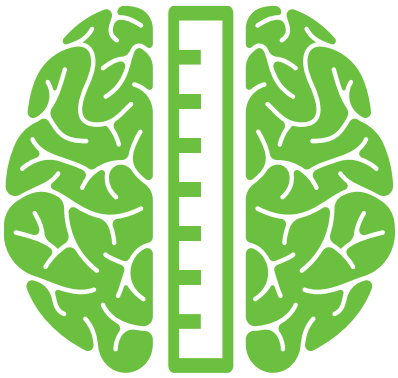

EUROTEQ

Clinical measurement literacy for
EUROpean Transparency and EQuality in health

THEORETICAL REPORT
PART I

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

Research basics

This report summarizes some key aspects of clinical research in the health sciences. Such research may be of very different kinds, and can very broadly be classified as either quantitative, qualitative or mixed-methods research, depending on the type of methods that are used to collect and analyze data. All three have their definite role and place and neither one is "better" than another; the choice ultimately depends on the objectives or research questions that underpin a particular study.

In essence, quantitative research is based on numerical data (that in turn may be either qualitative or quantitative by nature, which we will return to later). It aims to describe and summarize the amounts or magnitudes of various phenomena, and to understand how various phenomena are related. This is done with the intention to draw generalizable conclusions that hold not only in the particular study, but also in other situations or groups of people, and to be able to make predictions (e.g., regarding the effectiveness of a certain intervention).

In contrast, qualitative research is not based on numerical data but on audio or text data (it may involve other types, e.g., observations, pictures or movies, but that is less common). The most common type of qualitative data in the health sciences are text transcripts of interviews or stories regarding the phenomenon that one wants to study. In general, these are analyzed in order to summarize and gain a deeper understanding of the nature of various phenomena. That is, in qualitative research one wants to understand what something is like, what it means to live with a certain condition, people's experiences, views or thoughts regarding certain things or situations, and so on. There is a wide range of methods and approaches to conduct qualitative research and to analyze qualitative, including (but not limited to) content analysis, thematic analysis, grounded theory, hermeneutic and phenomenological. Finally, with mixed-methods research, a combination of quantitative and qualitative approaches is used within the same study.

This report focuses on quantitative research, and particularly on studies that intend to evaluate the effectiveness or value of a certain intervention or treatment, i.e., interventional studies. This is not because quantitative research is "better" (again, the "best" approach is governed by the study objectives), but because this is the most common approach in clinical research and because it is typically this type of studies that underpin clinical guidelines and recommendations. In addition, there are a number of basic aspects of quantitative research that are necessary to understand in order to understand and critically review research reports – some of these aspects are not seldom neglected or misunderstood, but important to grasp in order to enable effective and correct communication with patients and their family members as well as with other health care professionals (HCPs).

In addition, we refer to Appendix 1, which provides brief explanations of common terms and concepts in clinical research, some of which may not be explained in the text.

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

Evidence-based practice

Over the past several decades, the principles of evidence-based practice (EBP) have gained increased importance and typically underpin clinical guidelines and recommendations. EBP can be defined as the integration of clinical expertise with the best available external clinical evidence from systematic research (Sacket et al. 1995). In terms of the best available clinical evidence from systematic research, a key aspect is research design, which is about how a study is conducted. This includes how participants are recruited (or sampled) to the study, how many people that participate in the study, by what means data are collected, what the study procedures look like (e.g., when, how and by whom data are collected, and when, how and by whom the study intervention is provided), and how data are analyzed. Factors such as these all contribute to the extent to which study results can be trusted, are interpretable, representative, and generalizable.

There is a general hierarchy in terms of the levels of evidence, i.e., the extent to which research results are considered solid and trusted. This hierarchy is illustrated in Figure 1 and is largely based on the employed research design. The highest level is represented by systematic reviews of published research, preferably coupled by a meta-analysis of data from the original studies included in the review. The term meta-analysis means that a new statistical analysis of published data is conducted. The results from a meta-analysis synthesizes previously reported data into new results that represent the current knowledge level regarding, for example, the efficacy or value of a certain intervention (Impellizzeri & Bizzini 2012; Haidich 2010). If systematic reviews and meta-analyses are lacking, results from individual randomized controlled trials (RCTs) are considered the second-best source of evidence, followed by nonrandomized trials, observational studies, case series and, at the lowest level of the evidence hierarchy, expert opinions.

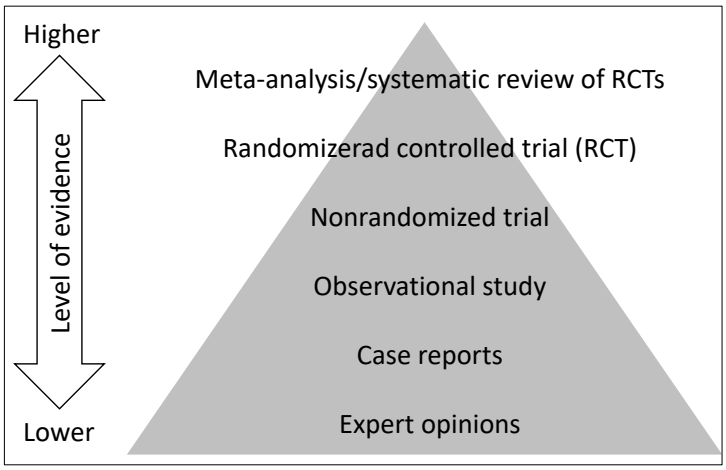

Fig. 1 Hierarchy of evidence

Research design

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

As is evident from this hierarchy, randomization is a key research design aspect, particularly in interventional studies. However, there are several other research designs (Grimes & Schulz 2002; Miller et al. 2020; Noordzij et al. 2009), of which some of the most common will covered later. The importance of randomization is because randomizing study participants to, for example, the active intervention group or the control group is the best and most effective way to control for any bias in study results (Kang et al. 2008). By letting chance decide which group each participant will belong to, the likelihood of having comparable groups at the outset of a study is maximized; if groups differ in any potentially important respect (e.g., age) there is a risk that this may account for some aspects of any differences in study results that may be observed between the groups (i.e., there is a bias). The intervention group (sometimes also referred to as the active group) consists of participants that receive the intervention that is being studied, and the control group receives an alternative intervention that is used as a comparator in the study. This may be no intervention at all, usual care, or placebo. Placebo is an inactive intervention that is as similar as possible to that delivered to the intervention group but lacks the critical component(s) that is believed to account for the outcome. The classical placebo-controlled study is a drug trial, where the placebo pill looks, tastes, and feels exactly like the active pill but does not contain the pharmacological active substance. But other examples also exist, such as in information- or education-based interventions where the placebo condition does not include the aspects believed to be important for achieving the desired outcome. To further strengthen the design, it is also desirable to leave participants as well as clinicians unaware of which group they have been randomized to. This is referred to as single (study participants do not know which group they belong to) and double blind (neither participants nor clinicians know who has been allocated to what group) studies. For ethical reasons, it is however important that all participants are aware of exactly what it means to be randomized to either of the study groups and that they will not be aware of their allocation until the study has been concluded. Furthermore, if the study results are positive, participants in the control group should be offered the intervention if results are favorable at the conclusion of the study (World Medical Association 2018).

In certain contexts, particularly pharmacological drug trials, studies are classified as according to four phases (Phase I-IV), reflecting the stage of development of the intervention being studied. Phase I studies concern the safety of the new intervention and is conducted with relatively small groups (up to maybe 100) of people, typically (in the case of a new drug) health volunteers, whereas Phase II studies focus on determining whether the intervention works as intended (i.e., to see if it has the intended effect, which is not clear at the outset of a Phase II study). If the intervention passes Phase II (i.e., is effective), Phase III studies are conducted to compare it to other interventions (or placebo/standard care) and to further monitor side effects. With commercial interventions (e.g., new drugs) Phase III studies are the basis for licensing and marketing of the product to the wider population. Finally, Phase IV studies are conducted after the intervention has been made available, typically to study its “real world” and/or long-term effectiveness.

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

In many instances it can be practically difficult or even impossible to randomize study participants to the active intervention and control groups or, indeed, to have a reasonably comparable control group at all, while at the same time ensure that there is no interaction or "spillover effects" between groups. Some examples of interventions that are difficult to control in such a manner include educational, psychosocial, care delivery and self-management interventions. Another situation may be when an intervention first is tested, and the investigators cannot afford the time, effort, or money to conduct a full-scale RCT. Regardless of the reason(s), study designs that do not include randomization are referred to as nonrandomized or quasi-experimental studies. A study that includes a control group but lacks randomization is sometimes also referred to as controlled clinical trial (CCT) but is still a quasi-experimental study. In essence, a quasi-experimental design refers to any type of study that lacks randomization, a control group, or both. This contrasts with true experimental designs, such as RCTs, which employ randomization and a control group.

Observational studies differ from the previously mentioned types of studies in that the investigator does not intervene or do anything except for collecting the data. There are three basic types of observational study designs: cohort studies (a group of people are followed prospectively over time), case-control studies (a group of people with a certain characteristic are studied retrospectively, typically together with a control group that is very similar to the study group except for a certain characteristic), and cross-sectional studies (data are collected at one point in time). Cohort and case-control studies are typically used in epidemiological research, where it is of interest to study what exposures or characteristics that may predict future events (cohort studies) or what exposures or characteristics in the past that may be associated with developing something (case-control studies). In addition, cohort studies are employed to study the natural history of various conditions. Since cross-sectional studies rely on data collection at one point in time, they lack the time perspective and only provide a "snapshot" of the studied variables. As such, their use is limited to descriptive purposes and to explore various associations.

Similar to the important role of random allocation to study groups in interventional studies, random selection of study participants is the best and most effective way to ensure representative samples and representative results in non-interventional studies, such as surveys. This is because letting chance decide who to be considered for participation in a study maximizes the likelihood of the sample (i.e., the persons selected for participation) to mirror the population (i.e., the kind of persons that the study intends to learn about) with respect to characteristics such as age, sex, and other relevant variables. This is important because the more representative the study sample is of its population, the more likely it is for study results to be generalizable. Here it should be noted that it is the investigator that defines what the study population is. This is typically done by means of inclusion and exclusion criteria, which thus articulate what kind of people the study concerns, attempts to learn and draw conclusion about, and hence generalize to. In contrast, the group of people who actually participate in a study is referred to as the sample.

