Journal of Electrical Bioimpedance
www.bioimpedance.net

J Electr Bioimp, vol. 5, pp. 14–27, 2014
Received: 5 Mar 2014, published: 16 Apr 2014
doi:10.5617/jeb.830

TUTORIAL ARTICLE

Statistical methods for bioimpedance analysis

Christian Tronstad[1,3] and Are H. Pripp[2]

1. *Department of Clinical and Biomedical Engineering, Oslo University Hospital, Oslo, Norway*
2. *Department of Biostatistics, Epidemiology and Health Economics, Oslo University Hospital, Oslo, Norway*
3. *E-mail any correspondence to: christian.tronstad@gmail.com*

Abstract
This paper gives a basic overview of relevant statistical methods for the analysis of bioimpedance measurements, with an aim to answer questions such as: How do I begin with planning an experiment? How many measurements do I need to take? How do I deal with large amounts of frequency sweep data? Which statistical test should I use, and how do I validate my results? Beginning with the hypothesis and the research design, the methodological framework for making inferences based on measurements and statistical analysis is explained. This is followed by a brief discussion on correlated measurements and data reduction before an overview is given of statistical methods for comparison of groups, factor analysis, association, regression and prediction, explained in the context of bioimpedance research. The last chapter is dedicated to the validation of a new method by different measures of performance. A flowchart is presented for selection of statistical method, and a table is given for an overview of the most important terms of performance when evaluating new measurement technology.

Keywords: Statistics, research methodology, bioimpedance, data analysis, data reduction, parameterization, prediction, classification, performance, validation

1. Introduction

When doing measurements, statistics are needed if we want to describe the data (descriptive statistics) or if we want to draw conclusions based on the data (inferential statistics). There is a vast amount of statistical methods in the literature, and the choice of method depends on what we want to know and what type of data we have. In this paper we give an overview of the most basic and the most relevant methods for bioimpedance analysis along with examples within the bioimpedance field. Because bioimpedance measurements often are done as frequency sweeps, producing large amounts of correlated and possibly redundant data, the implications for inferential statistics are discussed together with data reduction solutions. A goal of bioimpedance research is often to develop methods for prediction of a biological variable or state, and an overview is given for the most relevant methods for development and testing of a prediction model. At last, the validation of a new measurement technology employs distinct statistical methods, and an overview is given on the concepts, terms and methods for evaluating performance.

2. Hypothesis and research design

Instead of beginning with the type of measurement as a basis for selecting the statistical method, we expand the perspective by beginning with the hypothesis and research design that should come before the measurements are acquired. The reason is that we generally have an idea about what we want to investigate with our bioimpedance measurement, and we do not perform measurements completely randomly. In order to do our investigation properly, it begins with a research hypothesis where we formulate what we want to investigate in a testable way. For instance, if we want to find out whether gel electrodes provide lower bioimpedance measurement than textile electrodes, our hypothesis can be formulated as: "Bioimpedance is lower when using gel electrodes compared to using textile electrodes". We now have a testable hypothesis, and our hypothesis can be either be accepted or rejected by experiments. It is much easier to dismiss a hypothesis than to prove a hypothesis, because it takes only one piece of solid evidence to reject it, but an endless amount to prove it correct. That is why the statistical methods are based on rejecting an opposite hypothesis, called a *null-hypothesis*, instead of attempting to prove the hypothesis. In our example, we test whether "Bioimpedance is not lower when using gel electrodes compared to using textile electrodes", which is our null-hypothesis. We reject the null-hypothesis and thereby accept our original hypothesis if the statistical analysis of our measurements find that the null hypothesis is improbable. The statistical analysis provides a *p-value*, which is the probability of our measurement result or larger deviations from the null hypothesis, assuming that the null hypothesis is actually true. Whether or not to reject the null hypothesis is based on whether the p-value is lower than a predetermined threshold, the *alpha* (α) (i.e. the significance level). α is conventionally set to 0.05 in medicine and biology, implying that we reject the null hypothesis if our measurement result is less than 5% probable with the assumption that the null hypothesis is true.

The hypothesis example above is very general and the testability could be improved by making it more specific, i.e. "Trans-thoracic bioimpedance is lower when measured by gel electrodes than measured by textile electrodes using a two-electrode setup" if this is the relevant setup we want to test. It is easier to test this hypothesis because it implies only one certain type of measurement, and reduces the chance of an inconclusive result. A hypothesis should be

simple, specific, and stated in advance [1]. It is the hypothesis that determines the research design, the type of experiments we need to conduct and the type of measurements we need to take. It is also the hypothesis which mainly determines what type of statistical test is appropriate. As an example, if we want to investigate whether the bioimpedance of two types of tissue samples are significantly different, this means that we have to assess the difference between two groups of bioimpedance measurements from a number of tissue samples of each type, and that the appropriate statistical test will be a test for comparison of two means in the groups such as the Student's t-test. We can also do a power analysis in order to estimate how many tissue samples we need in order to have a good chance of finding a difference if there actually is one. Hence, the planning of a study should begin with a clear hypothesis. The whole process from the planning of the study to the statistical analysis of the measurements is illustrated in figure 1.

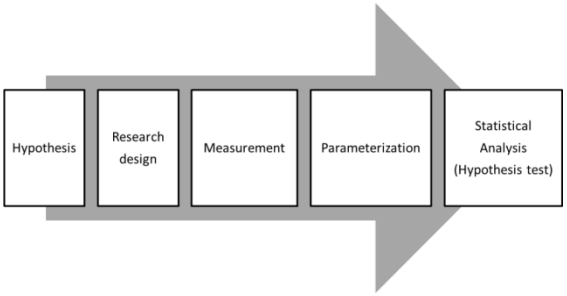

Fig.1: The steps from the formulation of a hypothesis to the testing of the hypothesis by statistical methods.

1.2. How many subjects are needed?

It is a good idea to know how many units (i.e. items or subjects) are needed in order to test our hypothesis. Unless we test all the units in a population, we are only testing a sample of the whole population. In order to make a general conclusion about the population, we need to show that the effect that we observed was not likely due to chance from random variation in our sample. If we choose too few units, we may end up with an inconclusive result and a worthless study, and if we choose too many, we are wasting resources (e.g. sacrificing more animals than needed). Hence, sample size consideration is of ethical relevance [2].

In hypothesis testing, we want to reduce the chances of two types of errors: incorrectly rejecting a true null hypothesis (Type I error), and the failure to reject a false null hypothesis (Type II error). The Type I error probability is determined by the α (i.e. 5% usually). With a given α, we can also calculate the *beta* (β), which is the probability of a type II error. As an example, let us say you want to replicate a pilot study you did on the bioimpedance of two materials, material A and B. You want to test whether they are different based on a t-test, but you are not sure what sample size (N) you need in order to test the hypothesis. Based on the pilot data, you can estimate how large difference between the materials you expect, and also how much variation there is within each material. With an expected difference in means of 10 Ohms and a standard deviation also of 10 Ohms as an example, N=5 (for each material) gives a beta of 71%, which is a too large chance of a Type II error. Increasing N to 10, we obtain a beta of 44%, and with N=20 the β is down to 13%. The statistical power of the test is 1-beta, and can be viewed as the ability to correctly reject the null hypothesis when it is false. The power requirement in a study depends on the type of investigation, but a power >0.8 is often considered acceptable.

