlier findings for the ICCs.

Considerations for Sample Size of the Reliability Studies

Validation studies designed to support the analysis of reliability need to ensure an adequate sample size [46]. Inadequately sized studies may lead to incorrect conclusions or results that are not informative (e.g., a very wide CI for ICC that does not allow to conclude whether the novel clinical measure meets minimum reliability requirements).

Closed-form formulas constructed using large-sample approximation techniques for calculation of the sample size requirements for estimation of the ICC can be found in the literature [47–49] and are implemented in several statistical software packages, e.g., PASS and R. The assumptions necessary to estimate the required sample size, given the planned number of repeated measurements from each study participant, include the target ICC (assumed population parameter value) and a desired precision of the estimated ICC in terms of either a CI width or assurance probability that the actual CI width will not exceed the planned bound. As illustrated in [48], the sample size requirement increases as the assumed population ICC decreases. Sample size can also be calculated for a hypothesis testing analysis that aims at showing that the ICC is larger than some minimum threshold of interest [49]. The methodology in [50] provides an exact sample size estimation method and incorporates the study cost constraints and determination of optimal number of repeated measurements.

A post hoc power calculation for our example dataset using the approach in [48] and the R package ICC.-Sample.Size shows that the HFpEF data (n = 191) could provide 82.1% power to demonstrate that an ICC of at least 0.64 (the lowest observed estimate) was significantly different from a minimum threshold of 0.50 (i.e., at least a “moderate” ICC; two-tailed test with alpha = 0.05). The HFrEF data (n = 359) had 97.7% power to demonstrate the same effect. In other words, given the assumptions stated above regarding the target ICC and a desired precision, increase in the sample size from 191 to 359 participants would result in an increase in power from 82.1% to 97.7%. When reducing the null hypothesis from 0.5 to 0, power increased to >99.9% for both studies, although the null hypothesis of ICC = 0 has little practical utility.

For a categorical clinical measure where reliability is assessed using a Kappa coefficient, sample size can be calculated as the minimum number of participants required to detect a statistically significant Kappa different from zero [51]. The assumptions needed for the calculation include the value of Kappa under the alternative hypothesis and the proportion of patients in the reference category (sample size requirements are greatest when that proportion is either high or low). Sample size calculation can also be performed to detect a higher degree of agreement that is to test that the Kappa coefficient is statistically significantly larger than a prespecified threshold [52]. A precision-based sample size estimation technique is described in [53] for studies aiming to achieve a prespecified lower and upper limit for a CI for the Kappa coefficient. The abovementioned methods are implemented in several R packages.

Conclusion

Reliability analyses evaluate the extent to which individuals can be distinguished from each other using a particular clinical measure given the magnitude of natural variability in the COI and measurement error. Utility of a novel digital clinical measure further depends on its ability to satisfy a number of other validation requirements [1–3], but assessment of reliability can be viewed as a prerequisite for further investigations of clinical validity. Reliability metrics and SEM can be used to compare several measurement modalities or candidate definitions of a clinical measure in order to select the most reliable in the target population (see, e.g., [8, 36]).

The example of actigraphy-based digital measures of physical activity in HF patients used in this paper was presented for illustration only and not meant to support any conclusions about the validity of the discussed measures. However, using this example we highlighted several aspects that should be considered when performing formal reliability analyses: (a) availability of data from a cohort of patients representative of a well-described target population in terms of important patient and disease characteristics; (b) a digital clinical measure targeting a known COI relevant for drug development in the therapeutic area of interest; (c) availability of sensor data allowing us to capture the desired digital measure; and (d) availability of repeated measurements (at the defined aggregate level of the digital clinical measure) during a period of stable disease and collected in conditions similar to those of the intended context of use (free living in clinical trial participants in our case). Our example illustrated that alternative granular features (two activity indices summarizing the granular sensor data in our case) may capture different aspects of COI and the associated variability and may affect reliability metrics of the aggregated digital measure. We were also able to observe that the reliability is affected when going from an underlying continuous measure to a categorical one (reliability loss may be expected due to information loss).

In general, when validating a novel digital clinical measure for any context of use, assessment of reliability is an essential step that should be carried out with the following recommendations in mind:

- *Follow the V3 validation framework* [2]. When embarking on the clinical validation component, ensure that the DHT chosen as a measurement tool underwent adequate verification and analytical validation.
- *Conduct reliability assessment as the first step of clinical validation.* This provides a characterization of the signal-to-noise ratio and measurement error, which are necessary for subsequent interpretations on other aspects of clinical validation.
- Assess reliability at the level of a *precisely defined* clinical measure as intended to be used in a specific context (e.g., the clinical measure may be an aggregate derived from multiple granular parameters measured over a specified time interval).
- To assess reliability, *design a study* or use an existing data source where:
  - Clinical measure values are collected from each participant at least two times over a period where participant's disease status can be assumed or verified to be stable with respect to the target COI.
  - Repeated measurements are scheduled to minimize practice effects and other avoidable sources of systematic bias; however, if these biases are expected to occur in the intended future context of use, they should be evaluated.
  - Study sample is representative of the target population and covers relevant categories of demographic and disease characteristics, to ensure sufficient between-individual variability.
  - Measurement modality and environment are similar to the future intended context of use.
  - Sample size is adequate for a robust estimation of reliability metrics.