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

The process of identifying and selecting people for participation in a study is referred to as sampling (Elfil & Negida 2017; Shorten & Moorley 2014). There are many types of sampling methods available but most of them can be classified as either probability (i.e., random) or nonprobability (i.e., nonrandom) sampling methods. Essentially, probability sampling means that every person in the population has an equal probability of being selected to the study sample. As touched upon earlier, this is generally preferable to nonprobability sampling because it maximizes representativeness of the population. However, in many instances it may be difficult or even impossible to employ this type of sampling and some kind of nonprobability sampling will be used instead. Common probability sampling methods include simple random sampling (random selection of X number of people from a list of all people in the population), systematic random sampling (random selection of one person followed by selection of, e.g., every 10^{th} person from the initially randomized person), stratified random sampling (random selection from different pre-defined relevant subgroups of people), and cluster sampling (random selection of units, e.g., hospitals or geographical areas, rather than people, followed by selection of persons from those units). Common nonprobability sampling methods include convenience sampling (selection of participants according to their availability and accessibility, e.g., all persons who attend a certain clinic and fulfills inclusion-/exclusion criteria), quota sampling (similar to stratified sampling but without random selection), and snowball sampling (asking initially selected persons to identify other persons of the same kind; typically used in studies involving people that are difficult to access, e.g., homeless persons).

As mentioned earlier, research design is about how a study is conducted and underpins the extent to which study results can be trusted, are interpretable, representative, and generalizable. These aspects go back to what is conserved the validity of a study, which refers to how confident we can be regarding a study's results, conclusions, and related implications. There are several types of validity described in the literature, of which three main types are considered here: internal validity, external validity, and statistical conclusion validity. Internal validity concerns the extent to which we can trust the results in a study, whereas external validity concerns the extent to which those results may be generalized to other situations beyond the study at hand. Obviously, external validity thus requires internal validity, which in turn relies on research design issues such as those addressed above, i.e., randomization, control groups, sampling, and so forth. Additional aspects include methods and procedures related to data collection. For example, whether study outcomes are clinically relevant, if data were collected using instruments and protocols that are valid and reliable, and if results are interpretable and meaningful. These aspects will be addressed in some more detail later. The third validity aspect, statistical conclusion validity, concerns whether the number of participants in the study was appropriate to meet the objectives of the study and in relation to the methods used to analyze the data, as well as whether those methods are appropriate given study objectives and the type of data in the study. Such aspects will also be addressed later.

Validity issues become particularly pertinent in interventional and other studies that attempt to address causation, i.e., whether some effect or outcome is caused a particular thing (e.g.,

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

an intervention). To support a causal relationship, there are some basic requirements that need to be fulfilled (see, e.g., Greenhalgh 1997b). These are several and we will cover only a few here. First, there should be an association between the cause and the effect. However, this is not enough since an association may be seen for various reason. For example, if there is a strong association (e.g., a correlation) between X and Y, this does not in itself tell whether it is X causing Y or the other way around. Furthermore, there may well be something else (let's call it Z) that has a causal relationship with both X and Y, which makes it look like X and Y are strongly associated when they in fact are not (in this case Z would be considered a confounder). Second, there needs to be a temporal relationship between the cause and the effect. Specifically, the cause must pre-date the effect, otherwise it cannot be responsible for the effect. One implication of this is that it is impossible to draw any causal conclusions from cross-sectional data. Third, it should be possible to rule out alternative explanations for effect. This is an important reason why it is important to have a control group (to rule out, e.g., that the effect would not happen just by letting time pass) and to use random assignment of participants (to ensure the groups are comparable).

Numbers: Measurement and assessment

As mentioned initially, quantitative research is based on numerical data. Therefore, it is relevant to consider the meaning of numbers, what they may represent, and what they may not represent. In principle, numbers *in themselves* are abstract and meaningless, and will require outside information to be meaningful. Consider for example the number 39. Without any further information all we can say is essentially that it is a number between 38 and 40. To give it meaning we would at least need to know what variable it represents (e.g., time, temperature, distance) and the unit it expresses (e.g., years, degrees C, km). In addition, knowledge regarding its possible range and context (e.g., a person's age, body temperature, or the distance to the local GP office) will be needed. As mentioned before, this type of information needs to come from outside the data, and any calculations or statistical analyses will remain abstract and meaningless without it.

The meaning of numbers

While quantitative research concerns numerical data, it is important to be aware that numerical data can be either quantitative or qualitative (Sevens 1946). This distinction relates to what numbers represent (Table 1).

Table 1. Types and levels of numerical data

Type	Data level	Characteristics	Information provided	Example
Qualitative	Nominal	• Categories/numerical labels • No quantitative meaning	• Classification	Assessments (e.g., sex, ethnicity, blood type)

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

				0 = Female; 1 = Male
	Ordinal	• Ordered categories • Unknown (unequal) intervals between numerals	• Rank order	Assessments, rating scales (e.g., well-being, symptom severity) 0 1 2 3 4 5
Quantitative	Interval	• Linear • Equal intervals between numbers • Arbitrary (no absolute) zero	• Rank order • Quantitative difference	Measurements (e.g., temperature/Celsius) … -1 0 1 2 3 4…
	Ratio	• Linear • Equal intervals between numbers • Absolute zero	• Rank order • Quantitative difference • Ratios	Measurements (e.g., length, biomarkers, blood pressure) 0 1 2 3 4 5…

Qualitative numerical data represent different categories of a variable. There are two kinds of qualitative numerical data, referred to as nominal and ordinal. Ordinal level data are numerical codes without any quantitative meaning (i.e., there is no “more” or “less” implied), e.g., sex may be categorized as female or male and labeled by numerals. Ordinal data also are also numerical codes but in this instance an order from less to more is implied. This type of data is very common in clinical practice and research and is for example what we see with various rating scales (e.g., 0=none, 1=mild, 2=moderate, 3=severe). While ordinal data are ordered from less to more, it does not tell *how* much less or more but just that, e.g., 2 is more than 1 and less than 3. This is because the actual differences between ordered categories are unknown and, in general, unequal (e.g., the difference between 1 and 2 may be twice as much as that between 0 and 1).

Quantitative data, on the other hand, represent linear measurements. This means that the intervals (or differences) between two numerical values are the same across the range of possible values, which in turn means that calculations such as addition and subtraction are meaningful and that differences can be expressed. There are two kinds of quantitative numerical data, referred to as interval and ratio data. The difference between these two is that with interval level data there is no absolute zero. In practice interval data are not very common, and one of very few common examples is temperature measured according to Celsius or Fahrenheit (whereas temperature according to Kelvin is on a ratio level scale). Ratio level data, on the other hand are very common in clinical practice as well as in everyday life (e.g., length, weight, blood pressure, various biomarker concentrations in blood samples, and so forth). In practice, the difference between interval and ratio level data is marginal. Both allow for the same type of mathematical calculations, except that ratios and percentages cannot be validly expressed based on raw interval data. For example, if the outdoor temperature is +8 °C today and it was +2°C yesterday, we cannot say that it is 4 times warmer today (8/2=4), because the 0 is not an absolute zero. This is clear if one considers the same relationship but with 3 degrees lower temperatures, i.e., +5°C and -1°C (5/-1=-5). If, on the other hand, we consider two differences on an interval scale (say the difference between +5°C and -1°C, i.e., 6 °C, and that between +8 °C and +4°C, i.e., 4 °C), ratios and percentages may well be calculated and expressed in a meaningful way (in this example, 6/4=1.5, i.e., the first difference is 50% greater than the second).

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

The distinction between qualitative and quantitative numerical data has a long history. For example, in his 1920 textbook *Physics: The elements*, physicist Norman Campbell distinguished between numerals and numbers, where numbers represent quantitative data that allow for arithmetical calculations, whereas numerals represent qualitative data that do not (Campbell 1920). The implications are important, but not seldom forgotten.

In general, and from a clinical research perspective, quantitative data involve measures and measurement, whereas qualitative data are assessments and ratings. While this distinction is important in a practical sense, these are not inherently different or separate from each other at a more fundamental level. This is because all measurement begins with some sort of systematic observation and based on this an assessment of the amount observed (from less to more) is made. Going back to the taxonomy reviewed above, these observations and assessments are qualitative by nature, and represent all measurement, even physical measures such as length and weight that we tend to take for granted but at some point in history they were invented and devised through this type of process. Transformation of observations and assessments into quantification and measures of the amount requires a measurement process that, among other things, involves the definition and expression of a unit that is stable and linear. Examples of such units include the meter, the kilogram and the second. Indeed, while there are several definitions of measurement available, the international definition states that measurement is a process of obtaining quantity values (i.e., units) that can reasonably be attributed to a quantity (Joint Committee for Guides in Metrology 2012), which is in line with the definition of objective measurement provided by the Institute of Objective Measurement: “the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument is used and no matter who or what relevant person or thing is measured” (Program Committee of the Institute of Objective Measurement 2000). This clarifies that it is important to distinguish between numbers that represent assessments and those that represent measurement, since their meanings are fundamentally different.

Rating scales

Rating scales are commonly used in clinical practice as well as to assess outcomes in clinical research. The term rating scale is often used in a very broad sense to represent structured assessments based on observations or self-reports. The widespread use of rating scales in means that they play a central role in high stakes decisions regarding, e.g., the usefulness of therapies. However, the quality of clinical study results is directly dependent on the quality of the instruments used to collect data. Therefore, it is relevant to consider rating scales and their use in clinical practice and research (Cano & Hobart 2011).

The method or instrument used to assess or measure a particular variable needs to exhibit certain quality criteria for the results to be trustworthy and interpretable. This is often taken for granted. For example, when we measure the temperature of something we just do it and

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

use the result without further consideration of the quality assurance processes (e.g., calibration of the thermometer to a reference standard and the establishment of its measurement uncertainty) that ensures that we can do so. In the measurement of physical and chemical entities this type of quality control is built into the system of instrument development and is, in general, of no concern for end users such as HCPs. While quality criteria also apply to rating scales and other assessment methods and instruments, these cannot be taken for granted in the same way since quality control processes such as those mentioned above for physical measures are not available. In addition, rating scales are associated with some complexities since they typically are intended to target so called latent variables that are not directly observable (e.g., symptom severity, quality of life, support, disability).

Rating scales rely on observable manifestations of the variable they intend to target or represent. These manifestations are expressed as items (questions, statements, observed behaviour or performance, etc.) that make up the instrument, and variations in item responses are assumed to reflect variations in the variable that the rating scale intends to represent (Hagell 2019). Basically, the use of a rating scale involves the presentation of items to the individual (either for self-report or by observation), who is asked to affirm one of two or more ordered response categories that each describe a certain level. Each descriptive response category (e.g., none – mild – moderate – severe) is assigned an integral numeral (e.g., 0 – 1 – 2 – 3) as a means of partitioning the underlying latent continuum into various amounts. Item responses are then typically summed into a total score intended to locate the respondent on a continuum from less to more. These raw total scores are then used to describe the level of individuals or groups of people, make comparisons, and evaluate changes.

