This is an add-on resource generated by the Validating Novel Digital Clinical Measures project.

Statistical methodology considerations for analytical validation studies where measures have directly comparable units

When selecting appropriate statistical methodologies for your analytical study, in line with the steps laid out in the Framework for Validating Novel Digital Clinical Measures, it is important to carefully consider your data types and your study objectives.

Below are some suggested methodologies and agreement statistics to use in situations where the data from both the digital clinical measure of interest and the reference have directly comparable units.

Directly comparable units are either identical, or can be translated for the purposes of comparison, such as via calibration.

The suggestions are presented based on whether your digital clinical measure of interest collects categorical or continuous data.

Categorical data (including binary outcomes)

In assessing agreement between your digital clinical measure of interest and your reference measure, consider producing agreement statistics for:

- True Positives/True Negatives/False Positives/False Negatives
- Sensitivity and Specificity
- Accuracy and Misclassification
- Positive and Negative Predictive Value
- Recall
- F_1 score and Micro F_1 score

These values, and other related values, can be reported using a confusion matrix. Using receiver operating characteristic curves, and using a multi-class classification approach such as the one detailed in Yang (2009) when your data has more than two categories, to visualize and analyze your results can also be informative.

If you have ordinal data, then Kendall's τ rank distance can be used to understand and quantify how similarly the digital clinical measure and the reference measure categorize their data. Item-weighting the Kendall's τ rank distance by, for example, adapting the approach in van Doorn et al. (2021) may be even more informative, by penalizing greater levels of misclassification more strongly than lower levels.

Continuous data, or categorical with fine categories

Often continuous data or a continuous score is produced as the algorithm output, or discrete data is produced with many levels or large counts such that you may appropriately treat it as continuous. Data of this type are typically volumes or durations. In the case where data are continuous in nature, classification tables are no longer useful without coarsely discretizing the data, leading to a loss of power and test sensitivity.

Instead, agreement statistics for continuous data can be used. The familiar Bland-Altman plots can be used to assess agreement between your digital clinical measure of interest and the reference measure, however intraclass correlation coefficients (ICCs) can also be considered in this case. ICCs for absolute agreement between two raters can be used to assess the agreement between the digital clinical measure of interest and the reference measure, by treating each measure as one of the two raters. Along similar lines, the Concordance Correlation Coefficient (CCC) can also be employed as an agreement statistic between your digital clinical measure of interest and your reference measure.

Be conscious of the distribution and heterogeneity of your data if using ICCs in this context. Typically, ICCs are used to assess ratings or questionnaire scores, where data is bounded and often normally distributed. These properties constrain the heterogeneity that exists in a scale, in the absence of excessive floor or ceiling effects. Data arising from sensor-based digital health technologies, however, is often skewed in its distribution, and either partially

or fully unbounded (for example, step count is only bounded from below). This unboundness means that heterogeneity in patient ability could lead to an increase in between-subject variance when compared to questionnaire scores, which may artificially inflate the ICC statistic. Using traditional interpretation thresholds, your digital clinical measure of interest may erroneously appear to be in strong agreement with your reference measure in this case. Therefore, considering adjustments in the ICC thresholds used (such as using more conservative thresholds for acceptability) would be encouraged, based on an assessment of your data's distribution and heterogeneity.

Statistical methodology considerations for analytical validation studies where measures do not have directly comparable units

After using the Framework for Validating Novel Digital Clinical Measures to select reference measures, develop novel comparators, or identify anchor measures for your analytical validation study, you may have chosen **measures that do not have directly comparable units** to your digital clinical measure of interest, and have chosen lower-ranked reference measures such as **reported measures, comparators, or anchors**. In such situations, investigators are generally limited to assessing associations and correlations by using metrics such as the Pearson Correlation Coefficient. While this and other established methods remain suitable for such an analytical validation study, we offer additional statistical methodology considerations that may complement these, and give a broader understanding of the agreement between your digital clinical measure of interest and your reference measures.

Construct validity

Investigators may find that employing ideas and methods from the field of construct validity, and in particular convergent validity, are useful in this scenario. Evidence can be derived from demonstrating theoretically expected outcomes of your digital clinical measure of interest, and relationships with your chosen reference measures.

