# 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 all information gathered during Section 1 of the Framework, including your data types, 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. The suggestions are presented based on whether the data your digital clinical measure of interest collects are categorical or continuous.

### 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 by using a confusion matrix. Using receiver operating characteristic (ROC) curves with a multi-class classification approach (such as a "one vs rest" approach) to visualize and analyze your results can also be informative.

If you have ordinal data, then Kendall's tau rank distance can be used to understand and quantify how similarly the target measure categorizes the data to the reference measure. Item-weighting the Kendall distance can allow for greater levels of misclassification having a larger penalty imposed upon them.

**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 we 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.

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 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.

Along similar lines, the Concordance Correlation Coefficient can also be employed as an agreement statistic between your digital clinical measure of interest and your reference measure.

## 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, or develop novel comparators and anchors, 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, novel reported comparators, and reported 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 the digital clinical measure of interest and relationships with the 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 the 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 single reported reference measure, comparator, or anchor.

The digital 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 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 the other features of the Simulation Toolkit for Validating Novel Digital Clinical Measures. Using these features, it can be seen that CFA factor correlation is less biased, albeit less precise, than Pearson correlation 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 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 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 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.

To explore how CFA may be useful in your analytical validation study, download the other resources in the Simulation Toolkit for Validating Novel resources, and simulate an analytical validation study that is tailored to your scenario.

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

**Known-groups validity**

Under the same latent trait framework 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 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 diagnosed with rheumatoid arthritis, who are expected to display a lower level of physical activity. 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 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 categorized using an external measure such as a Patient Global Impression of Severity.

After collecting data on individuals who represent two or more groups with different expected outcomes, or after dividing a sample into groups of individuals expected to have different outcomes based on some external measure, analysis can 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. 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 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.

## Regression models

Linear regression models can also be employed in your analytical validation study in this scenario. Simple linear regression models can be built with a reference measure as a predictor and the mean values of the digital 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 the digital 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 can be observed in more detail by using the other features of the Simulation Toolkit for Validating Novel Digital Clinical Measures, to analyze simulated longitudinal step count data where the true relationship between the measures is strong. This trade-off is particularly important to consider when using a reference measure, comparator, or anchor that collects data on a daily basis.

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 entire participants' data being excluded, thus reducing the sample size.

In addition to these methods, Deming regression models can also be considered. As errors-in-variables models, they will account for errors in observations in both the digital measure ***and*** the reference measure. 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 the two measures.

## General study design considerations

When using a single-time-point 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 measure data collection is recommended. We further recommended that the reference measure should be assessed at the conclusion of the digital measure data collection period.

Capturing repeated measures of the digital measure during the recall period of a single-time-point reference measure allows for analysis of the digital measure's mean values against the total score reference measure. If any 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 important, it may be advantageous to be more granular and consider individual digital 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 the digital 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.