However, as discussed above, scores obtained from rating scales represent assessments that are qualitative and ordinal by nature rather than measures. Among other things, this means that a certain difference in raw total scores may represent diffident things depending on where along the possible rang of scores it is observed. For example, the difference between a score of 1 and 2 may not be any smaller than that between scores of 5 and 8, if those scores had been transformed into linear measurement. This is because raw rating scale total scores are ordinal, i.e., they are rank ordered but the intervals or distances between them are unequal (Hobart & Cano 2009). The general relationship between ordinal rating scale total scores and the linear measures that they represent is illustrated in Figure 2.

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

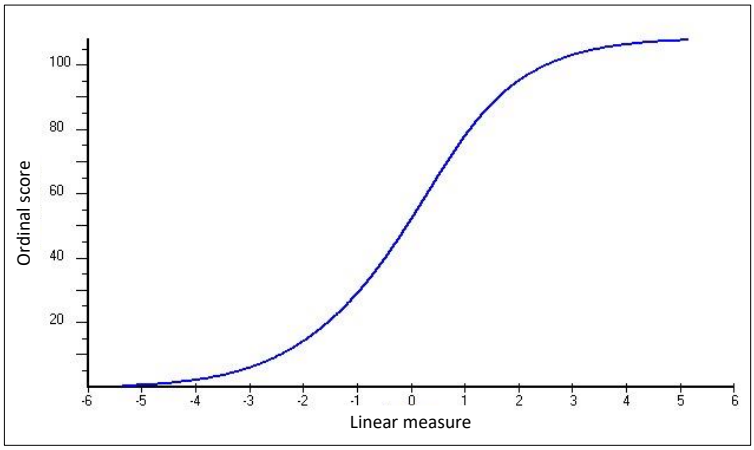

Fig. 2 The relationship between ordinal scores and linear measures

Transformation of ordinal raw rating scale total scores into linear measures is possible provided that the rating scale is of sufficient quality and meets the criteria measurement. The technique to achieve this is referred to as Rasch Measurement Theory (RMT) and is based on a measurement model developed by Georg Rasch (Rasch 1960). RMT has been applied in the health sciences for more than 30 years (Wright & Linacre 1989) and its use has increased considerably over the past couple of decades. However, its application has primarily been in testing the quality of rating scales to see whether they allow for linear measurement, whereas its use to obtain linear measurement has been less common. Regardless, if a rating scale has been found to work as intended and be of good quality when evaluated according to RMT, this is in general to be considered a good indicator that the rating scale is of good quality. Nevertheless, even if a scale meets the rigorous RMT critera, its raw total scores are still ordinal in nature. We will not consider any further details regarding RMT and its use but refer to available literature on the subject (see, e.g., sAndrich & Marais 2019; Cano & Hobart 2011; Hagell 2019; Hobart & Cano 2009; Hobart 2003; Tennant et al. 2004; Wright & Linacre 1989; Wright & Stone 1979; Wright & Masters 1982).

When assessing the quality of rating scales, there are certain criteria that typically are considered. Essentially, these are the same as those used for traditional physical measures (Pendrill 2019), but they differ in terms of the rigour by which they are tested, and the terminologies and scientific traditions differ. For example, the field of quality assurance of physical measures is referred to as metrology (Howarth & Redgrave 2008), whereas that of rating scales is referred to as psychometrics (Hobart & Cano 2009). Although RMT represents a substantial leap forward when it comes to quality assured rating scale based measurement, traditional methods (often referred to as Classical Test Theory, CTT) are still more commonly applied (Andrich & Marais 2019; Cano & Hobart 2011; Hagell 2019; Hobart & Cano 2009; Hobart 2003). However, regardless of the methods used, there are two main quality aspects of rating scales, typically referred to as psychometric properties: validity and reliability (Alrubaiy et al. 2014; Hobart & Cano 2009; Hobart et al. 1996; Hobart 2003; Streiner & Norman 2014; FDA 2009).

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

Whether rating scale scores confidently can be interpreted as good representations of the variable they intend to represent is a matter of validity. For a rating scale to be valid it needs to be based on a clear definition of the variable that it intends to target, and its items need to represent various aspects of that same variable. Consequently, and similarly to the case in physical measurement (Kuhn 1961) thorough theoretical and qualitative work is a typically a prerequisite for successful development of high-quality rating scales, and also an important aspect when judging the relevance and validity of a particular scale (Hobart 2003; FDA 2009). Reliability, on the other hand is not concerned with what scores represent but the precision of the obtained scores. As such, it is related to the concept of measurement uncertainty that is used in metrology (Howarth & Redgrave 2008). It can therefore be thought of as the extent to which observed scores are free from error. Consequently, the better the reliability, the easier it is to detect differences and changes. Reliability is expressed as a coefficient between 0 and 1 (where 1 represents perfect reliability and no error); a reliability coefficient around 0.80 or above is often considered acceptable. Sometimes the ability of scores to reflect changes (referred to as responsiveness) is considered a third psychometric property. The psychometric properties of rating scales are typically tested in studies conducted for this specific purpose (Hagell 2019; Hobart & Cano 2009; Hobart et al. 1996; Hobart 2003; Streiner & Norman 2014).

Analyzing numbers

Statistics, like metrology and psychometrics, is a scientific discipline of its own, and one of relatively long history (Stiegler 1986). What sets disciplines such as these apart from most (if not all) other disciplines is that they are applicable regardless of what subject area is being studied, whether it is health, agriculture, psychology, paleontology, or whatever. The basic purpose and function of statistics is to summarize, describe and analyze numerical data, typically regarding associations of some sort such as differences and correlations. The number of available statistical tests is vast and may seem infinite. However, for most situations a relatively limited number of tests and methods will suffice. Furthermore, the important principles to be aware of as a consumer of quantitative research are relatively few. Here we will not go into details concerning particular statistical tests, but review some of the common principles that are relevant to be aware of. For more detail, please refer to any of the many textbooks available (e.g., Norman & Streiner 2014, Swinscow & Campbell 1997 or Munro 2005).

Statistical analyses may be categorized as univariate, bivariate and multivariate. Univariate analyses concern the analysis of one variable at a time. In practice, this essentially involves descriptive statistics that is used to summarize data. Bivariate analyses concern the analysis of two variables at a time, e.g., the correlation between two variables or group comparisons where one variable defines the groups and the other is the one being compared (e.g., comparing ages between men and women). In multivariate analyses, more than two variables are considered. For example, determination of which of several variables that are associated

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

with an increased risk of something. A variable may be defined as characteristic that may vary, e.g., sex, age, blood pressure, anxiety, treatment, well-being. In other words, variables are the "things" that data are collected on. In the analyses of data, a distinction is often made between dependent and independent variables. The dependent variable can be thought of as the outcome or the effect, i.e., something that is influenced by (dependent on) differences or changes in other variables. Independent variables, on the other hand, are variables that may influence, explain, or cause the outcome or the effect. For example, in the comparison of ages between men and women, age is the dependent and sex is the independent variable. In a clinical trial of a therapy intended to improve sleep, the treatment group (e.g., receiving the active therapy or no therapy) is the independent variable and sleep is the dependent (outcome) variable. Finally, if studying which of the factors age, sex, and body mass index (BMI) that is most strongly associated with perceived well-being, the former three (age, sex and BMI) are independent and well-being is the dependent variable. With few exceptions, any variable may be either dependent or independent, since this is defined by the research question being asked. If, e.g., the last example above was re-framed as which of the factors age, sex, and perceived well-being that is most strongly associated with BMI, the dependent variable would be BMI instead of well-being (which would be an independent variable).

To complicate matters (?), the terms dependent and independent are also used to characterize the type of observations (data) in an analysis. Dependent observations means are observations are somehow linked and have a systematic relationship. The most common type of dependent observations is when data are collected from the same people at different time points, e.g., before and after an intervention. Independent observations, on the other hand, do not have this relationship. The perhaps easiest way to distinguish between dependent and independent observations is to consider whether data come from the same people or not. If they do, observations are dependent, and if they do not, they are independent. Sometimes dependent and independent observations are also referred to as paired and unpaired.

Parametric and nonparametric statistics

In addition to what type of analysis (descriptive, comparison, correlation, etc.), choosing an appropriate statistical test involves consideration of characteristics of the data and the situation (Greenhalgh 1997a). By situation, we mean whether observations are dependent or independent, and whether there are two or more than two groups involved. The characteristics of data that need to be considered are whether they are nominal, ordinal or interval/ratio, and their distribution (typically whether they are approximately normally, or symmetrically, distributed). This determines which of two main types of statistical tests that is appropriate, parametric or nonparametric. Parametric tests assume that data are quantitative (i.e., interval or ratio) and come from a particular distribution (typically the normal distribution), whereas nonparametric tests make no such assumptions but are based on rank order, frequencies or proportions. To illustrate, we will consider parametric and nonparametric descriptive statistics.

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

The main purpose of descriptive statistics is to summarize a large set of observations in a way that provides an efficient and representative picture of the data. To do so, there are two main types of descriptive statistics, central tendency and dispersion (McCluskey & Lalkhen 2007). Central tendency expresses the "average" value for a variable and depending on the characteristics of the data, either the mode, the median or the mean may be the best choice. Central tendency should be coupled by a measure of dispersion, which describes the spread of data around the central tendency. Common measures of dispersion are min-max, range, quartiles, and the standard deviation (SD). In addition, there is also the variance. Whereas variance is commonly used in the calculation of various statistical tests, it is rarely (if ever) used to describe data. Of the descriptive statistics mentioned above, the mean and the SD (as well as the variance) are parametric, and the others are nonparametric.

The descriptive statistics mentioned above are illustrated in Figure 3 using a data set with the ages (in years) of a small sample of 12 people (n=12). For ease of illustration the ages have been arranged from the oldest to the youngest. Starting with the parametric options, the mean is calculated as the sum of all values (in this case 474) divided by the number of observations (i.e., 12), which gives a mean age of 39.5. The measure of dispersion to use with the mean is the SD, which can be interpreted as the average spread below and above the mean. In this case the SD is 13.8, so on average the ages are spread 13.8 years below and above 39.5.