The *effect size* is the relative magnitude of the effect we are investigating. When comparing groups, the effect size could for example be the difference in |Z| between the groups divided by the pooled standard deviation, or for testing associations the effect size could for instance be the coefficient of determination R^2. It follows that smaller effect sizes require larger samples in order to be detectable and avoid a type II error. In the example above, the effect size was 1 (10 Ohms / 10 Ohms). If the effect size was 2, only N=6 would be required to obtain the same power as with N=20 for an effect size of 1. The minimum amount of information needed to do a sample size estimation for a given statistical test is:

- Desired α (probability of incorrectly rejecting a true null hypothesis)
- Desired β (probability of failing to reject a false null hypothesis)
- Expected sample distribution (type of distribution and variance)
- Expected magnitude of difference or association.

Estimation of sample sizes is not an exact science, and often these inputs will be "a qualified guess", even so it is still important to assess whether we need something like 10 or 100 subjects/items. We already have decided our type of test based on the hypothesis, and our α will conventionally be set to 0.05 with a beta somewhere above 0.8. What is left for us to provide is the effect size. If this is not known, the first place to look is in similar studies. Perhaps other investigators have published data with similar measurements on a similar sample. If no previous data are available, conducting a pilot study can give a good indication of these values. Perhaps we gather information which suggests that the sample variance may be somewhere between 100 to 500 Ohms, and that the difference between means is between 1k to 2k Ohms. In such cases it is best to account for the worst case (variance = 500 Ohms and difference between means = 1k Ohm) in the sample size determination.

In practice, the sample size calculation is not done by hand, but by computer programs (such as the free G*Power ©) which lets you choose a statistical test, asks for the necessary inputs (i.e. α, β, variance and effect size), and gives you the minimum required sample size. They can also be used to determine the power of your test given the sample size, α, β and effect size. Because of all these

unknowns, it is a good idea to consult a biostatistician on these matters if possible.

3. Multiple variables and data reduction

Often in bioimpedance measurements, we want to examine more than one bioimpedance variable per sample or measurement. Sometimes, we have limited knowledge beforehand on effects in bioimpedance in our study. In order to maximize the changes of a finding, we may acquire several bioimpedance parameters (i.e. |Z|, G, theta) at multiple frequencies from one measurement. Say this gives us a set of 100 variables for comparing two different tissue types, it is very probable that we will find a significant difference in at least one of the variables purely due to chance. It is possible to do adjustments for such multiple comparison tests by e.g. the Bonferroni correction method [3,4], which adjusts the significance level threshold by dividing the single-comparison level by the number of multiple comparisons included in the analysis. For bioimpedance analysis, this approach is often insufficient due to the large number of comparisons. Unless the numbers of comparisons are few, a better approach is to reduce the number of variables by data reduction or model based approaches. Quite often, and especially for bioimpedance frequency sweeps, the data will be highly correlated and can be reduced into a small set of variables which account for most of the information in the measurements. A common method in the bioimpedance field is to assume that the electrical properties of the sample can be described by an electrical equivalent model (see chapter 8 in [5]) such as the Cole model, and to estimate the component values by fitting the measurement to the mathematical expression of the model. With a good agreement between the model and the measurement, this approach reduces the measurement into a few uncorrelated parameters which are easier to handle statistically. Data reduction can also be done without equivalent model assumptions. One such way to reduce the data is to computationally transform the data into a set of uncorrelated components using principal component analysis (PCA). The transformation works in the way that linear combinations of the data are used to construct components which explain as much as possible of the variance in the data, with the constraint that all components must be uncorrelated. The PCA may provide a data subset by which almost all the information (i.e. 99%) is accounted for by just a few components. The disadvantage of PCA compared to the model-based approach is that the transformation is a "black-box" and the principal components are meaningless with respect to what we are measuring.

4. Choice of statistical method

After the data has been reduced to a practical set of parameters (if necessary), the next step is to perform a statistical analysis in order to test whether our null-hypothesis can be rejected or not. The choice of statistical method is mainly determined by the hypothesis, but the measurements may also influence the selection of the most appropriate method. Figure 2 provides a flowchart for selection of statistical method based on the type of study.

4.1. Comparing two groups

Let us go back to the example of the alpha parameter of two tissue types, with the hypothesis that the alpha is different between the two tissue types. Our natural choice of test is a two-sample Student's t-test, which is designed to test whether the means of two sets of data are different. If however, our hypothesis is also on the *direction* of the difference between the tissue types, such as "the alpha parameter of tissue A is *larger* than for tissue B", the statistical testing must also include this direction. Imagine if we throw five coins, the probability of getting all heads or all tails is 0.03 (0.55), but twice (0.06) for getting either all heads or all tails. In the first hypothesis (one-sided), getting all heads would be statistically significant by $p<0.05$, but not for the second hypothesis (two-sided). When comparing two groups for one direction of the difference, the one-sided (also called one-tailed) t-test takes this into account. In medicine, the two-sided hypothesis and tests should be used unless there is a very good reason for doing otherwise, and if one-sided tests are used, the direction of the test must be specified in advance [6].

The t-test belongs to the family of *parametric* tests, which assume that our data follows a mathematical probability distribution, in this case the normal distribution, which is something we usually do not know before all the measurements are done. Our distribution of alpha values may be asymmetrical with an overweight of alphas close to one and fewer and fewer alphas towards zero. We then have two options, either to mathematically transform our data into a normal distribution, or use a type of statistical test which does not require such a distribution. The alternative type of test for unpaired data which do not satisfy an assumption about normal distribution is the Wilcoxon ranksum test (also called the Mann-Whitney U test), which is based on comparing the ranks of the values within the groups. This type of test does not rely on any parameter for describing the distribution of data (such as the standard deviation), and belongs to the non-parametric family of statistical tests, which handle different types of hypothesis testing, typically based on data ranking.