There are several tests of construct validity, but all rely on a measurement assumption called the latent trait. A latent trait is the underlying level of severity or ability on a given construct. For example, a latent trait for physical

activity would represent the underlying level of physical activity ability. Crucially, although the latent trait is unseen and immeasurable, it influences an individual's behavior. This means that an individual's level of latent trait can be estimated through assessing "indicators" of that trait. In traditional psychometric research, the indicators are typically questionnaire items related to the topic under investigation, however, in the case of digital health technology measurement, the indicator of the underlying construct would be the output of your algorithm.

This means that we may be able to make testable hypotheses, either between measures that are assumed to target the same latent trait, or based on groups of people who are assumed to have a greater or lesser value of the latent trait.

A statistical technique highly suited to this approach is confirmatory factor analysis.

Confirmatory Factor Analysis (CFA)

CFA can be employed to assess how well the observed data from the digital clinical measure of interest and the reference measures fit a hypothesized latent trait theoretical model. A two-factor correlated factors model can be employed, where one factor concerns the digital clinical measure of interest, and one factor concerns a reported reference measure, comparator, or anchor.

The digital clinical measure factor in the CFA model is loaded with each day of subject data as a separate variable, with data summarized from epoch level as necessary; the reference measure factor is loaded with the individual items from the reported reference measure, comparator or anchor. Once model fit is verified, the correlation relationship between the digital clinical measure factor and the reference measure factor can be used to assess the strength of the relationship between a given indicator and the underlying latent trait.

To understand more about how this CFA model can be implemented in an analytical validation study, please refer to this manuscript, and to the Simulation Toolkit for Validating Novel Digital Clinical Measures on GitHub. Using the materials contained in the Toolkit, we established that CFA factor correlation is less biased, albeit less precise, than the Pearson Correlation Coefficient (PCC), when analyzing simulated longitudinal step count data where the true relationship between the measures is strong.

If you intend to use CFA in your analytical validation study, then there are some caveats to note. Firstly, CFA is known to require a larger sample size in order to produce stable estimates. While we cannot advise a uniformly applicable minimum sample size, the consensus is that a sample of participants in at least the hundreds is desirable. While this sample is not common in analytical validation studies of this type conducted so far, with the improving feasibility of conducting observational research in the out-of-lab environment, larger sample sizes are increasingly accessible.

CFA requires more than one variable loaded onto a factor in order for the model to be identified. Studies using this CFA approach must collect longitudinal data and repeated measures from their digital clinical measure. Any reference measure used must contain more than one item. In line with the established consensus, a minimum of three repeated measures or items is strongly recommended. Scaling the digital clinical measure variable to match scales with the scale used for the items of the reported reference measure, comparator, or anchor may also be required to achieve model convergence.

Known-groups validity

Another pertinent subtype of construct validity for an analytical validation study of this type is known-groups validity.

Under the same latent trait concept described above, it follows that data collected from a context where higher levels of the latent trait are hypothesized, should lead to data arising from the algorithm output in line with this hypothesis. In known-groups validity using questionnaire data, the different measurement contexts are typically different groups of individuals, some with a greater propensity for the underlying latent trait and some with less propensity. With a digital clinical measure this understanding of measurement contexts can be expanded.

Traditional known-groups analysis selects or defines groups of individuals who are known to vary on the level of the latent trait under assessment. For example, this could mean comparing a group of individuals from the general population assumed to be unimpaired in their physical activity, with a population who would be expected to display a lower level of physical activity, such as a population diagnosed with rheumatoid arthritis. From a latent trait perspective, these two groups are expected to have a distribution of physical activity ability that differs from one another, but exists on the same spectrum.

In this case, the hypothesis is that there is a numerical, interpretable and statistically significant difference in the mean level of the digital clinical measure recorded for the two groups, in line with expectations about their level of functioning on the latent trait.

An analytical validation study looking to leverage known-groups validity techniques could enroll groups of participants from different known groups in the population, each displaying a different expected level of the latent trait under examination. Alternatively, a single group of participants could be enrolled, who are known to cover a range of abilities on the latent trait, and are categorized using an external measure such as a Patient Global Impression of Severity. After this step, analysis could be conducted to compare the output between the groups.