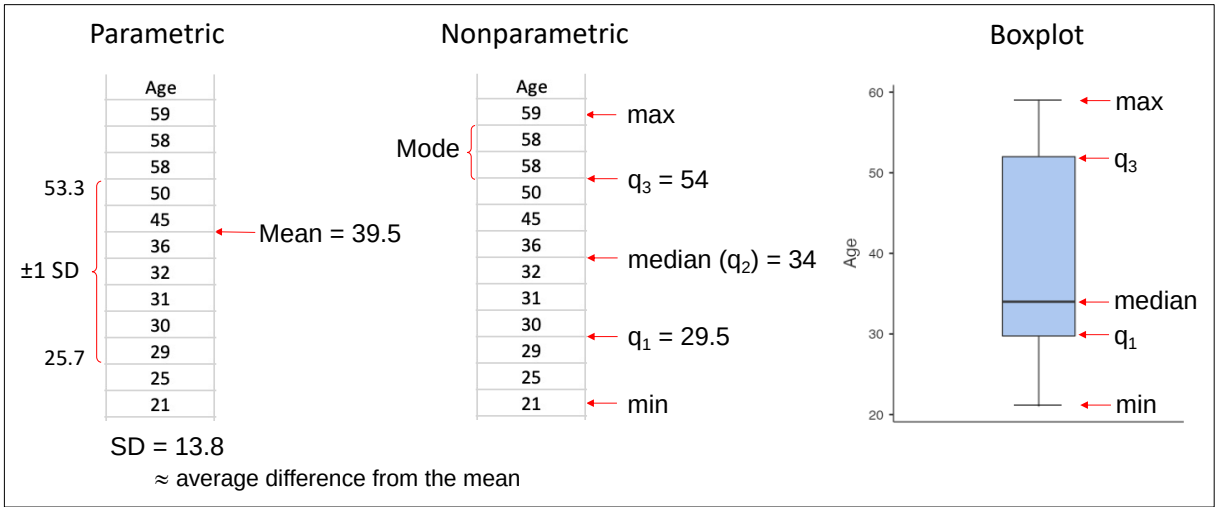

Fig. 3 Parametric and nonparametric descriptive statistics

There are two nonparametric central tendencies, the mode and the median. The mode is simply the most frequently occurring value, in this case 58. As can be seen (Fig. 3) this is not a particularly representative value, and it is therefore rarely used to describe the central tendency of quantitative data. It may be used with ordinal data but its most common use is with nominal data. The median is the value in the middle when data have been rank ordered. When the number of observations is even (as in Fig.3) the median is the mean of the two data points in the middle (in this case 34 is the mean of 32 and 36). The dispersion around the median is typically indicated by the first and third quartile values (q1 and q3, respectively).

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

These are based on the same logic as the median, which represents the second quartile (q2). Thus, the three quartiles are the values that split the observations into four equally sized groups. Therefore, q1 and q3 (which are used to indicate the dispersion around the median) define the range within which the mid half of all observations are. This is sometimes referred to as the inter-quartile range (IQR), although per definition the IQR is simply the difference between q3 and q1 (in our case 24.5, i.e., 54 – 29.5). Sometimes the three quartiles are referred to as the 25th (q1), 50th (q2) and 75th (q3) percentiles. Another way of expressing the dispersion is to give lowest and highest observed values, i.e., min-max. As with quartiles, this is often called the range, although per definition the range is simply the difference between the highest and the lowest values (in our case 38, i.e., 59 – 21). On its own, min-max is not very useful since it per definition is very influenced by extreme outliers and therefore may not provide very representative information. However, it may be useful to give the min-max in addition to, e.g., the SD or q1-q3, to provide a more complete description of data. To the far right in Figure 3 is a graphical representation of what just has been discussed. This type of graph is called a boxplot or box-and-whisker plot. The lower and upper limits of the box in the middle of the graph represent q1 and q3, respectively. The horizontal line inside the box is the median (q2) and the whiskers represent the lowest and highest observed values. Sometimes asterisks are used to indicate more extreme outliers beyond the whiskers.

Median and quartiles are used when data are either ordinal or when interval or ratio level data are not normally distributed. This is because these nonparametric statistics do not assume quantitative data or a particular distribution of data. To see why this is the case, we will review the characteristics and logic of the normal distribution (McCluskey & Lalkhen 2007). As illustrated in Figure 4 (panels a and c), the important feature of the normal distribution is that it is symmetrical, i.e., the frequency (y-axis) of observed values (x-axis) steadily increases and peaks in the middle of the range of values, before it decreases at the same rate as it increased. This curve is sometimes referred to as a bell-shaped curve since it resembles an old church bell.

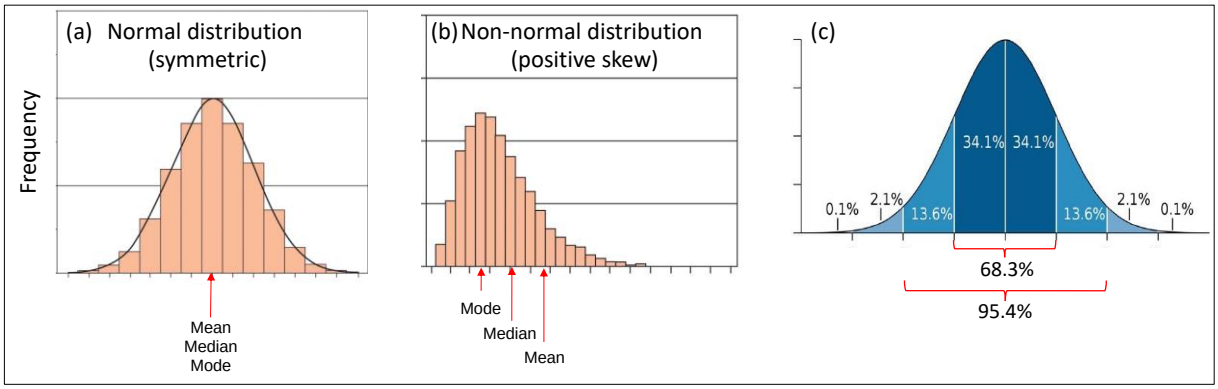

Fig. 4 The normal distribution (panels a and c) and a non-normal distribution with positive skew (panel b). A non-normal distribution with negative skew would show the same pattern but with the tail pointing to the left instead of to the right.

As illustrated in panel a, the peak of the normal distribution coincides with all three measures of central tendency. The value on the x-axis that represents the peak is the most common

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

value (i.e., the mode), divides the observations in two equally sized groups (i.e., the median) and it is also the mean of all observations. This is all due to the symmetry of the distribution. Panel c in Figure 4 also illustrates the meaning of the SD. If we go 1 SD above and below the mean, that range will cover about 68% of all observations (34.1% on either side of the mean) and if we take 2 SDs above and below the mean, that range will cover about 95% of all observations (47.7% on either side of the mean). However, when the distribution is not symmetrical these features fall apart, as illustrated in panel b of Figure 4. First, because the mean is sensitive to extreme values it will be influenced by the skewness of the distribution, which makes it less representative of the data. The median on the other hand is insensitive to the distribution and will still represent the value that divides the observations into two equally sized groups. Therefore, the median is the recommended central tendency measure when data do not conform with the normal distribution. Furthermore, since non-normal distributions are asymmetrical, the meaning of the SD falls apart. With severely skewed data, the SD may for example suggest the average difference from the mean is a value that cannot exist. Again, this will make the SD unrepresentative of the data and essentially uninterpretable, whereas quartiles will remain appropriate.

The above examination of parametric and nonparametric descriptive statistics is, in general terms representative also of other parametric and nonparametric statistics used for, e.g., group comparisons, correlations, and so forth. Therefore, we will not cover these in any detail but conclude that they operate according to the same principles as those reviewed above. Researchers often prefer parametric methods because they tend to offer some more flexibility and provide more useful information. However, with few exceptions, there are nonparametric tests available as an alternative to their parametric counterparts so there is no real need for inappropriate use of parametric statistics, particularly not given the risk of getting spurious and unrepresentative results because assumptions are not met. However, judging whether a correct choice has been made may be a difficult task since it may not be obvious whether data reported in a clinical study are compatible with the normal distribution or not. But there are some signs that may help the reader. First, if the authors for whatever reason present both the mean and the median this can provide a clue because if these two values differ considerably, this is a clear sign that data do not follow a normal distribution. Second, if the value of the SD is large relative to the mean (say larger than half the mean) this is also a sign that data probably is not normally distributed (recall the meaning of the SD as illustrated in panel c of Figure 4). Finally, and regardless of the distribution, parametric statistics such as the mean and SD are not appropriate for qualitative (i.e., nominal and ordinal) data (Greenhalgh 1997a; Svensson 2001).

Table 2 lists various common parametric and nonparametric statistical tests and methods according to the type of analysis they represent as well as to the characteristics (level and distribution) of data and situation.

Table 2. Statistical methods and their use according to data levels, distribution, and situation [a]

Data level / Distribution	Description	Correlation	Group comparison			
			Independent observations		Dependent observations	
			2 groups	>2 groups	2 time points	>2 time points

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

• Nominal	• Frequencies, % • Mode	• Phi coefficient • Cramér's *V*	• Chi^2 test • Fisher's exact test		• McNemar's test	
• Ordinal • Interval / non-normal • Ratio / non-normal	• Median (md) • Quartiles (q1-q3) • Min – Max • Range	• Spearman's rho (r_s) • Kendall's tau • Kendall's W	• Mann-Whitney U-test	• Kruskal-Wallis' H-test	• Wilcoxon's signed-rank test	• Friedman's test • Skillings-Mack test
• Interval / normal • Ratio / normal	• Mean • Standard deviation (SD)	• Pearson's *r* • Intra-class correlation (ICC)	• Independent samples t-test • Analysis of covariance (ANCOVA)	• Analysis of variance (ANOVA) • ANCOVA	• Paired t-test • ANCOVA	• Repeated-measures ANOVA • ANCOVA

[a] The shaded area represents parametric statistics.

Correlation and regression

Together with group comparisons analyses regarding the association between variables are among the most common in the health sciences. Although group comparisons also address associations (i.e., between groups of people and the dependent variable), the more specific types of analyses to study associations are correlation and regression. While correlation and regression have similarities, they also differ from one another. Correlation concerns the linear relationship between two variables. This may be addressed graphically by means of scatterplots, where each person's values on two variables (let's call them X and Y) is plotted and the shape of the plots is inspected regarding how close the plots are to an imagined straight line and whether the association tends to be positive (values on Y tend to increase as values on X increase) or negative (Figure 5).