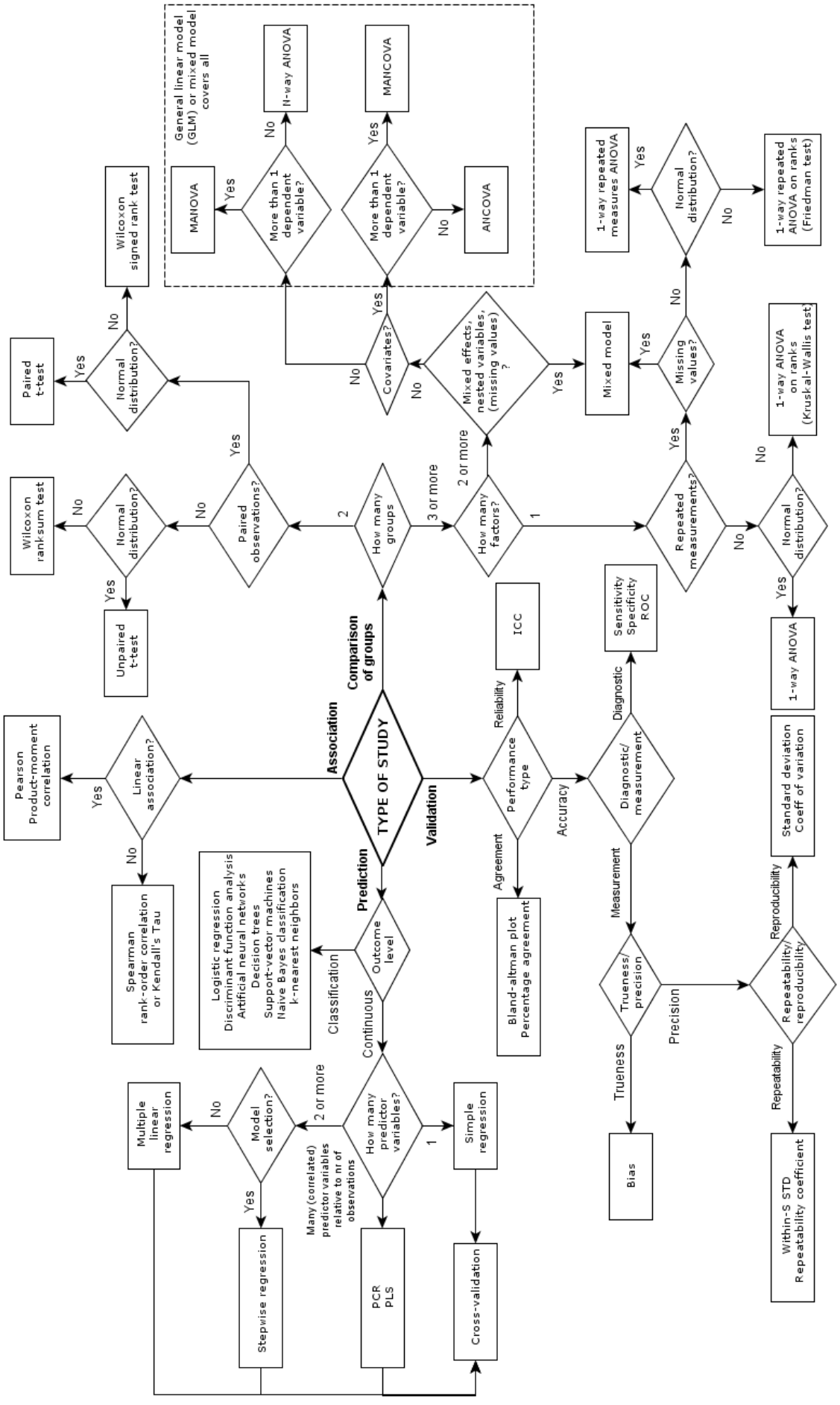

Fig.2: Flowchart for selection of statistical method based on the type of study.

In general, the parametric tests are a better choice if possible because of a higher statistical power. In bioimpedance analysis, we often do mathematical transformations of our measurements in order to interpret or graph them differently. When doing statistical analysis, we need to keep in mind that the transformations may also change the distribution of the data. For instance, when transforming a normally distributed set of |Z| measurements to |Y|, the distribution is likely to change into a non-normal one.

In the previous example, the measurements came from independent samples. If we have pairs of tissue types with each pair coming from the same animal, we cannot consider the samples from the two tissue types independently, and we have to use statistical tests which account for the correlations within each pair, such as the paired t-test or the non-parametric Wilcoxon signed rank test. A typical situation where these tests are recommended is for testing the change in bioimpedance before versus after a treatment.

4.2. Comparing more than two groups

If we want to statistically compare more than two groups of measurements, another type of test is better suited - the analysis of variance (ANOVA). This test compares the variance within each group to the variance between the groups, and also overcomes the problem of multiple pairwise comparisons (as described in chapter 3). The ANOVA has the following assumptions: independence between groups, normal distribution and equal variances within the groups. For more detail on theory, testing and violations of these assumptions, see e.g. [7]. For non-normal data, most statistical packages offer rank-based ANOVAs, and the ordinary ANOVA is also regarded as robust against violations of the normality assumption [8]. The one-way ANOVA, which is used for comparing more than two independent groups, first calculates an F-statistic (based on the ratio between between-group variability and within-group variability) which together with the degrees of freedom determines a p value for the null hypothesis that the data from all groups are drawn from populations with the same mean. Further on, the difference between each pairwise combination of the groups can be tested similar to the t-test but with correction for the multiple testing.

4.3. Factor analysis

The one-way ANOVA is useful when we study only one factor which groups the measurements (e.g. tissue type). If we for example want to study how electrode configuration in addition to tissue type affects the bioimpedance, we have a *factorial design* with two factors and may use the two-way ANOVA. The output of this test gives us the statistics (F-statistic, p-value) which tell us whether each of the two factors have a significant effect on the bioimpedance. In addition, the two-way ANOVA can test whether there is a significant interaction between the two factors; the difference in bioimpedance among tissue types may depend on the electrode system. As a procedure for two or more factors, it is advised to first test for all possible interaction terms, and to continue with an ANOVA without these terms if none are found significant. If there are significant interaction terms, the main effects (e.g. the influence of tissue type and electrode system on bioimpedance) may essentially be rendered meaningless, since both effects will have to be qualified in reference to another factor. The most logical approach in this case is to do one-way ANOVAs for evaluating all levels of one factor across only one level of the other factor (also called *simple effects*). If one or more simple effects are significant, additional comparisons between specific groups within given factor levels can be conducted (for instance tissue type A vs tissue type B when using the four-electrode system). The advantages of the factorial design and the factorial ANOVA are that it allows the same set of hypotheses to be evaluated (at a comparable level of power) by using only a fraction of the subjects which would be required if separate experiments were conducted, and also the possibility to evaluate the interaction between the experimental conditions [9].

In the same way as the unpaired or paired t-tests are suited for comparing independent and dependent groups respectively, there are also ANOVA methods which are suitable for dependent groups, the *repeated measures ANOVAs*. Consider the example of comparing the bioimpedance measurement from three different electrode positions. If all measurements are on different subjects, the one-way ANOVA is the appropriate test, but if you measure several times on each subject (with the different electrode positions), the one-way repeated measures ANOVA is the appropriate test. The repeated measures tests have a higher statistical power, and fewer subjects are needed with the repeated experimental design. Repeated measures ANOVA can also be done for factorial designs (e.g. two-way repeated measures ANOVA), and for non-parametric data (e.g. repeated measures ANOVA on ranks, Kruskal-Wallis test). However, some care should be taken when performing factorial ANOVA on ranks as the rank transform procedure may be erratic for certain designs [10].

In factorial repeated measures design, the effect of time (or the repeated experimental condition) can be investigated by including it as a factor in the two-way repeated measures ANOVA. It is important to know that the ANOVA does not consider the order of the time-points, only the difference between them, and if we want to evaluate a trend or relationship, it is better to use a regression approach.