The above aspects of the study design should be clarified in publications or other types of reports on reliability assessment. A more granular checklist for reporting reliability studies is provided by Kottner et al. [6], which we recommend adhering to. In terms of the specific statistical methods, we recommend the following for a thorough assessment of reliability, agreement, and measurement error:

- *Conduct analyses* of reliability using appropriate statistical methods as discussed in this paper and *report estimates of the reliability indices and SEM.*
- For *continuous digital clinical measures*, in line with recommendations by Liljequist et al. [16], report three types of intraclass correlation coefficient, ICC(1), ICC(A,1), and ICC(C,1), their CIs, as well as the *F* test based on Model 3. If there is close agreement between ICCs, interpretation can focus primarily on the ICC(1). Otherwise, the interpretation should rely on the ICC(A,1) and ICC(C,1) which provide complementary information. Report Bland-Altman plots to aid in the interpretation of agreement between repeated measurements and presence of systematic errors. Additionally, report SEM (both the consistency and agreement versions).
- For *categorical digital clinical measures*, report a Kappa statistic (Cohen's or PABAK), as well as the Index of Crude Agreement, with CIs. If a categorical measure is derived based on an underlying continuous measure (or a continuum is hypothesized to exist), also report the ICC derived from a GLMM. For ordinal measures with at least 5 categories, SEM can also be reported.
- For *composite (hybrid) measures* (e.g., composite scores combining ePRO and/or sensor-based data), additionally assess internal consistency reliability.
- Conclusions regarding reliability rest on the interpretation of the magnitude of the estimated reliability metrics, such as the ICC and Kappa coefficients. Some common rules of thumb exist for the interpretation of reliability metrics [17] (e.g., ICC >0.75 is considered good), but acceptable reliability may depend on the intended context of use [16] and the associated risks.
- If reliability is deemed acceptable, other aspects of clinical validity should be evaluated, as appropriate for the intended context of use [3].
- If reliability is not deemed acceptable, examine the extent of systematic biases and if that can be reduced in future applications of the DHT. If error variance is high, consider whether the clinical measure could be designed as an average over multiple values to minimize this. Alternatively, consider if the measurement process can be improved by better training (e.g., in consistent application of sensors, importance of adherence to the wear schedule, etc.). In the case of internal consistency, explore if any of the individual items could be removed to improve reliability.

Finally, it should be noted that reliability in general and its specific metrics are not intrinsic properties of a clinical measure and a measurement method but are also directly associated with the target population and measurement conditions. For example, the Canadian Longitudinal Study on aging evaluated five measures of physical function in community-dwelling Canadians over 50 years old and showed that some clinical measures have substantially different reliability depending on the age group [54]. Moreover, the reliability metrics depend on the composition of the study sample and can be generalized only to samples with a similar between-subject variability driven by relevant demographic and disease characteristics. This underlines the importance of proper representation of the target population in the sample enrolled in the reliability study. In contrast, the SEM characterizes the measurement instrument more directly and is likely to be more similar over different samples. Impact of measurement conditions should also be carefully considered when validating a novel clinical measure. Just like the underlying disease should be stable during the period of time when repeated measurements are taken from the same individual for a reliability study, drastic environmental, and lifestyle changes should be avoided. For example, seasonality has been shown to affect measures of physical activity [55]; therefore, repeated measurements for the reliability assessment should ideally be done within the same season for each participant, while different participants may be assessed across different seasons.

This paper provides an overview of an essential toolbox of statistical methods for reliability assessment that statisticians and data scientists should be equipped with when supporting development and validation of digital clinical measures. Examples of analysis code in R can be found in online supplementary Material.

Acknowledgments

The authors acknowledge Frank Kruesmann for reviewing the manuscript draft, Ciprian Crăiniceanu for insights into the processing of granular accelerometry data, and Francesca Giorgolo and Ilaria Ceppa for work on the data preparation.

Conflict of Interest Statement

All authors are salaried employees of Bayer.

Funding Sources

The authors have no funding sources to disclose.

Author Contributions

Bohdana Ratitch, Andrew Trigg, Madhurima Majumder, Vanja Vlajnic, Nicole Rethemeier, and Richard Nkulikiyinka made substantial contributions to the conception or design of the work, drafted this work and revised it critically for important intellectual content and gave final approval of the version to be published and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. Bohdana Ratitch, Andrew Trigg, Madhurima Majumder, Vanja Vlajnic, and Nicole Rethemeier contributed to the analysis of the data.