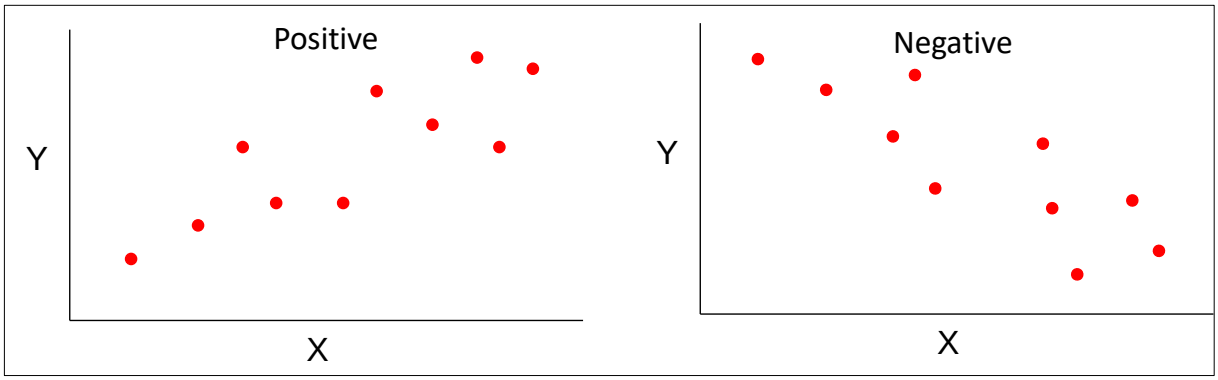

Fig. 5 Hypothetical scatterplot of positive and negative linear relationships between variables X and Y.

Correlation is a means to quantify the kind of relationships illustrated in Figure 5. This quantification is made by a correlation coefficient, which may take values between -1 and 1 and expresses the extent to which the dots follow a straight line. Negative values represent a negative relationship and positive values represent a positive relationship; the closer to -1 or 1 the coefficient is, the stronger the linear relationship. It is important to keep in mind that the coefficient expresses linearity (i.e., the proximity of dots to an imagined straight line) and nothing else. This means that certain patterns of associations, such as U-shaped relationships, are not reflected by the correlation coefficient. It is therefore wise not to rely solely on correlation but also to plot data since this can reveal important patterns not reflected by the coefficient. The most common correlation coefficient is Pearson's product-moment correlation coefficient (r). However, this is a parametric method that, among other things

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

assumes that both variables are quantitative (interval/ratio) and normally distributed. When this is not the case nonparametric correlation, such as Spearman's rank correlation (rho) should be used instead.

As touched upon earlier correlation does not tell whether the association is causal. With regression we are in a somewhat better position to address causality, although the extent to which causal inferences can be made also relies on design related and other aspects (Greenhalgh 1997b; Guyatt et al. 1995). The reason is that regression does not only concern the relationship between variables but allows the variation in one variable (the dependent variable) to be explained or predicted from another (the independent variable). There are many types of regression analyses available and the choice large depends on the type of dependent variable that one wants to predict the value of. Examples include logistic regression (the dependent variable is dichotomous, e.g., ill or healthy), ordinal regression (the dependent variable is ordinal), multinominal regression (the dependent variable is nominal with more than two categories), and poisson regression (the dependent variable is a count of events). However, the basic logic is the same regardless of the type of regression and will be illustrated here by means of linear regression, i.e., when the dependent variable is quantitative. This is illustrated in Figure 6.

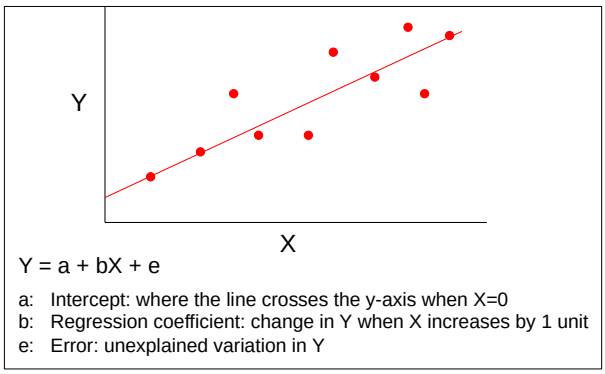

Fig. 6 Linear regression to predict the dependent variable Y from the independent variable X.

Regression identifies the straight line that follows the observations (dots) as closely as possible, which is expressed as an equation that tells the estimated value of the dependent variable Y based on the independent variable X: Y = a + bX. In this equation, a represents the intercept (i.e., the value where the straight line crosses the y-axis, in other words the value of Y when the value of X is 0), and b is the regression coefficient (or the slope). Since b is multiplied by X, this means that b represents the change in Y when X increases by one unit. Since the independent variable never explains all variation in Y, there is also an error term (e) in the equation. In addition to the information mentioned above, a regression analysis will also provide information on the proportion of variation in Y that can be explained by X. This value is called the R^2 value.

The powerful thing with regression lies in the fact that we are not limited to consider just one independent variable but are able to simultaneously study the influence of several variables

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

on the variation in Y. This is called multiple regression and represents one of the most common types of multivariate statistical methods. In multiple regression, we will get the same type of information as that reviewed above, but for each one of the independent (explanatory) variables while also taking account of the influence of the other independent variables in the analysis. For example, if we are interested in explaining the variation in body weight from physical activity, we may also include, e.g., age and sex in our analysis since there is reason to believe that these also are associated with weight. The results will then tell us to what extent we may explain variations in weight from physical activity, while at the same time controlling for the influence of age and sex.

Other common multivariate statistical methods in the health sciences include (but are not limited to) factor analysis (identification of variables that share something in common), path analysis (essentially a variation of regression where it is possible to account also for indirect associations), and analysis of covariance (ANCOVA; group comparison that controls for or takes account for additional variables; e.g., comparing the effect of an intervention between study groups while taking account for differences between groups at the beginning of the study).

Hypothesis testing and the P-value

Most statistical analyses are concerned with hypothesis testing. For example, do groups differ? Is there an association? Is there an increased risk? And so on. Formally, what is being tested is something called the null hypothesis, typically denoted H_0. If we consider an analyses of a potential group difference, the null hypothesis says that there is no effect, i.e., no difference between the groups. Theoretically, the null hypothesis is considered to hold unless there is strong enough evidence against it. In that case, the null hypothesis is rejected in favor of the alternative hypothesis (called H_1). The alternative hypothesis is the opposite of the null hypothesis and means that there is an effect, in this case a difference between groups. This logic does not only apply to group comparisons but to virtually all kinds of statistical analyses. For example, in correlation the null hypothesis says that there is no association between the variables, in regression the null hypothesis says that Y cannot be explained by X, in risk assessment the null hypothesis says that there is no increased risk, and so on. The general term used regardless of situation is “effect”, which thus can refer to differences, associations, explanatory power, risk, or whatever is being studied.

So, how do we determine when there are strong enough evidence against null hypothesis? This is done by means of assessing the probability that the observed (or a larger) effect had occurred if the null hypothesis is true. That probability is referred to as the P-value. Like all probabilities, the P-value can in principle range between 0 and 1, although in practice it is never 0 or 1 since there always is a slight chance that it is wrong. The general rule of thumb since about 100 years is that the null hypothesis is rejected if the P-value is <0.05, i.e., if there is less than 5% probability to see an effect (e.g., difference) of the size observed, or larger, if the null hypothesis is true. When P-values are <0.05, this is typically referred to as a

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

statistically significant result (Greenhalgh 1997b). In high stakes analyses such as interventions where the outcome concerns whether people will survive or not, this threshold is usually set at a lower level (e.g., P<0.001) to be as certain as possible. Also, in studies where many hypothesis tests are conducted it is common (and advisable) to adjust the resulting P-values downward since the influence of chance increases for each test that is conducted. One common such adjustment is something called the Bonferroni correction (Bland & Altman 1995; Greenhalgh 1997b). However, it is recommended that exact P-values (not only whether above or below, e.g., 0.05) should be reported and their interpretation should not be a binary one. For example, the practical difference between P-values of 0.052 and 0.049 is negligible (Whitley & Ball 2002a; Cohen 1994).

It is of utmost importance to be aware that the P-value, no matter how small, only concerns statistical significance. In other words, it says *nothing* about the size of the observed effect or about whether it is of any clinical significance or not (Matthews & Altman 1996; Altman & Bland 1995). This is because the P-value depends not only on the effect (e.g., the size of a difference) but also on the sample size as well as on the dispersion. As the sample size increases and the dispersion decreases, the P-value will decrease. For example, with 9 people in each of two groups with an SD of 15, a mean difference of about 10 will be significant (P<0.05), whereas the same level of statistical significance will be observed with a mean difference of about 6 if there are 25 people in each group and with a mean difference of about 1 if there are 900 people in each group (Norman & Streiner 2014). In other words, very large samples are more or less bound to yield statistically significant results, even though they may be trivial in practice. To get a better sense of the meaning of results, it is common to not only present P-values and associated descriptive statistics, but also something called confidence intervals. A confidence interval tells us within what range the true effect is likely to be. The most common is to report 95% confidence intervals, which represent the range of values within which we can be 95% confident that the true effect (e.g., group difference) is (Greenhalgh 1997b; Sheldon 2000). If this range includes or is close to what can be considered a trivial or no effect, there is probably good reason to question the practical relevance of the findings. It should be noted that just like P-values, confidence intervals are also influenced by sample size and dispersion.

How many people?

The issue of sample size has been touched upon earlier in this report and will now be explored in slightly more detail. As touched upon in the previous section, an important aspect in planning for a study that involves group comparisons (e.g., a study to assess the effectiveness of an intervention) is to make sure that the study groups are sufficiently large to detect such an effect statistically. On the other hand, if the groups are unnecessarily large the analyses may indicate statistically significant effects that in fact are trivial. It is thus desirable to have large enough, but not too large samples. Apart from technical statistical reasons, another important reason to have a well-balanced sample size is an ethical one. It would be of questionable ethical legitimacy to conduct a study with a sample so small that it is highly

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

improbable to detect an effect even if it is there. And on the other hand, it would also be questionable if it is good ethical practice to expose many more people than necessary to the extra burden that it means to participate in a clinical study.

The way that researchers deal with the issue of obtaining well-balanced sample sizes is by doing a power analysis, which is used to determine how many participants that is needed to detect a certain effect (Whitley & Ball 2002b; Chan 2003). Power refers to the likelihood of statistically detecting an effect. Investigators typically strive for at least 80% and often 90% power. While the exact procedure differs depending on the planned type of analysis, the basic ingredients needed are the size of the effect of interest (e.g., the difference that is clinically relevant), at what P-value the effect is to be detected (e.g., <0.05), the expected dispersion of the outcome measure (e.g., SD), and the desired power (e.g., 80%). Once these values have been set, one can calculate the number of people needed in each group. The tricky parts of the procedure typically are to decide the minimum effect or difference that is of clinical importance and the expected dispersion. With some well-established outcome variables, such as blood pressure, it may be relatively easy to decide the minimum effect size, whereas it may be more difficult with rating scale based outcome measures. For some such scales, studies where a minimally important difference in scores have been estimated may be available but often it will be a matter of clinical experience-based judgement. The expected dispersion of the outcome measure is obviously also difficult to know before any data have been collected, and typically this is based on educated guesses from previous studies. However, whatever considerations and assumptions that were made when conducting the power analysis should be summarized when the results from the study eventually are reported. In general, there is a plateau effect in power analyses. That is, a lot of power is gained when increasing a relatively small sample size, whereas less is gained once a certain number of people are in the study. This is illustrated in Figure 7.