In experiments, there may be other observable variables than the experimental factors, which have an influence on the dependent variable. This variable may be continuous, and therefore problematic to add as a factor in the design. In this case, the variable can be added as a covariate in the design. Let us use the example of measuring impedance in solutions during different chemical reactions. The temperature changes may be unknown and uncontrollable, but possibly influence the impedance. The temperature can

then not be added as a factor in the analysis, but as a covariate. The appropriate statistical method is the ANCOVA (analysis of covariance), which is a combination of ANOVA and regression. With this method, we will find out whether there is a significant difference between the impedance of the different chemical reactions when also controlling for the temperature effect. In some cases, we want to examine more than one dependent variable. If they are related, such as the |Z| and phase of the same measurement, both the dependent variables can be studied in the same test while controlling for the correlation between them by the MANOVA (multivariate analysis of variance). If the dependent variables are not correlated, separate ANOVAs are appropriate. The MANOVA assesses the effect of each factor on each of the dependent variables (with p-values for each case), and also the interactions both among the independent variables and among the dependent variables. The advantages of this method is that several dependent variables can be studied in one test which avoids the increased Type I error rate from multiple comparisons, the correlations between the dependent variables will be incorporated in the analysis, and the test may even find a significant result for the combined effect of all dependent variables when the effect on each of them are not strong enough. For adding covariates to the MANOVA, the appropriate test is the MANCOVA (multivariate analysis of covariance), which is the same analysis as the MANOVA, but adds control of one or more covariates that may influence the dependent variables.

One type of test that incorporates all of the above (t-test, ANOVA, ANCOVA, MANOVA and MANCOVA and also ordinary linear regression) is the *general linear model* (GLM). This method is included in several statistical software packages and is a convenient tool for analyzing many different types of data.

4.4 Mixed models

Until now, we have discussed group comparison or factor analysis with *fixed effects*. Fixed effects means that the levels of our independent variables will be the same (fixed) in any attempted replication of our experiment. We have chosen a certain selection of levels which are of interest, and do not attempt to generalize beyond these levels. As an example, let us consider a comparison of the impedance of electrode types. We could do a study where the aim is to compare a certain selection of electrode types with different characteristics, and electrode type would then be a fixed factor. We could also do a study where the aim is to assess whether the electrode type has an effect on the measured impedance in general. The electrode type would then be a random factor, and our sample of electrode types (levels) would be treated as a random selection of the overall population of possible electrode types/manufacturers. In the first case, our test result will be the explicit differences between the impedance of the selected electrode types, but for the second case the test result will be the general effect of electrode type on the impedance.

In some cases, our experiment may include both fixed and random factors, and the analysis model is then called a mixed-effects model. Tremendous advances have been made over the last years in the methods for mixed model analysis, and the current tools available offer a lot of different features and advantages over other “traditional” methods. For instance, a mixed model analysis is not weakened by missing values in the same way as repeated ANOVA. The mixed model can also deal with hierarchies in our data. For instance, we may study samples of different electrode types from different producers, and have different types of electrodes from each of the different producers. In the statistical terms, we have two factors (electrode type and producer) where different levels of one factor do not occur at all levels of the other factor, which is called a *nested* (or *hierarchical*) design. A third advantage of the mixed model method is that our measurements do not need to be taken at the same time points. For instance if we are following the impedance of different materials during a time-dependent process, but are unable to measure at all materials simultaneously, we can add all measurements with their individual timestamp to the mixed model and examine the effects of both material type and time even though we only have one measurement per time point. To sum it up, mixed model analysis is recommended for repeated (with missing values) and/or nested data. The analysis can be performed in statistical packages such as R, SPSS, STATA and SAS. The procedure is more advanced compared to methods such as the ANOVA due to the number of choices and settings. Before embarking on the mixed model, it is advised to read up on the subject or consult a biostatistician.

4.5 Association analysis

Until now, we have been dealing with methods for investigating differences between groups and the effect that different factors have on these differences. Now we move over to the methods that assess associations between variables. The most basic case is testing for a linear relationship between two variables (also called bivariate association) by the Pearson Product-Moment correlation coefficient. The output of this test is the r statistic, which indicates the strength (0-1 in absolute value) and direction (positive or negative depending on the sign of r) of the relationship between the two variables. The r does not say anything about the causality or dependency of the relationship, r will be the same whether X was dependent on Y or Y was dependent on X, or if they were independent of each other but dependent on another factor. The r also does not say anything about the agreement between X and Y. You can get an r=1 (perfect correlation) with paired observations on completely different scales. This makes the r insufficient for testing agreement between two methods, but is useful for exploring associations between two variables under a linear assumption. Statistical inference (p-

value and hypothesis test) on the correlation can be conducted (see e.g. eq 28.3 in Sheskin 2011). The squared r is also frequently used and is called the *coefficient of determination*. R^2 is more convenient to interpret, especially for larger r, because it expresses the proportion of variance of one variable which can be accounted for by the variance of the other variable. For instance if we measure BMI and total body |Z| and get an r=0.5, then we can say that 25% of the variance in |Z| could be explained by the BMI.

The Pearson Product-Moment correlation coefficient (and also the coefficient of determination) is based on an assumption of equal variance (in statistical terms called *homoscedasticity*). If we graph two variables against each other, and the scatter is roughly of the same size across the whole range, then the data is homoscedastic. In violation of this (called *heteroscedasticity*), the correlation will not be consistent across the full range of both variables. Different tests for homoscedasticity are available in most statistical software. Another pitfall of this method is that it is sensitive to outliers or extreme values in a way that they lead to over-estimation of the correlation. Another matter is that a restriction of the range in either of the variables directly reduces the value of r. In general it is recommended to begin with plotting the variables against each other in order to investigate the relationship and these assumptions, and also provide these graphs together with the statistics in publication of the results (completely different graphs can correspond to the same r!). Graphing the variables also lets us know whether a linear relationship should be assumed not. Perhaps we see a quadratic or exponential relationship, and in this case we may simply transform one variable and use the linear methods for simplicity. If we do not want to consider the function of the relationship, but simply test whether there is a monotonic relationship (both variables increasing in the same or the opposite direction), we can apply the Spearman's rank-order correlation.

4.6 Regression methods

The relationship between two variables can also be described mathematically by *regression* methods. Perhaps we have found a significant association between bioimpedance and total body water (TBW), and we want to find the function that expresses their relationship. The basic method is simple linear regression. It estimates the straight line that best fits with the variables plotted against each other, and provides the intercept and slope of this line. Using the intercept and slope values together with the bioimpedance measurements, we can make *predictions* about the TBW. The difference between the measured TBW and the predicted TBW is called the *residuals*. The residuals can be used for inspecting whether or not the selected model (in this case linear) is valid. A scatterplot of the residuals (in this case bioimpedance vs residuals) provides information on linearity, homoscedasticity, normality and independence of the error terms, which are all assumptions for doing regression analysis. If for instance we see that the distribution of points is asymmetrical around the x-axis (also called skewed), the normality assumption might be violated (the *robust regression* method is an alternative to the ordinary method using least squares, which can be appropriately used for non-normal distributions or outliers). From the residuals we also obtain a value for the error of the prediction, such as the root-mean square error (RMSE) which is the average of how close the regression line is to all of the points. Both the r, R^2 and RMSE are measures for goodness of fit, but the RMSE is better at telling us how large an error is in terms of the quantity we are trying to predict. It is important to note that the RMSE or similar error values represent the error in the sample of measurements we have analyzed, and not the whole population.