References

1 FDA. Digital health technologies for remote data acquisition in clinical investigations. Draft guidance for industry, investigators, and other stakeholders. U.S. Food and Drug Administration; 2021. [cited 2021 December 22]. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/digital-health-technologies-remote-data-acquisition-clinical-investigations.

2 Goldsack JC, Coravos A, Bakker JP, Bent B, Dowling AV, Fitzer-Attas C, et al. Verification, analytical validation, and clinical validation (V3): the foundation of determining fit-for-purpose for Biometric Monitoring Technologies (BioMeTs). NPJ Digit Med. 2020 Apr 14;3(1):55–15.

3 Ratitch B, Rodriguez-Chavez IR, Dabral A, Fontanari A, Vega J, Onorati F, et al. Considerations for analyzing and interpreting data from biometric monitoring technologies in clinical trials. Digit Biomark. 2022;6(3):83–97.

4 Looney SW. Statistical methods for assessing biomarkers. Methods Mol Biol. In: Looney SW, editor. Biostatistical Methods. 2002. 184. p. 81–110.

5 Lord FM, Novick MR. Statistical theory of mental test scores. Reading, MA: Addison-Wesley; 1968.

6 Kottner J, Audigé L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. Int J Nurs Stud. 2011 Jan;48(6):661–71.

7 Manta C, Patrick-Lake B, Goldsack JC. Digital measures that matter to patients: a framework to guide the selection and development of digital measures of health. Digit Biomark. 2020 Sep–Dec;4(3):69–77.

8 Kossi O, Lacroix J, Ferry B, Batcho CS, Julien-Vergonjanne A, Mandigout S. Reliability of ActiGraph GT3X+ placement location in the estimation of energy expenditure during moderate and high-intensity physical activities in young and older adults. J Sports Sci. 2021 Jul;39(13):1489–96.

9 Voors AA, Shah SJ, Bax JJ, Butler J, Gheorghiade M, Hernandez AF, et al. Rationale and design of the phase 2b clinical trials to study the effects of the partial adenosine A1-receptor agonist neladenoson bialanate in patients with chronic heart failure with reduced (PANTHEON) and preserved (PANACHE) ejection fraction. Eur J Heart Fail. 2018 Nov;20(11):1601–10.

10 Voors AA, Bax JJ, Hernandez AF, Wirtz AB, Pap AF, Ferreira AC, et al. Safety and efficacy of the partial adenosine A1 receptor agonist neladenoson bialanate in patients with chronic heart failure with reduced ejection fraction: a phase IIb, randomized, double-blind, placebo-controlled trial. Eur J Heart Fail. 2019 Nov;21(11):1426–33.

11 Shah SJ, Voors AA, McMurray JJV, Kitzman DW, Viethen T, Bomfim Wirtz A, et al. Effect of neladenoson bialanate on exercise capacity among patients with heart failure with preserved ejection fraction: a randomized clinical trial. JAMA. 2019 Jun 4;321(21):2101–12.

12 Butland RJA, Pang J, Gross ER, Woodcock AA, Geddes DM. Two-six-and 12-minute walking tests in respiratory disease. BMJ. 1982;284(6329):1607–8.

13 Bai J, Di C, Xiao L, Evenson KR, LaCroix AZ, Crainiceanu CM, et al. An activity index for raw accelerometry data and its comparison with other activity metrics. PLoS One. 2016 Aug 11;11(8):e0160644.

14 Taioli E, Kinney P, Zhitkovich A, Fulton H, Voitkun V, Cosma G, et al. Application of reliability models to studies of biomarker validation. Environ Health Perspect. 1994 Mar;102(3):306–9.

15 Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979 Mar;86(2):420–8.

16 Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation: a discussion and demonstration of basic features. PLoS One. 2019 Jul 22;14(7):e0219854.

17 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016;15(2):155–63.

18 Streiner DL, Norman GR, Cairney J. Health Measurement Scales: a practical guide to their development and use. 5th ed. Oxford: Oxford University Press; 2015.

19 de Ron J, Fried EI, Epskamp S. Psychological networks in clinical populations: investigating the consequences of Berkson's bias. Psychol Med. 2021 Jan;51(1):168–76.

20 McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30–46.

21 de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59(10):1033–9.

22. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986 Feb 8;327(8476):307–10.

23 Cicchetti DV. Assessing inter-rater reliability for rating scales: resolving some basic issues. Br J Psychiatry. 1976 Nov;129:452–6.

24 Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther. 2005;85(3):257–68.

25 Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull. 1969 Nov;72(5):323–7.