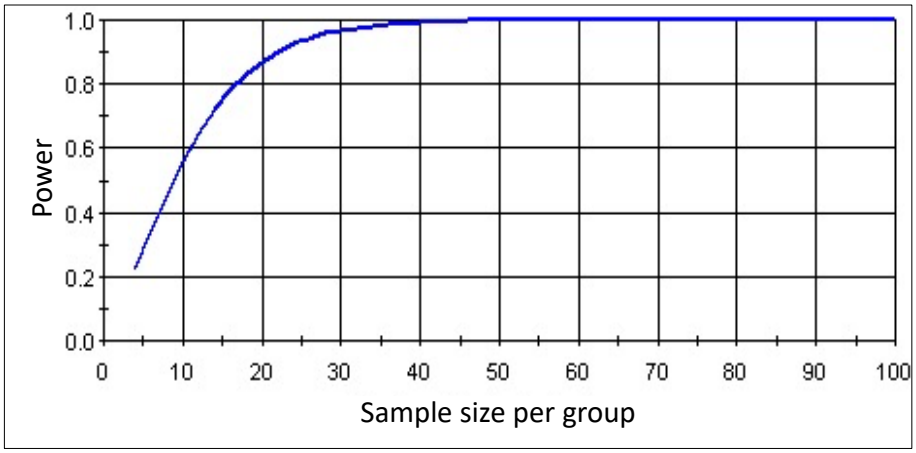

Fig. 7 Example plot illustrating the relationship between power and the required sample size per group to detect a mean difference of 10 at P<0.05 with an SD of 10 in each group.

Once a power analysis has been conducted an estimation of the expected drop-out rate needs to be made since it typically cannot be expected that all recruited participants will complete

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

the study. If, for example, it has been estimated that 25 people are needed in each group to have a 90% power (Figure 7) and it turns out that 10 people for whatever reason do not complete the study, the actual power at the end will be about 70%. To compensate for this, most studies recruit more participants than estimated in the power analysis.

Estimating sample size requirements by means of power analysis is primarily done in interventional and other studies primarily concerned with group comparisons. In more explorative studies, using for example regression or other multivariate statistical methods, the main issue is typically not so much concerned with having enough participants to detect a certain effect, but rather to have sample that is large enough to obtain interpretable and stable results. While this requirement is complex and differ between methods, as a very general rule of thumb statisticians tend to recommend a minimum of 50-100 + at least 10 or so per variable or estimated parameter (Norman & Streiner 2014). For example, if we plan to do multiple regression analyses with 10 independent variables, we will need to have complete data from at least 50-100 + 100 (10 variables x 10 observations), i.e., 150-200 people.

Putting it all together

The principles reviewed above come with a number of caveats for the practicing HCP. For example, it is not sufficient for a study to be designed as an RCT, but its internal and external validity must also be taken into consideration. This includes, but is not limited to questions such as

- Was the study sample representative for patients encountered in daily clinical practice?
- Is the primary study outcome variable clinically relevant?
- Was the primary study outcome assessed or measured using a reliable and valid tool?
- Were the methods(s) used to analyze the data appropriate to meet study objectives and for the data available in the study?
- Are study results clinically meaningful and interpretable?

While issues such as these should have been addressed during the peer review process that foregoes the publication of research papers, this may not necessarily be the case and even if it is, there may be issues remaining when judging a study in view of one's own clinical practice and context. Therefore, it is important for HCPs and other end-users of clinical studies to be aware of these challenges. This does not mean that each and everyone needs to be an expert in research methodology; it often suffices to be aware of the issues and apply this awareness with common sense and experience. In the following section we will exemplify the issues mentioned above by reviewing a real word example.

A real world example

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

To illustrate the points outlined above, key aspects of a typical double blind, placebo controlled RCT will be reviewed. The RCT involved people with Parkinson's disease (PD)[1] who experience wearing-off, i.e., episodes with insufficient efficacy of levodopa and increased PD symptoms before the onset of the next dose. The RCT ran for 13 weeks and concerned a drug to be used together with levodopa to reduce wearing-off problems. The aim of the study was to determine the effect of the combination therapy on activities in daily living (ADL), which thus was the primary study outcome.[2] The primary outcome measure in the study was part II of the Unified PD Rating Scale (UPDRS; Fahn et al. 1987), which is a gold standard clinical assessment scale in PD and consists of four main parts: Mentation, behavior and mood (part I), ADL (part II), Motor symptoms (part III), and Complications of therapy (part IV).

The study and its main design features and results is illustrated in Figure 8. Based on a power calculation to detect a difference between treatments with a power of 90% at $P<0.05$ based on results from previous studies using part II of the UPDRS (UPDRS II) the study recruited 270 participants, of whom 174 were randomized to receive the active study drug and 96 were randomized to receive inactive placebo pills (the control group). The sample was assessed at baseline and 13 weeks later and the change within each group was compared at the conclusion of the study according to the intention-to-treat principle.[3] Data were analyzed using analysis of covariance (ANCOVA), controlling for baseline data, group (active vs. placebo), study centre and group x study centre interaction. The difference in UPDRS II scores between study groups was significant ($P<0.0001$) in favor of the active study drug and it was concluded that the adjunct therapy ha good efficacy in terms of ADL, which thus is the conclusion to be generalized to the larger population of people with PD outside of the study.

[1] PD is a common progressive neurodegenerative disorder associated with motor symptoms of bradykinesia (slowness of movement), muscle rigidity, tremor, and postural impairments, as well as non-motor features such as depression, anxiety, sleep disorders, fatigue, cognitive impairment, dysautonomia and pain. The core motor symptoms are believed to be due to loss of dopamine in the brain and symptomatic drug treatment aims to restore brain dopamine levels, where levodopa (a precursor of dopamine) is the main compound.
[2] Reference to the study or the name of the tested drug is not provided since this is irrelevant for the purpose of this illustration.
[3] A strategy for analyzing data in interventional studies, where participants are analyzed according to the group they were assigned at the beginning of the study, regardless of whether they continued that treatment or not.

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

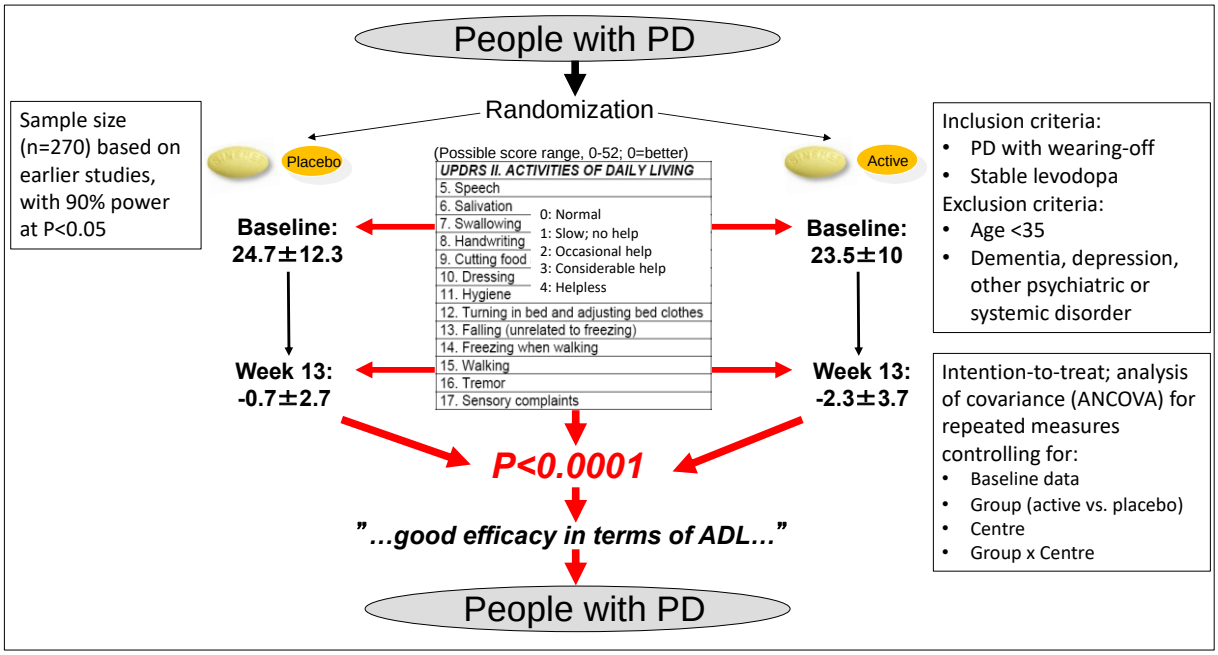

Fig. 8 A post-marketing (Phase IV) Northern European double-blind placebo controlled RCT of an adjunct (levodopa add-on) drug for Parkinson's disease (PD) regarding its effects on activities of daily living.

Was the study sample representative for patients encountered in daily clinical practice?
The study was conducted in Northern Europe. When identifying participants for the study, those who were under the age of 35, suffered from dementia, depression or other psychiatric, systemic or metabolic disorders were excluded. This means that the study population (i.e., the kind of persons that results are intended to be generalized to) does not include people with, e.g., mental comorbidities, which are common in PD. While not reported, it may further be assumed that a majority of participants in the study were Caucasians of Northern European heritage. An implication of this is that it is uncertain to what extent results are representative of, and applicable to, people from other geographical and cultural areas. This may be of relevance with respect to, e.g., how the primary outcome (ADL) was assessed in the study (i.e., the relevance and meaning of various ADL tasks may differ across cultural settings), as well as from a biomedical perspective since drug pharmacology may differ by ethnicity (Johnson 1997; Shas 2015).

Is the primary study outcome variable clinically relevant?
Arguably, it appears reasonable to consider activity performance in daily life (ADL) to be a clinically relevant outcome in studying the value of an intervention for people living with a long-term progressive disorder such as PD. Indeed, it can be argued that this outcome is more relevant from a clinical and person-centered perspective than, e.g., the effects on motor symptoms, since reduced symptoms *may* ease ADL performance but the latter is also influenced by other factors (World Health Organization 2001).