Perhaps we have found that the RMSE of our TBW prediction based on bioimpedance was rather large, and that we need to reduce this error if we want to develop a TBW device. We might know of other factors which are also related to TBW, or factors which give changes in bioimpedance but are not related to TBW (confounding variables). If these variables are independent, we may be able to reduce the prediction error by including them as predictor variables in a *multiple linear regression*. For instance we may add BMI and Age together with the assumption that they will correct for the discrepancies in the estimation of TBW from bioimpedance. The multiple regression gives us an R^2, and the same error quantities such as the RMSE. Again, these statistics represent the sample and not the population, and the results from a multiple regression are optimized for the sample it is based on. The chance of spurious inflation of the statistics and a type-I error increases as more predictor variables are added, and it has been recommended that the number of data points should be at least 10 times the number of predictor variables [11, page 1456]. It is advised to keep the number of predictor variables low (5 or less), and that these variables are close to uncorrelated, in order to find the simplest possible predictive model. Including all points of a bioimpedance frequency sweep is an example of a bad prediction set for multiple regression.

Selection of the prediction variables may be done based on preexisting empirical data or theory, but there are also semi-automatic methods which can assist in sifting out redundant predictors. One of these methods is called *stepwise regression*, which suggests a model based on successively adding or removing variables based on the t-statistics of their estimated coefficients. This procedure can either be *forward* or *backward*. In the forward procedure, variables are added one by one based on the degree to which they produce a significant increase in prediction until the addition of more variables no longer make significant contributions. A similar procedure is employed for backward selection, but beginning with a model including all predictor variables. In general, the forward procedure is best suited for finding few significant predictors among

many candidates, while the backward procedure is best suited for elimination of few variables for fine-tuning of a pre-selected model. Still, there is always a chance of over-fitting (obtaining a poor model with spuriously inflated statistics), and it is always recommended that the resulting regression model is validated (more on this in chapter 5).

For data as e.g. bioimpedance frequency sweeps, when we may have many predictor variables compared to the number of observations, and/or highly correlated predictor variables, there are other type of regression methods which could be more suitable such as *principal components regression (PCR)* and *partial least squares regression (PLS)*. Both methods employ transformations of the initial set of predictor variables by constructing linear combinations of these into a new set of orthogonal (independent) components. In PCR, these components, which may explain most of the variance of the predictor variables, are regressed against the dependent variable. Often the first few principal components explain most (typically more than 90%) of the variance, and these components are used in the PCR model. However, this selection of regression set is based on including the largest variance of the predictor variables, not the independent variable. The PLS deals with this problem by finding the components of the predictor variable set which are most relevant to the dependent variable (by a simultaneous decomposition of both the predictor variables and the dependent variable with the constraint of explaining as much as possible of the covariance between them). Selecting how many components to include is a matter for both PCR and PLS and has for long been subject for discussion in the statistical field [12]. In general, some kind of cross-validation (see chapter 5.2) should be used in order to compare the predictive ability as a function of the number of components included.

4.7 Classification methods

Until now, we have been dealing with predictions of a continuous outcome as the dependent variable. Suppose we have investigated bioimpedance with respect to tissue type and found a significant difference between two types of tissue, and now we want to test how well bioimpedance could be used to discriminate between these tissue types. *Logistic regression* is such a method, which finds an optimal model based on the predictors and calculates the percentage of correct classification for each of the categories and for the overall classification. For more than two outcomes, the method is referred to as *multinomial logistic regression*. The independent variables can be continuous and real-valued, binary, categorical or a combination of these types. The R^2 statistic (calculated as in linear regression) gives a misleading indication of the goodness of fit for logistic regression, and several alternative analogues have been suggested (see e.g. ch 9.5.1. in [13]). In the case that the outcome is ordinal (i.e. good, better, best), another variant called *ordered logistic regression* is suitable. The logistic regression methods do not require normal distribution of the predictor variables [14, p.575].

Another classification method is the *discriminant function analysis*, which instead of regression as the mathematical framework is based on the same principle as the MANOVA. While the MANOVA deals with whether a number of groups differ significantly with respect to differences on a number of dependent variables, the discriminant function analysis deals with whether a linear combination of predictor variables can differentiate between the groups. The assumptions are normality of the predictor variables, homoscedasticity, linear relationships between all predictor variables within each group and absence of multicollinearity and outliers in the predictor variables (same as for MANOVA) [9]. The discriminant function analysis is in general reasonably robust against violations of these assumptions, especially for large samples and equal number of observations per group [9].

Other classification methods used in biomedical research include *artificial neural networks (ANN), decision trees, support-vector machines (SVM), naïve Bayes classifier* and *k-nearest neighbors*. These methods are important approaches in the field of machine learning, where algorithms are being developed by learning from sample data in order to classify unseen data. These methods are increasingly being adopted in the biomedical engineering fields, including bioimpedance.

ANN is a classification method inspired by the workings of the central nervous system. By constructing a network of interconnected nodes (“neurons”) organized in layers, a classification algorithm is developed (also called trained) by optimizing the weights of each node-to-node connection, representing the connection strength between them. As new inputs of selected features are fed through the network, the output layer at the end of the network will provide the suggested classification.

Decision trees are based on an algorithm for splitting the input data in a way that maximizes the separation of the data, resulting in a tree-like structure [15]. There are algorithms that can suggest the structure of the tree such as the Hunt’s algorithm, but these algorithms usually employ a greedy strategy that grows a tree by making a series of locally optimum decisions. Another drawback with the method is that continuous variables are implicitly discretized by the splitting process, losing information along the way [16].

SVM has become popular due to the performance the method has demonstrated in problems such as handwriting recognition. The principle is based on representing the data as points in space, and then finding an optimal surface called a hyperplane, which maximizes the margin between the classes. If the classes are not linearly separable in the original data space, the data is mapped into a much higher dimensional space (called *feature space*) by employing a mathematical projection called the *kernel trick*, where a new

hyperplane is found. This makes the SVM also efficient for non-linear classification.

Naïve Bayes classification uses the Bayes' theorem together with a "naïve" independence assumption to calculate probabilities of class membership. Predicting class membership can be done directly using Bayes' theorem with only one feature and the prior probability. As an example, consider we are investigating bioimpedance as a marker for wound healing. Suppose we gathered 100 measurements after wounding in an experiment where 50 of the wounds healed by themselves. Among the wounds that healed, the impedance increased during the healing process in 35 of the 50 wounds, and in 5 out of the 50 wounds that did not heal. We can now calculate the conditional probability of a wound belonging to the "healing" class based on whether or not the impedance increases using the Bayes' Theorem, giving us 88% if impedance increases and 25% if the impedance does not increase. When including several conditional features, the mathematics would normally become problematic due to the relations between the features, but the naïve Bayes classifier assumes that all the features are independent which allows for easy computation.

The k-Nearest Neighbors (kNN) algorithm predicts the class of a point in a feature space based on the known attributes of the neighboring k number of points in this space. For instance, say we want to predict tissue status based on a set of independent bioimpedance features, such as the Cole parameters. Using a dataset of measurements with known tissue states, the kNN algorithm will first construct class-labeled vectors in a multidimensional space with one dimension for each feature. Class prediction of a new measurement will then be done based on the majority of class-memberships of the k number of nearest neighbors based on the distance (usually Euclidian) to the new point.