For example, one could derive the mean output for each group and the associated effect sizes of the difference between them, as a test of whether the groups display the expected differences. Mean scores, standard deviations and confidence intervals would allow interpretation of the magnitude of any differences, and these could be supported by a calculation of the effect sizes of the differences, using the formula

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

where s is the pooled standard deviation.

A note of caution: effect size is typically interpreted in line with guidance by Cohen, suggesting that effect sizes of 0.2 are small, 0.5 are moderate and 0.8 and above are large. One issue in the digital health field is that the effect size formula considers the difference standardized through using the pooled standard deviation as a denominator. When the variance is expected to be large in both groups (as discussed above in the section on ICCs) this can lead to low magnitude of effect size, even when the mean differences are stark. It may be appropriate to accept lower effect size values as evidence of analytical validation in novel digital clinical measure scenarios, as the between-subject variance is likely to be larger than the contexts for which the thresholds were initially derived.

Other analysis techniques, such as logistic regression assessing the change in the outcome variable between groups, may lead to complementary outcomes while also considering covarying factors. An advantage of such methodology is its ability to exhibit the power of the outcome variable to show differences, even when controlling for covariates which may not be equal between groups, but are known to have a potential impact on the output.

Outside of the field of construct validity, linear regression can also be employed in an analytical validation study of this type.

Linear Regression models

Simple linear regression models can be built with a reference measure as a predictor and the mean values of your digital clinical measure as the outcome, using R^2 as the agreement statistic.

If multiple reference measures, comparators, or anchors are chosen for your study, then multiple linear regression models can also be built by including each reference measure as a predictor for your digital clinical measure outcome, using adjusted R^2 as the agreement statistic.

When introducing additional predictors for a multiple linear regression model, a trade-off must be considered: the adjusted R^2 may increase at the cost of the model precision. This trade-off is particularly important to consider when using a reference measure, comparator, or anchor that collects data on a daily basis. This trade-off can be observed in more detail by using the Simulation Toolkit for Validating Novel Digital Clinical Measures, to analyze simulated longitudinal step count data where the true relationship between this measure and each of several reported reference measures is strong.

When using regression models, extra care should be taken to minimize data missingness, particularly to minimize situations in which participants complete some but not all of the reference measure assessments. Data missingness particularly affects regression models, where incomplete cases will lead to an entire participants' data being excluded, thus reducing the sample size.

In addition to these simple and multiple linear regression models, Deming regression models can also be considered. As errors-in-variables models, they will account for errors in observations in both your digital clinical measure ***and*** the reference measures. However, care must be taken to accurately estimate

the ratio of variances between the measures when using this method by, for example, using the ratio of the sample variances of the data from your two measures.

General study design considerations

There are three key concepts that you should consider when deciding on your statistical strategy for your analytical validation study of a novel digital clinical measure: **temporal coherence**, **construct coherence**, and **data completeness**.

Temporal coherence describes the similarity between the time periods of data collection for two measures. Poor temporal coherence between measures may decrease the values estimated with agreement statistics, because a participant's meaningful aspect of health assessed by the measures may have changed or fluctuated over time.

When using a single-time-point reported reference measure in your study, such as a PRO with a two-week recall period that is administered to each participant only once, aligning the recall period with the duration of the digital clinical measure data collection is recommended, and is expected to increase temporal coherence. We further recommended that the reference measure should be assessed at the conclusion of the digital clinical measure data collection period, in order to increase temporal coherence.

In addition to improving temporal coherence, capturing repeated measures of the digital clinical measure during the recall period of a single-time-point reported reference measure allows for analysis of the digital clinical measure's mean values against the total score from the reference measure. If any reported reference measures collect daily data (such as a daily Patient Global Impression of Severity), then consider if analysing mean values or being more granular is the best approach. For example, in situations where events such as breakthrough pain are a factor, it may be advantageous to be more granular and consider individual digital clinical measure days correlated against your daily reference measure, especially if such events are assumed to strongly affect a participant's recall for a reported reference measure.

If your digital clinical measure captures data at the epoch level, consider whether passing to a summary level, such as a total count of events per day, is most appropriate for analysis of your measure. This may be particularly appropriate if using a reported reference, comparator, or anchor with a daily recall period, as the temporal coherence between the measures would be expected to increase.