Was the primary study outcome assessed or measured using a reliable and valid tool?
The primary outcome variable was ADL and this was assessed using UPDRS II, which purports to represent ADL. It consists of 13 items that are assessed by HCPs based on information from

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

a clinical patient interview. Each item is scored from 0 (“normal” or absence of problems) to 4 (“severe” or inability to perform task), and item scores are summed to provide a total score that is taken to represent the person’s level of ADL limitation, from 0 (no limitations) to 52 (severe limitations). While the UPDRS II is widely used and has been deemed “recommended” for assessing disability in PD (Shulman 2016) it has also been criticized. While the reliability of the UPDRS II generally has been found acceptable, its validity has been questioned repeatedly (Forjaz & Martinez-Martin 2006; Hagell 2019; Hariz et al. 2003; Ramaker et al. 2002; van Hilten et al. 1994). The main issue is that only six of the 13 items of the scale represent activity performance, while the remaining items concern impairments. This means that it is unclear what scores on the UPDRS II actually represent.

Furthermore, when reading the full publication of the RCT considered here, it turns out that whereas the UPDRS II was the primary outcome measure, the investigators also used two additional rating scales to assess ADL performance. However, in contrast to the primary outcome measure, the UPDRS II, neither one of these showed any significant differences between the study groups. To

Were the methods(s) used to analyze the data appropriate to meet study objectives and for the data available in the study?
Data were analyzed according to the intention-to-treat principle, which is generally recommended since it better resembles clinical practice and potentially avoids the risk of claiming a treatment effect that actually does not exist due to, e.g., adverse effects that make people stop treatment.

A power analysis was conducted at the outset of the study to determine the required sample size to detect a difference in outcome with a power of 90% at $P<0.05$. Since the study result was highly statistically significant with a very small P-value (<0.0001), it is clear that the study was not under powered. However, it is less clear if it actually involved more people than necessary. This is because of the very small P-value in combination with the fact that we do not know what size of an effect (i.e., how large a difference between the groups) that was used as a clinically relevant difference in the power analysis (just that it was based on previous studies using the UPDRS II). We will return to this point shortly.

The outcome variable was assessed using the UPDRS II, which is a rating scale and thus represents qualitative (ordinal) data. However, data were presented using mean and SD and analyzed using ANCOVA, which all are parametric statistics that assume quantitative data. While the extent to which this may have influenced (biased) the results is essentially unknown. However, since rating scale data are not suitable parametric statistics (Greenhalgh 1997a; Svensson 2001) this potentially challenges the validity of results.

Are study results clinically meaningful and interpretable?
At the end of the study a mean difference of 1.6 on the UPDRS II (which represents about 3% of the possible 0-52 score range) was observed. One clinically pertinent question is therefore what this actually means, and whether it represents a clinically meaningful difference.

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

Previous studies addressing the minimal important difference in UPDRS II scores have suggested that improvements of anything between about 2 and 5 may be meaningful, depending on stage of disease (Hagell 2019). The observed difference of 1.6 is below this range and it is unknown how representative this reported mean difference is, since the median may have been more appropriate. Furthermore, it is unknown what this difference means in practical terms. That is, how people with this improvement actually have benefitted. The fact that the UPDRS II represents both activity performance and impairments further complicate this matter.

Finally, it is important to be aware that average scores and results from clinical studies that ae based on group comparisons say nothing about the outcomes of individual persons. This is because they are group data. In principle, failure to demonstrate an effect may be because some participants improved and some deteriorated, resulting in no average change or difference, when in fact a subgroup of people may experience clear benefits. This is a general concern that is difficult to circumvent without carefully conducted subgroup analyses or by taking measurement error at the individual person level into account (Hobart et al. 2010).

Co-funded by the
Erasmus+ Programme
of the European Union

Reference number: 2020-1-SE01-KA202-077806

References

Alrubaiy L, Hutchings HA, Williams JG. Assessing patient reported outcome measures: A practical guide for gastroenterologists. *United European Gastroenterology Journal* 2014; 2(6): 463–470. https://doi.org/10.1177/2050640614558345

Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995; 311: 485. https://doi.org/10.1136/bmj.311.7003.485

Andrich D, Marais I. *A course in Rasch measurement theory: Measuring in the educational, social and health sciences*. Singapore: Springer Nature Singapore Pte Ltd, 2019.

Bland JM, Altman DG. Multiple significance tests: the Bonferroni method; *BMJ* 1995; 310(6973): 170. https://doi.org/10.1136/bmj.310.6973.170

Campbell NR. *Physics: The elements*. London: Cambridge University Press, 1920. https://archive.org/details/physicstheelemen029733mbp/page/n11/mode/2up

Cano SJ, Hobart JC. The problem with health measurement. *Patient Preference and Adherence* 2011; 5: 279-290. https://doi.org/10.2147/PPA.S14399

Chan YH. Randomised controlled trials (RCTs) - Sample size: The magic number? S*ingapore Medical Journal* 2003; 44(4):172-174. www.sma.org.sg/smj/4404/4404bs1.pdf

Cohen J. The earth is round (p < .05). *American Psychologist* 1994; 49(12): 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Elfil M, Negida A. Sampling methods in clinical research; An educational review. *Emergency* 2017; 5(1): e52. https://doi.org/10.22037/emergency.v5i1.15215

Fahn S, Elton RL, members of the UPDRS development committee. Unified Parkinson's Disease Rating Scale. In: Fahn S, Marsden CD, Calne DB, Goldstein M (Eds.) Recent Developments in Parkinson's Disease, Vol. 2. Florham Park: MacMillan Healthcare Information, 1987, pp. 153-163.

FDA. *Patient-reported outcome measures: use in medicinal product development to support labelling claims*. Washington DC: Food and Drug Administration, 2009. www.fda.gov/media/77832/download

Forjaz MJ, Martinez-Martin P. Metric attributes of the Unified Parkinson's Disease Rating Scale 3.0 battery: Part II, construct and content validity. *Movement Disorders* 2006; 21(11): 1892–1898. https://doi.org/10.1002/mds.21071

Co-funded by the Erasmus+ Programme of the European Union

Reference number: 2020-1-SE01-KA202-077806

Greenhalgh T. Statistics for the non-statistician. I: Different types of data need different statistical tests. *BMJ* 1997a; 315(7104): 364-366. www.ncbi.nlm.nih.gov/pmc/articles/PMC2127256/

Greenhalgh T. Statistics for the non-statistician. II: "Significant" relations and their pitfalls. *BMJ* 1997b; 315(7105): 422-425. www.ncbi.nlm.nih.gov/pmc/articles/PMC2127270/

Grimes DA, Schulz KF. An overview of clinical research: The lay of the land. *Lancet* 2002; 359(9300): 57-61. https://doi.org/10.1016/S0140-6736(02)07283-5

Guyatt G, Walter S, Shannon H, Cook D, Jaeschke R, Heddle N. Basic statistics for clinicians: 4. Correlation and regression. *Canadian Medical Association Journal* 1995; 152(4): 497–504. www.ncbi.nlm.nih.gov/pmc/articles/PMC1337703/

Hagell P. Measuring activities of daily living in Parkinson's disease: On a road to nowhere and back again? *Measurement* 2019; 132: 109-124. https://doi.org/10.1016/j.measurement.2018.09.050

Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. *International Journal of Nursing Studies* 2009; 46(3): 380-393. https://doi.org/10.1016/j.ijnurstu.2008.10.007

Haidich AB. Meta-analysis in medical research. *Hippokratia* 2010; 14(Suppl 1): 29–37. www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/

Hariz GM, Lindberg M, Hariz MI, Bergenheim AT. Does the ADL part of the Unified Parkinson's Disease Rating Scale measure ADL? An evaluation in patients after pallidotomy and thalamic deep brain stimulation. *Movement Disorders* 2003; 18(4): 373–381. https://doi.org/10.1002/mds.10386

Hobart J. Rating scales for neurologists. *Journal of Neurology Neurosurgery & Psychiatry* 2003; 74 (Suppl 4): iv22-iv26. https://doi.org/jnnp.74.suppl_4.iv22

Hobart JC, Lamping DL, Thompson AJ. Evaluating neurological outcome measures: the bare essentials. *Journal of Neurology Neurosurgery & Psychiatry* 1996; 60(2): 127-130. https://doi.org/10.1136/jnnp.60.2.127

Hobart J, Cano S. Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. *Health Technology Assessment* 2009; 13(12). https://doi.org/10.3310/hta13120

Hobart JC, Cano SJ, Warner TT, Thompson AJ. Effect sizes can be misleading: is it time to change the way we measure change? *Journal of Neurology Neurosurgery and Psychiatry* 2010; 81: 1044-1048. http://dx.doi.org/10.1136/jnnp.2009.201392

Co-funded by the Erasmus+ Programme of the European Union

Reference number: 2020-1-SE01-KA202-077806

Howarth P, Redgrave F. *Metrology – In short, 3rd edition*. Euramet: Albertslund, 2008. www.dfm.dk/wp-content/uploads/2020/12/Metrology_in_short_3rd_ed.pdf

Impellizzeri FM, Bizzini M. Systematic review and meta-analysis: A primer. *International Journal of Sports Physical Therapy* 2012; 7(5): 493-503. www.ncbi.nlm.nih.gov/pmc/articles/PMC3474302/

Johnson JA. Influence of race or ethnicity on pharmacokinetics of drugs. *Journal of Pharmaceutical Sciences* 1997; 86(12): 1328-1333. https://doi.org/10.1021/js9702168

Joint Committee for Guides in Metrology (JCGM). *International vocabulary of metrology - Basic and general concepts and associated terms (VIM), 3rd edition*. JCGM, 2012. www.bipm.org/en/committees/jc/jcgm/publications

Kang M, Ragan BG, Park JH. Issues in outcomes research: an overview of randomization techniques for clinical trials. Journal of Athletic Training 2008; 43(2): 215-221. https://doi.org/10.4085/1062-6050-43.2.215

Kuhn TS. The function of measurement in modern physical science. *Isis* 1961; 52(2): 161-193. https://doi.org/10.1086/349468

Matthews JNS, Altman DG. Interaction 2: compare effect sizes not P values. *BMJ* 1996; 313: 808. https://doi.org/10.1136/bmj.313.7060.808

McCluskey A, Lalkhen AG. Statistics II: Central tendency and spread of data. *Continuing Education in Anaesthesia Critical Care & Pain* 2007; 7(4): 127–130. https://doi.org/10.1093/bjaceaccp/mkm020

Miller CJ, Smith SN, Pugatch M. Experimental and quasi-experimental designs in implementation research. *Psychiatry Research* 2020; 283: 112452. https://doi.org/10.1016/j.psychres.2019.06.027

Munro BH. *Statistical methods for health care research*. Philadelphia: Lippincott Williams & Wilkins, 2005.

Noordzij M, Dekker FW, Zoccali C, Jager KJ. Study designs in clinical research. *Nephron Clinical Practice* 2009; 113(3): c218-221. https://doi.org/10.1159/000235610

Norman GR, Steiner DL. *Biostatistics: The Bare Essentials*. Shelton: PMPH-USA Limited, 2014.