These classification methods use different principles and rules for learning and prediction of class membership, but will usually produce a comparable result. Some comparisons of the methods have been given [i.e. 17, 18]. Although the modern methods such as SVM have demonstrated very good performance, the drawback is that the model becomes an incomprehensible "black-box" which removes the explanatory information provided by e.g. a logistic regression model. However, classification performance usually outweighs the need for a comprehensible model. Principal component analysis (PCA) has been used for classification based on bioimpedance measurements. Technically, PCA is not a method for classification but rather a method of data reduction, more suitable as a parameterization step before the classification analysis.

1. Validation methods

Until now, we have been dealing with exploratory methods, where the bioimpedance measurements have been used to explore differences between groups of measurement, effects of different factors or associations between bioimpedance and other parameters. We have also been dealing with predictions of either continuous or discrete outcomes, but not the *validation*[1] of these. If we have come one step further and developed a potentially useful method based on our research, we need other types of testing and statistical methods to validate its performance. These statistics are very important, as they will mainly determine how good the developed method is, together with for instance availability, usability, price etc.

5.1. Evaluating performance

For bioimpedance measurements, the performance is in most cases determined by the agreement between a developed bioimpedance parameter and a reference ("gold standard"). For example, if we develop a probe to detect breast cancer, we need to find out how often it correctly detects cancerous tissue (the sensitivity) and how often it correctly detects healthy tissue (the specificity). These two statistics are what the potential users will mainly look for when considering the method. Our approach will then be as follows. We have already explored the difference in bioimpedance between healthy and cancerous tissue by the procedure shown in figure 1, and based on our previous results we have also selected which bioimpedance parameters and algorithm we will use for discriminating between the two tissue types. We now do a new study using the selected method on a new sample of subjects. The sample size should be adequate in order to obtain an estimate with acceptable precision, and can be estimated based on the prevalence and the anticipated sensitivity and specificity [20, 21]. Say we did 500 measurements, among which 100 were confirmed positive by a reference measurement. Among these 100, our method detected 85 as positive, and among the 400 negative, our method detected 350 as negative. We can now calculate the sensitivity and specificity by:

$$Sensitivity = \frac{TP}{TP + FN} \quad (1)$$

$$Specificity = \frac{TN}{TN + FP} \quad (2)$$

Where TP is the number of true positives (85), FN is the number of false negatives (100-85=15), TN is the number of true negatives (350) and FP is the number of false positives (400-350=50). Our sensitivity and specificity then becomes 85% and 87.5% respectively. It is also often of interest to see how the sensitivity and specificity depends on the decision threshold (i.e. the level of our bioimpedance parameter which separates healthy and cancerous tissue). The ROC (receiver operating characteristic) curve is

[1] Provision of objective evidence that a given item fulfils specified requirements, where the specified requirements are adequate for an intended use [19]

constructed by plotting the sensitivity (the true positive rate) against 1-specificity (the false positive rate) for the whole range of decision thresholds. This plot allows us to see what kind of sensitivity and specificity we may obtain according to what we consider important with respect to the application. The area under the curve (AOC) is usually reported together with the ROC curve as a measure of total classification performance.

A very relevant case in the field of bioimpedance is the validation of an estimate of a continuous physiological parameter, where we want to find out how well our estimate agrees with a reference measurement of this parameter. Among the methods used for evaluating agreement in medical instruments measuring continuous variables, the Bland-Altman method [22] is the most popular [23]. By this method, a plot is constructed with the means of all measurement pairs (estimate and reference pairs) on the x-axis and the difference between them on the y-axis. In addition to this, the mean difference line is plotted along with two lines representing the 95% limits of agreement (LOA), given by the mean difference ± 1.96 standard deviation of the difference. By this simple method, the reader can easily see how much the two measuring methods differ according to the magnitude of the measurement, and also inspect for systematic differences such as bias or trends. The LOA tell us that most (95%) of the measurements had a difference within the upper and lower LOA. It is not possible to give any general criterion for an acceptable LOA because it depends on the intended use of the proposed method. As an example, limits of agreement of up to ±30% have been recommended as acceptable for introducing new techniques within cardiac output measurements [24]. It is important to note that the correlation coefficients or the coefficient of determination is not sufficient for reporting agreement, as two variables may have a perfect linear relationship but at the same time be very different in magnitude. Another type of correlation which avoids this problem is the *intraclass correlation coefficient* (ICC). Although it was originally devised to assess reliability (see chapter 5), it has also been used to assess agreement [23]. The appropriateness for this use of the ICC has been criticized [25] and considered doubtful [23], but also regarded as the best traditional approach for assessing agreement [26].

5.2. Cross-validation

Models should always be validated in order to avoid overfitting and inflated performance results. When a predictive model, such as a regression model, is developed and tested based on the same sample of measurements, there is a chance that the model parameters are optimized in a way that fits better with the sample than the population it comes from and produces an overoptimistic result. This is more relevant the more complex (i.e. number of independent variables) the model is. Therefore, the model should always be tested against an independent sample in order to see how well the model will generalize and perform in practice. The model can be validated by replicating the results on one or more independent samples from the same population, but in most cases it is more practical to split the data in one part which is used to develop the model parameters (the training sample) and use the remaining data (the validation sample) to test the performance of the model. The validation data can then play the role of "new data" as long as the data are independent and identically distributed [27]. This is called the *hold-out method* [28] and is performed by splitting the data in a training sample and a validation sample with e.g. 2/3 for training and 1/3 for validation. The training sample is used to fit the model parameters against the independent variable, and the model along with its fitted parameters is then used to predict the independent variable in the validation sample, based on the predictor variables in the validation sample. A large difference between the results of training and validation is an indication that the model is wrong. Whereas one such split yields a validation estimate, averaging over several splits is called *cross-validation* [29]. Cross-validation overcomes the risk of a misleading performance (i.e. accuracy, sensitivity, specificity) estimate due to an "unfortunate" split in the hold-out method. There are several procedures which can be employed for cross-validation, among these the most popular are: (See [27] for a comprehensive list of cross-validation procedures):

Random subsampling. The data is split by random selection of data points without replacement into training and validation samples. The model is then fit to the training sample and the performance of the model is evaluated on the validation sample. This is repeated a number of times, and the total performance is found by averaging the performance of all iterations. This process is similar to the hold-out method and is also referred to as repeated hold-out. A drawback of this method is that different validation sets may overlap.

k-Fold Cross-Validation. The data is first divided into k (e.g. 10) parts as evenly as possible. Each part is used in turn as a validation sample and the remaining for training. The performance results are then averaged from the k runs to provide an overall estimate.

The leave-one-out method. Each data-point is successively "left out" from the sample and used for validation. The performance is calculated for the left-out data point based on the fitted model obtained from the remaining data. The average of all iterations gives an estimate of the overall performance. This method makes maximum use of the data, but is also computationally expensive because the number of iterations is the same as the number of data points.