Construct coherence is the level to which your digital clinical measure and your reference measures assess the same underlying concept or latent construct. Poor construct coherence is likely to lead to weak or non-meaningful relationships between measures, no matter the statistical methods employed.

Data completeness is the extent to which both your digital clinical measure and your reference measure(s) avoid data missingness. Patterns of data missingness in the digital clinical measure and the reference measure may distort the results of your statistical analyses, and so steps should be taken to identify:

- Likely causes of data missingness across all of your measures.
- The expected patterns of data missingness, and their likely effect on the statistics or estimates obtained from your methods.
- A strategy to maximize data completeness in all of your measures. This may include technical considerations, or social considerations such as a patient engagement strategy where an investigator reminds participants the day before the beginning of a device wear period via phone call.

The table below presents a summary of further recommendations to aid in your design of a rigorous analytical validation study for a novel digital clinical measure.

Category	Sub-category	Considerations
Digital measure data collection	Number of days of data captured by the measures	Longitudinal collection on consecutive days allows for the use of CFA methods, as long as at least three days are collected. Have an enactable participant engagement strategy to minimize data missingness.
Study Design	Rigor and quality of reference measures	The quality of a reference measure affects the claims that can be made about the performance of your digital clinical measure. High-quality and high-rigor reference measures enable the possibility for the strongest claims to be made about your digital clinical measure. Use the reference measure hierarchy, as seen in the Interactive Guide for Validating Novel Digital Clinical Measures, to guide you in selecting appropriate reference measures of the highest possible rigor for your analytical validation study.
	Objectivity of reference measures	Objective data capture in a reference measure improves accuracy by reducing the possibility of measurement error. Objective data processing, and standardized and trained data interpretation, reduces ambiguity and avoids issues with inter-rater variability.
	Construct coherence of measures	Good construct coherence between measures may strengthen the values estimated from agreement statistics. Poor construct coherence may cause issues, even if the methods are well suited to assessing agreement. Consider the effect of construct coherence at the item and instrument

		level if using a reported reference, comparator, or anchor.
	Temporal coherence of measures	Good temporal coherence aligns data capture, meaning the measures assess a subject over the same time period. Poor temporal coherence may decrease the values estimated with agreement statistics, because the measures assess the construct at different times and the level of the construct is subject to change. If using a reported reference, comparator, or anchor, then consider the benefit of using a daily recall period and assessing on the same days as your digital clinical measure, if for example, the digital measure collects daily summary count data. For a reported reference, comparator or anchor with a multi-day recall period, applying the reference measure at the end of the period of digital measure data collection, and collecting digital measure data on each day of the recall period, is expected to increase temporal coherence.
	Miscellaneous	Review the assumptions and requirements of the statistical methods used, and consider any likely violations of assumptions that your digital clinical measure data and your reference measure data may incur (such as the distribution of your data). Consider how any likely assumption violations may distort the statistics or estimates obtained from your chosen methods, and plan to minimize violations of assumptions, or avoid them entirely where possible.
		Identify factors that may increase missingness and measurement error in data capture. Such factors may include issues such as smartphone compatibility for a digital measure with smartphone integration, or usability issues such as discomfort in wearing a device for the required

		study duration. Seek to minimize the identified factors where possible.
Statistical methods for assessing agreement when the measures do not have directly comparable units	CFA	CFA can account for measurement error and variance at the item level when working with reported references, comparators, or anchors. This is because it can assess the latent correlation between measures, and correlation between latent variables is not attenuated by measurement error.
	Pearson Correlation Coefficient (PCC)	The PCC is stable, easier to compute, and relatively robust with respect to violations of parametric assumptions - see Havlicek and Peterson (1976) for details. However, the PCC is known to underestimate the true correlation between measures, because of attenuation by measurement error.
	Linear regression models	If multiple reference measures, comparators, or anchors are being used in your study, then multiple linear regression models may provide a stronger assessment of agreement between measures than individual simple linear regression models. This may be a particularly useful approach if your study uses a reported reference, comparator, or anchor with a daily recall period.
Sample size	-	The statistical methods employed in an analytical validation study affect the appropriate minimum sample size for analysis. Methods such as CFA often require a large sample numbering in at least the hundreds, however, this could be fulfilled by including repeated assessment periods from each participant.