Pendrill L. *Quality assured measurement: Unification across social and physical sciences*. Cham: Springer Nature Switzerland AG, 2019.

Co-funded by the Erasmus+ Programme of the European Union

Reference number: 2020-1-SE01-KA202-077806

Program Committee of the Institute of Objective Measurement. Definition of objective measurement. Institute of Objective Measurement, 2000. www.rasch.org/define.htm

Ramaker C, Marinus J, Stiggelbout AM, van Hilten BJ. Systematic evaluation of rating scales for impairment and disability in Parkinson's disease. *Movement Disorders* 2002; 17(5): 867–876. https://doi.org/10.1002/mds.10248

Rasch G. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Sacket DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. *BMJ* 1996; 312: 71-72. https://doi.org/10.1136/bmj.312.7023.71

Shas RR. Inter-ethnic differences in drug response: Implications for drug development and complying with drug regulation. *Clinical Research and Regulatory Affairs* 2015; 32(3): 90-100. https://doi.org/10.3109/10601333.2015.1064131

Sheldon TA. Estimating treatment effects: Real or the result of chance? *Evidence-Based Nursing* 2000; 3(2): 36-39. http://dx.doi.org/10.1136/ebn.3.2.36

Shorten A, Moorley C. Selecting the sample. Evidence Based Nursing 2014; 17(2): 32-33. http://dx.doi.org/10.1136/eb-2014-101747

Shulman LM, Armstrong M, Ellis T, Gruber-Baldini A, Horak F, Nieuwboer A, Parashos S, Post B, Rogers M, Siderowf A. Disability rating scales in Parkinson's disease: critique and recommendations. *Movement Disorders* 2016: 31: 1455-1465. https://doi.org/10.1002/mds.26649

Stevens SS. On the theory of scales of measurement. *Science* 1946; 103(2684): 677-680. https://doi.org/10.1126/science.103.2684.677

Stiegler SM. *The History of Statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press, 1986.

Streiner DL, Norman GR. *Health Measurement Scales. A Practical Guide to Their Development and Use (5th edition)*. Oxford: Oxford University Press, 2014.

Svensson E. Guidelines to statistical evaluation of data from rating scales and questionnaires. *Journal of Rehabilitation Medicine* 2001; 33(1): 47-48. https://doi.org/10.1080/165019701300006542

Swinscow TDV, Campbell MJ. *Statistics at square one (9th edition)*. London: BMJ Publishing Group, 1997. www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one

Co-funded by the Erasmus+ Programme of the European Union

Reference number: 2020-1-SE01-KA202-077806

Tennant A, McKenna SP, Hagell P. Application of Rasch analysis in the development and application of quality of life instruments. Value in Health 2004; 7(Suppl 1): S22-26. https://doi.org/10.1111/j.1524-4733.2004.7s106.x

van Hilten JJ, van der Zwan AD, Zwinderman AH, Roos RAC. Rating impairment and disability in Parkinson's disease: Evaluation of the Unified Parkinson's Disease Rating Scale. *Movement Disorders* 1994; 9(1): 84-88. https://doi.org/10.1002/mds.870090113

Whitley E, Ball J. Statistics review 3: Hypothesis testing and P values. *Critical Care* 2002a; 6(3): 222-225. https://doi.org/10.1186/cc1493

Whitley E, Ball J. Statistics review 4: Sample size calculations. *Critical Care* 2002b; 6(4): 335-341. https://doi.org/10.1186/cc1521

World Health Organization. *International classification of functioning, disability and health: ICF*. Geneva: World Health Organization, 2001. https://apps.who.int/iris/handle/10665/42407

World Medical Association. WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. World Medical Association, 2018. www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation* 1898; 70(12): 857-860. www.rasch.org/memo44.htm

Wright BD, Masters GN: Rating scale analysis. Chicago: Mesa Press, 1982. www.rasch.org

Wright BD, Stone MH: Best test design. Chicago: Mesa Press, 1979. www.rasch.org

Co-funded by the Erasmus+ Programme of the European Union

Reference number: 2020-1-SE01-KA202-077806

Appendix 1: Brief explanations of common terms in clinical research

Term	*Explanation*
Absolute risk	The risk of experiencing a certain event (e.g., a disease or complication) during a certain time period. Expressed as the percentage of people who experience the event of those that potentially could experience it. For example, if 26 out of 100 people will get dementia in their lifetime, the absolute risk is 26/100 or 26%.
Assessment	Structured or unstructured estimation of something. Unstructured assessments are made routinely, for example "I feel stressed" and "this person appears to experience pain". Structured assessments in clinical research typically involve the presence or absence of something (e.g., a diagnosis) based on some criteria, as well as the use of various self- or observer-reported rating scales regarding the presence/absence, frequency or magnitude/severity of something (e.g., symptoms, support, confidence, well-being). Resulting numerical values are either purely qualitative in nature (e.g., 0=sick, 1=healthy; nominal data) or rank ordered (e.g., 0=none, 1=mild, 2=moderate, 3=severe; ordinal data) that do not support valid calculation of, e.g., magnitudes or percentages of differences or changes.
Bias	Something that distorts results of a study. For example, if a certain subgroup of people is over- under-represented in a study this may bias the results.
Clinical importance	The degree to which a study addresses a problem of importance to clinical practice. Alternatively, the degree to which a certain study result (e.g., the difference in outcome between people who receive different treatments) is important in clinical practice.
Focus group	A group of individuals assembled to discuss a topic or answer certain questions regarding a given topic. Typically used in qualitative and mixed-methods studies (focus group interviews).
Grounded theory	A qualitative research approach aiming to develop theories grounded in real-world observations. Data are typically analyzed continuously during data collection and used to inform future interviews.
Informed consent	An ethical principle that requires researchers to obtain the voluntary participation of subjects, after informing them of the study and its possible risks and benefits, and what it means to participate.

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

Intention-to-treat (ITT)	A principle for analyzing data in interventional studies, where data are analyzed according to the group that each participant was assigned to at the beginning of the study, regardless of whether they continued that treatment or not, dropped out prematurely, or even took the wrong study treatment. That is, every patient randomized to the clinical study should be included in the primary analysis within the respective treatment group they were assigned to at randomization. The opposite is referred to as per-protocol (PP) analysis.
Interview	Data collection method where an interviewer asks questions of a respondent, either face-to-face or by other means (e.g., telephone, video call). May be structured (strict compliance to a standardized questionnaire), semi-structured (a list of topics to be covered) or unstructured (no predetermined plan regarding the content or flow of information).
Measurement	Linear quantification of something that is expressed using a defined unit that maintains its size no matter which instrument is used and no matter who or what relevant person or thing is being measured. Resulting numerical values are linear, either without an absolute zero (e.g., temperature according to Celcius or Fahrenheit; interval data) or with an absolute zero (e.g., weight in kg).
Meta analysis	Quantitative integration of results from several quantitative studies addressing a certain research question. For example, several clinical studies regarding the effect of a certain treatment may be summarized in a meta-analysis to provide new and integrated results on its effectiveness. Its qualitative counterpart is called meta-synthesis.
Mixed-methods research	Research collecting and analyzing both qualitative and quantitative data.
Null hypothesis	A hypothesis that says there is no effect, e.g., difference between groups.
Odds ratio	Odds are the probability of an event occurring divided by the probability of that event not occurring. An odds ratio is the comparison of odds for two outcomes describing their association.
Participant observation	A data collection method where the researcher participates in and observes a certain group or culture.
P-value	The probability that a statistical summary of the data (e.g., the sample mean difference between two groups)

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

	would be equal to or more extreme than its observed value under the null hypothesis.
Qualitative research	The investigation of phenomena, typically in an in-depth and holistic fashion, through the collection and analysis of narrative materials, typically verbatim transcriptions of interviews (but also other texts). A common objective is to gain a better and in-depth understanding of something, for example the impact of living with a certain disorder.
Quantitative research	Studies that describe and investigate associations between phenomena using numerical data, typically based on assessments and/or measurements. Aims at making predictions and generalizations.
Randomised controlled trial (RCT)	A truly experimental design, typically regarding the effect of an intervention or treatment. Apart from the treatment being studied, an RCT involves a control condition (a group of participants receiving an alternative or no treatment) and randomization, i.e., letting chance decide whether participants will receive the treatment under study or alternative/no treatment.
Relative risk	Sometimes also called risk ration. An estimate of risk of something in one group compared to (relative to) another, computed by dividing the absolute risk for one group by the absolute risk for another. For example, if people with a family history of a certain condition have a 12% risk of developing the condition themselves and those without a family history have a 3% risk, the relative risk of developing the condition is 4 times (12/3) higher among people with a family history.
Reliability	Defined as the ratio of the true value over the observed (true + error) value, reliability is an index of the absence of error in the quantification or categorization of something. The index can vary between 0 and 1, where higher values represent better reliability and less error. As such, reliability is related to precision and agreement. Primarily used with rating scales and other structured assessment tools, and related to the concept of uncertainty in measurement science (metrology). Reliability can be estimated in various ways and from various perspectives, for example, internally to a rating scale (internal consistency), between raters (inter-rater) or between time points (test-retest).
Right to withdraw	An ethical principle meaning that any participant in a research study has the right to withdraw from (leave) the

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806

	study at any time without having to explain why and without any implications for her/his clinical care.
Saturation	A concept usually used in qualitative data collection, indicating a point when additional data (typically interviews) do not yield any new information.
Statistical significance	Traditionally defined as a probability value (P-value) <0.05 (5%). The P-value is defined as the probability that, e.g., the actual mean difference between two groups would be equal to or more extreme than its observed value.
Systematic review	Rigorous summary of research findings regarding a particular research question or area, based on a systematic search of data bases of research literature, systematic sampling of sources, and systematic data extraction procedures. May be accompanied by a meta-analysis (quantitative) or meta-synthesis (qualitative).
Thematic analysis	A method for analyzing qualitative data (texts, typically interview transcripts) that involves identifying descriptive themes and developing overarching analytic themes that develop new concepts and understanding of the phenomenon under investigation. As such, it usually involves an inductive approach. Closely related to qualitative content analysis.
Validity	(a) A central aspect of the quality of quantitative research studies. Two central aspects are internal and external validity. Internal validity concerns the extent to which results in a study can be trusted (e.g., that it is the studied treatment that is responsible for the effects) and includes aspects such as the statistical methods and sample used in the analyses. External validity concerns the extent to which results will hold over variations in persons, setting, time, or outcome measures. External validity is also referred to as generalizability. (b) A central aspect of the quality and psychometric properties of rating scales and other quantitative data collection methods, where it refers to the degree to which data generated by an instrument represents the variable that it is intended to represent.

Co-funded by the Erasmus+ Programme of the European Union
Reference number: 2020-1-SE01-KA202-077806