5.3 Concepts of performance

There are several terms which are important in validating a new measurement method. A list of the most relevant aspects of validation is given in table 1, along with a

definition of each term and how it is usually reported. The definitions of the terms vary among different fields and standards, sometimes giving an inconsistent meaning. The table is an attempt at giving an unambiguous overview of the terms based on the most common uses.

The concept of *error* in a measurement is quite straightforward, and is the difference between the measured value and a reference value. If the error in replicate measurements remains constant or varies in a predictable manner, the error is referred to as a *systematic measurement error*. If the error varies in an unpredictable manner, it is referred to as *random measurement error*. The measurement error can be a combination of the two.

The term *agreement* can be regarded as a general term for the degree to which the measurements are identical (either in nominal, ordinal or continuous variables) and it is of main interest in method comparison studies.

Accuracy is the closeness of agreement between the result of a measurement and a true value, and depends on both *trueness* and *precision*. The difference between trueness and precision is easiest explained through the example of throwing darts. Trueness is high if the darts are centered around the middle, but low if they are all on one side of the board (bias), regardless of how much they are spread. Precision is high if they are close and low if they are spread far apart, regardless of the center they are spread around. Precision is further divided into *repeatability* and *reproducibility* according to the measurement condition. When new measurements are taken with the same setup by the same operator on the same items/subjects (i.e. replicated), the repeatability of the method is tested. When new measurements are taken with the same method on the same items/subjects but with different devices and operators, the reproducibility is tested. Repeatability can be thought of as the minimum variability between results, and reproducibility the maximum variability between the results. With measurement of thoracic bioimpedance as an example, the repeatability of the method can be assessed by replicating the measurement by the same operator using the same equipment (i.e. device and electrodes) on the same subjects, with measurements taken in quick succession such as on the same day. When clinical implementation is considered, it is also important to know how large this variation becomes under realistic conditions. Factors such as electrode positioning (operator related), calibration (device-related) and ambient humidity (laboratory-related) may cause variations in the measurement. The reproducibility of the method can then be assessed by performing measurements on the same subjects at two or more different laboratories having different operators and equipment (but of the same type), providing a realistic estimate of the precision. Specific reproducibility, such as inter-electrode reproducibility, can be assessed for the factors which influence the measurement, telling us how these factors influence the measurement precision.

Agreement and *reliability* are two distinct concepts in the medical literature [30, 31, 32]. While agreement is the degree to which scores or ratings are identical, reliability is the ability of a measuring device to differentiate among subjects or objects [31]. Agreement concerns the measurement error while reliability relates the measurement error to the variability between the subjects or items which are tested [30]. Reliability is assessed during certain conditions such as different equipment or users (inter-rater reliability) or with the same equipment and users (intra-rater or test-retest reliability). As an example, if we test our impedance measurement system against a set of calibration resistors once each month, and each time measure a 10% positive offset, the system has a low agreement (and accuracy), but a high test-retest reliability. Given these definitions, the test-retest reliability may seem to be the same as the repeatability of a measurement, but we make a distinction here. Repeatability is assessed through repeated measurements on identical subjects/items within a short time relative to any changes in the property being measured, whereas test-retest reliability is assessed from measurements taken at different occasions with the same conditions, and allowing changes in the property being measured. The same goes for reproducibility vs inter-rater reliability in that reproducibility is assessed using identical test items under different conditions (which is the source of variation) while inter-rater reliability also involves testing under different conditions, but in addition allows changes in the property being measured. This also implies that precision and reliability are two different concepts.

An advantage of using reliability to compare measurement methods is that it can be used to compare methods when their measurements are given on different scales or metrics [32]. For continuous variables, reliability is usually determined by the ICC. The ICC is a ratio of variances derived from ANOVA, with a maximum value of 1.0, indicating perfect reliability. There are different types of the ICC, including one- or two-way model, fixed or random-effect model, and single or average measures (see [33] for more on selection), and the type should be reported in a reliability study [34]. For assessing reliability in categorical data, *kappa* statistics such as Cohen's kappa provide useful information [31]. Instead of simply taking the percentage of equal decisions relative to the total number of cases, Cohen's kappa provides a measure of association which is corrected for equal decisions due to chance.

Which of these measures to report should be chosen based on how the measurements are to be used in the future. The same goes for the importance of the measurement performance. A certain degree of measurement error may be acceptable if measurements are to be used as an outcome in a comparative study such as a clinical trial, but the same errors may be unacceptably large in individual patient management such as screening or risk prediction [32]. For some applications, there are specific ways of reporting performance which have become standard, such as the Clarke-Error Grid together with MARD (Mean absolute relative deviation) for blood glucose measurement.

Table 1. List of important terms in the validation of new measurement technology along with the most usual and recommended ways of reporting. [1]There are numerous different definitions in the literature, which can be inconsistent and confusing. These definitions provide one version with the aim of reducing ambiguity. [2]Accuracy has previously been defined as the same as trueness only, but with ISO 5725-1 [37], and reflected in the JCGM 200:2012 [19], the definition of accuracy has for the most changed to include both trueness and precision as given here. The old definition is still in use in some areas.

Term	Definition[1]	Reported as
Measurement error	Measured quantity value minus a reference quantity value [19] Systematic measurement error: Component of measurement error that in replicate measurements remains constant or varies in a predictable manner [19] Random measurement error: Component of measurement error that in replicate measurements varies in an unpredictable manner [19]	Quantity on the same scale as the measurement scale, relative error, percentwise error, mean square error, root mean square error. Systematic measurement error: Bias Random measurement error: Standard deviation, variance, coefficient of variation.
Sensitivity	The sensitivity of a clinical test refers to the ability of the test to correctly identify those patients with the disease. [36].	Eq. (1) A part of the ROC curve which shows the relation between sensitivity, specificity and the detection threshold.
Specificity	The specificity of a clinical test refers to the ability of the test to correctly identify those patients without the disease. [36]	Eq. (2) A part of the ROC curve which shows the relation between sensitivity, specificity and the detection threshold.
Agreement	The degree to which scores or ratings are identical [31]	Continuous: Bland-Altman plot Discrete: Percentage agreement
Trueness	Closeness of agreement between the average value obtained from a large series of results of measurement and a true value [37].	Bias (i.e. the difference between the mean of the measurements and the true value)
Precision	Closeness of agreement between independent results of measurements obtained under stipulated conditions [37].	Standard deviation, coefficient of variation
Repeatability	Precision determined under conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time [37]	Within-subject standard deviation [38] Repeatability coefficient [38]
Reproducibility	Precision determined under conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment [37]	Standard deviation, coefficient of variation
Accuracy[2]	Closeness of agreement between the result of a measurement and a true value (both trueness and precision) [37] Measurement accuracy: Closeness of agreement between a measured quantity and a true quantity value of a measurand [19]	Bias (trueness) and standard deviation/ coefficient of variation (precision) Diagnostic accuracy: Sensitivity and specificity Sensitivity and specificity corrected for prevalence as: *(sensitivity)(prevalence)+(specificity)(1-prevalence)* [39]
Reliability	Ratio of variability between subjects or objects to the total variability of all measurements in the sample [31]	Intraclass correlation coefficient Kappa statistics (categorical data)

At last, it is important to also mention the concept of *validity*, originating from psychometrics and addresses the inference of truth of a set of statements [35]. A study may provide perfect test results on accuracy, but if the experiments are not testing what it is supposed to, the results are not valid. For instance, testing the agreement between a new method and an existing method with barely acceptable clinical accuracy may provide a good agreement between the two, but the results are not valid with respect to the accuracy of the new method. Validity is also used to describe the same concept as trueness within psychometrics.