26 Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement (chapter 18). In: Shewart WA, Wilks SS, Fleiss JL, Levin B, editors. Statistical methods for rates and proportions. John Wiley & Sons, Ltd; 2003. p. 598–626.

27 Fayers PM, Machin D. Quality of life: the assessment, analysis and reporting of patient-reported outcomes (Chapter 4). 3rd ed. Chichester, England: Wiley Blackwell; 2016.

28 Altman DG, Royston P. The cost of dichotomising continuous variables. BMJ. 2006 May 6;332(7549):1080.

29 Berk RA. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. Am J Ment Defic. 1979 Mar;83(5):460–72.

30 Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient. Psychometrika. 1979;44(4):443–60.

31 Muthén B. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika. 1984;49(1):115–32.

32 Rabe-Hesketh S, Skrondal A. Generalised linear mixed models (Chapter 4). In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. Longitudinal data analysis. Handbooks of Modern Statistical methods. Boca Raton, FL: Chapman & Hall/CRC; 2008. p. 79–106.

33 Goldstein H, Browne W, Rasbash J. Partitioning variation in multilevel models. Understanding statistics: statistical issues in psychology. Educ Soc Sci. 2002;1(4):223–31.

34 Bates D, Zimlichman E, Bolker B, Walker S. Finding patients before they crash: the next major opportunity to improve patient safety. BMJ Qual Saf. 2015;24(1):1–3.

35 Brage S, Brage N, Franks PW, Ekelund U, Wareham NJ. Reliability and validity of the combined heart rate and movement sensor Actiheart. Eur J Clin Nutr. 2005;59(4):561–70.

36 Aili K, Åström-Paulsson S, Stoetzer U, Svartengren M, Hillert L. Reliability of actigraphy and subjective sleep measurements in adults: the design of sleep assessments. J Clin Sleep Med. 2017 Jan 15;13(1):39–47.

37 de Vet HCW, Mokkink LB, Mosmuller DG, Terwee CB. Spearman–Brown prophecy formula and Cronbach's alpha: different faces of reliability and opportunities for new applications. J Clin Epidemiol. 2017 May;85(85):45–9.

38 Bland JM, Altman DG. Statistics notes: Cronbach's alpha. BMJ. 1997;314(7080):572.

39 Flora DB. Your coefficient alpha is probably wrong, but which coefficient omega is right? A tutorial on using R to obtain better reliability estimates. Adv Meth Pract Psychol Sci. 2020;3(4):484–501.

40 Cheng Y, Yuan K-H, Liu C. Comparison of reliability measures under factor analysis and item response theory. Educ Psychol Meas. 2012;72(1):52–67.

41 Bjorner JB. State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. J Patient Rep Outcomes. 2019 Jul 30;3(1):49.

42 Brennan RL. Generalizability theory. New York, NY: Springer-Verlag; 2001.

43 Pettersen KI, Reikvam A, Rollag A, Stavem K. Reliability and validity of the Kansas City Cardiomyopathy Questionnaire in patients with previous myocardial infarction. Eur J Heart Fail. 2005;7(2):235–42.

44 Joseph SM, Novak E, Arnold SV, Jones PG, Khattak H, Platts AE, et al. Comparable performance of the Kansas city Cardiomyopathy questionnaire in patients with heart failure with preserved and reduced ejection fraction. Circ Heart Fail. 2013;6(6):1139–46.

45 Rhemtulla M, Brosseau-Liard PÉ, Savalei V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychol Methods. 2012 Sep;17(3):354–73.

46 Borg DN, Bach AJE, O'Brien JL, Sainani KL. Calculating sample size for reliability studies. PM R. 2022 Aug;14(8):1018–25.

47 Bonett DG. Sample size requirements for estimating intraclass correlations with desired precision. Stat Med. 2002 May 15;21(9):1331–5.

48 Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med. 2012 July 04; 31(29):3972–81.

49 Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med. 1998;17(1):101–10.

50 Shieh G. Sample size requirements for the design of reliability studies: precision consideration. Behav Res. 2014;46(3):808–22.

51 Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med. 1992;11:1511–9.

52 Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

53 Rotondi MA, Donner A. A confidence interval approach to sample size estimation for interobserver agreement studies with multiple raters and outcomes. J Clin Epidemiol. 2012 Jul;65(7):778–84.

54 Beauchamp MK, Hao Q, Kuspinar A, D'Amore C, Scime G, Ma J, et al. Reliability and minimal detectable change values for performance-based measures of physical functioning in the Canadian longitudinal study on aging. J Gerontol A Biol Sci Med Sci. 2021 Oct 13;76(11):2030–8.

55 Garriga A, Sempere-Rubio N, Molina-Prados MJ, Faubel R. Impact of seasonality on physical activity: a systematic review. Int J Environ Res Public Health. 2021 Dec 21;19(1):2.