References

1. Hulley SB, Cummings SR, Browner WS, Grady D, Hearst N, Newman TB. 2nd ed. Philadelphia: Lippincott Williams and Wilkins; 2001. Getting ready to estimate sample size: Hypothesis and underlying principles In: Designing Clinical Research-An epidemiologic approach; pp. 51–63.
2. Bacchetti P, Wolf LE, Segal MR and McCulloch CE. Ethics and sample size. Am. J. Epidemiol. 2005;161(2):105-110. http://dx.doi.org/10.1093/aje/kwi014
3. Bonferroni CE. Teoria statistica delle classi e calcolo delle probabilit `a. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936;8:3-62.
4. Miller, RG. Simultaneous statistical inference. 2nd ed. Springer Verlag. 1981, pp. 6-8. http://dx.doi.org/10.1007/978-1-4613-8122-8
5. Grimnes S and Martinsen ØG. Bioimpedance and Bioelectricity Basics. 2nd edition. Academic Press. 2008.
6. Bland JM and Bland DG. Statistics Notes: One and two sided tests of significance BMJ. 1994;309:248. http://dx.doi.org/10.1136/bmj.309.6949.248
7. Dowdy S, Wearden S and Chilko D. Statistics for research. 3rd Edition, Wiley-Interscience. 2011.
8. Schmider E, Ziegler M, Danay E, Beyer L and Bühner M. Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. Methodology: Europ. J. Res. Meth. Behav. Social Sci. 2010;6:147-151.
9. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. CRC Press. 2011.
10. Sawilowsky SS, Blair RC and Higgins JJ. An Investigation of the Type I Error and Power Properties of the Rank Transform Procedure in Factorial ANOVA. J. Educ. Stat. 1989;14:255-267. http://dx.doi.org/10.2307/1165018
11. Mariscuilo LA and Levin JR. Multivariate statistics in the social sciences: A researcher's guide. Brooks/Cole Pub. Co. (Monterey, California). 1983.
12. Peres-Neto P, Jackson DA and Somers KM. How many principal components? Stopping rules for determining the number of non-trivial axes revisited. Comput. Stat. & Data Anal. 2005;49:974-997. http://dx.doi.org/10.1016/j.csda.2004.06.015
13. Ryan TP. Modern Regression Methods. 2nd ed. Wiley series in probability and statistics. Wiley. 2008
14. Tabachnick BG and Fidell LS. Using multivariate statistics (3. ed.). Pearson. New York. 1996.
15. Breiman L, Friedman J, Olshen R and Stone C. Classification and Regression Trees. Belmont, California: Wadsworth. Dowdy, Wearden, Chilko. Statistics for Research. Wiley Series in Probability and Statistics. 1984.
16. Dreiseitl S and Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J. Biomed. Inform. 2003;35:352-359. http://dx.doi.org/10.1016/S1532-0464(03)00034-0
17. Kotsiantis SB. Supervised Machine Learning: A Review of Classification Techniques. Informatica. 2007;31:249-268.
18. Rani P, Liu C, Sarkar N and Vanman E. An empirical study of machine learning techniques for affect recognition in human-robot interaction. Pattern Anal Applic. 2006;9:58-69. http://dx.doi.org/10.1007/s10044-006-0025-y
19. JCGM 200:2012. International vocabulary of metrology – Basic and general concepts and associated terms (VIM) 3. ed.
20. Buderer NM. Statistical methodology: I. Incorporating the prevalence of disease into the sample size calculation for sensitivity and specificity. Acad. Emerg. Med. 1996;3(9):895-900.
21. Malhotra RK and Indrayan A. A simple nomogram for sample size for estimating sensitivity and specificity of medical tests. Indian J Ophthalmol. 2010;58:519-522. http://dx.doi.org/10.4103/0301-4738.71699
22. Bland JM and Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;8:307-310. http://dx.doi.org/10.1016/S0140-6736(86)90837-8
23. Zaki R, Bulgiba A, Ismail R and Ismail NA. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. PLOS one. 2012;7:e37908. http://dx.doi.org/10.1371/journal.pone.0037908
24. Critchley LAH and Critchley JAJH. A Meta-Analysis of Studies Using Bias and Precision Statistics to Compare Cardiac Output Measurement Techniques. J. Clin. Monitor. Comput. 1999;15:85-91. http://dx.doi.org/10.1023/A:1009982611386
25. Bland JM and Altman DG. A Note on the use of the Intraclass Correlation Coefficient in the Evaluation of Agreement between two Methods of Measurement. Comput. Biol. Med. 1990;20:337-340. http://dx.doi.org/10.1016/0010-4825(90)90013-F
26. Lin L. Overview of Agreement Statistics for Medical Devices. J. Biopharm. Stat. 2007;18:126-144. http://dx.doi.org/10.1080/10543400701668290
27. Arlot S and Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys. 2010;4:40-79. http://dx.doi.org/10.1214/09-SS054

28. Devroye Luc P and Wagner TJ. Distribution-Free performance Bounds for Potential Function Rules. IEEE Trans. Inform. Theory. 1979;25:601-604. http://dx.doi.org/10.1109/TIT.1979.1056087

29. Geisser S. The predictive sample reuse method with applications. J. Amer. Statist. Assoc. 1975;60:320-328.

30. De Vet HCW, Terwee CB, Knol DL and Bouter LM. When to use agreement versus reliability measures. J. Clin. Epidem. 2006;59:1033-1039. http://dx.doi.org/10.1016/j.jclinepi.2005.10.015

31. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, Roberts C, Shoukri M and Streiner DL. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int. J. Nursing Studies 2011;48:661-671. http://dx.doi.org/10.1016/j.ijnurstu.2011.01.016

32. Barlett JW and Frost C. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. Ultrasound Obstet. Gynecol. 2008;31:466-475. http://dx.doi.org/10.1002/uog.5256

33. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J. Strength Cond Res. 2005;19:231-240.

34. Zaki R, Bulgiba A, Nordin N and Ismail NA. A systematic review of statistical methods used to test for reliability of medical instruments measuring continuous variables. Iranian J. Basic Med. Sci. 2013;16:803-807.

35. Nunnaly JC and Bernstein IH. Psychometric Theory. New York: McGrwaw-Hill. 1994.

36. Lalkhen AG and McCluskey A. Clinical tests: sensitivity and specificity. Contin. Educ. Anaesth. Crit. Care. Pain. 2008;8:221-223. ttp://dx.doi.org/10.1093/bjaceaccp/mkn041

37. ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results – Part 1: General principles and definitions

38. Bland JM and Altman DG. Measuring agreement in method comparison studies. Stat. Methods Med. Res. 1999;8:135-160. http://dx.doi.org/10.1191/096228099673819272

39. Metz CE. Basic principles of ROC analysis. Semin. Nucl. Med. 1978;8:283-298. http://dx.doi.org/10.1016/S0001-2998(78)80014-2