Statistical checklist for the design and analysis of quantitative empirical studies

(author: Gerard van Breukelen, Dept. of Methodology and Statistics, Maastricht Feb 2020)

Introduction: scope, purpose and limitations

This is a provisional checklist for the Quality Assurance page on the CAPHRI website (see https://www.maastrichtuniversity.nl/research/school-caphri-care-and-public-health-research-institute/our-research/quality-assurance
This checklist will be updated annually, based upon feedback from users. This checklist is limited to quantitative empirical studies and it does not provide any guidelines for studies involving small-scale in-depth interviews, systematic reviews, or philosophical studies. Further, this checklist serves as an aid in the design and analysis of quantitative studies, not as a mandatory tool. The recommendations in this list are based on decades of broad experience in statistical consultancy and analysis, but can be discussed with a member of the Methodology & Statistics dept. in case of doubt (for details of whom to contact then, see https://stat.mumc.maastrichtuniversity.nl/consultancy). Finally, until further notice, statistical quality control is no part of the audits of research quality in CAPHRI.

The checklist below has three parts: design, analysis, sample size calculation. Importantly, the choice of a statistical method of analysis (e.g. logistic regression) and its use (e.g. which specific model) must be made already in the design phase before data collection. This is both to prevent type I errors (= false positive, finding an effect that does not exist) due to data dredging, and to enable sample size calculation to prevent type II errors (= false negative, overlooking an existing effect due to a too small sample size). Sample size calculation is only possible after the study design, primary research question and outcome, and method of analysis have been specified. A proper study protocol specifies all of these and can be used as an analysis plan for the actual analyses.

1. Choice of design

1.1. Most common designs and purposes

The most commonly encountered designs are

a) Designs for intervention studies (parallel groups randomised trial, cross-over design, cluster randomised trial, multicenter trial, stepped wedge design, quasi-experiment/nonrandomised group comparison),
b) Longitudinal observational (cohort, panel, case-control a.o.) studies,
c) Cross-sectional studies.

Of these, a) and b) are suitable for the study of effects of treatments and exposures, and of the time courses of e.g. health outcomes, albeit with confounding as a major challenge in all nonrandomised designs. In contrast, cross-sectional studies are not suitable for either purpose. *Cross-sectional studies do not allow causal inference*, not even with advanced statistical methods like structural equations modeling, which at best can rule out some causal models. Instead, *cross-sectional studies are suitable for survey purposes* (inventory of e.g. lifestyle, health, or health care at a certain time point) *and for developing and evaluating measurement instruments* (factor structure, validity, reliability, interrater agreement). The remainder of this section provides a brief and provisional checklist per study type, pending updates based on feedback from users.

1.2. Design of intervention and exposure studies

1) **Always have a control group to compare the treated (or exposed) group with.** Treatment and exposure effects cannot be evaluated by within-group comparisons, i.e. by testing whether significant outcome change occurs in a treated (or exposed) group. Within-group changes can have multiple causes, among others seasonal effects, spontaneous recovery, placebo effects, regression to the mean. The relevant comparison is between-groups after treatment (or exposure), that is, between a treated (exposed) and untreated (non-exposed) group.
2) **In intervention studies, use randomised (or cluster randomised) treatment assignment if possible to prevent confounding**. Nonrandomised assignment may easily lead to inconclusive study results (see section “choice of method of analysis”)
3) **In randomised intervention studies**: choose between the following designs (in order of decreasing power and decreasing risk of treatment contamination/carry-over): crossover design, parallel groups (incl. multicenter) trial, cluster randomised trial. Justify your choice. Note: a stepped wedge design can be seen as a compromise between a parallel groups trial and cross-over trial. A stepped wedge cluster randomised trial is likewise a compromise between a cluster randomised trial and cluster randomised cross-over trial. Before choosing a stepped wedge design, contact a statistician to discuss the pros and cons, and the analysis and sample size needed.
4) In a parallel groups (incl. multicenter) trial and a cluster randomised trial: choose a randomisation method (simple, block/stratified, minimisation) and justify it; specify any interim analyses in terms of time point and stopping rules, and justify these.

5) In a cross-over trial: choose treatment sequences (e.g. AB/BA, or ABBA/BAAB) and justify the choice.
6) **Measure covariates, including a baseline measurement of the outcome of interest, before, not after, randomisation.** This is to ensure that drop-outs after randomisation provide a baseline measurement and can be included into the analysis to prevent selection bias (intention to treat analysis).
7) **In a nonrandomised intervention study, if feasible also in exposure studies: include multiple control groups and/or multiple baseline measurements if possible.** These can help to distinguish between a treatment effect and confounding effects (for details, see e.g. Rosenbaum (1995). Observational Studies. New York: Springer; and Van Breukelen, J Clin Epid 2006 page 925, Multivar Behav Res 2013 page 916).
8) **Interim analyses in RCTs**: In RCTs where participants are enrolled sequentially, interim analyses of the available results after a certain % of the planned sample size has been obtained can help to decide whether to continue or to stop recruitment of participants (patients). Such analyses and decisions must be based on sound statistical rules to prevent type I and II errors. **If early stopping is only allowed when interim analysis gives a significant result**, and the number and timing of interim analyses (denoted here as k) is fixed beforehand, then a safe and simple method is to distribute the α (usually 0.05) for significance testing between the k analyses, either equally (so α/k) or with a larger share for later analyses (e.g. if k = 2, one may let α = 0.01 at interim and α = 0.04 at the end). The first option (equal α) requires a larger sample size to have sufficient power in the final analysis. The second option (increasing α) gives low power at interim. **If early stopping for futility (i.e. non-significance at interim and low chances of significance upon study completion) is also allowed**, then the cut-offs for stopping at interim are more difficult and require the help of an expert**.** Note: **sample size recalculation** (e.g. extension) based on interim analyses can easily lead to incorrect data analysis and an inflated type I error risk after trial completion, see e.g. Proschan, Biometrical Journal 2009, pp. 348-357.

1.3. Design of any longitudinal study aiming at causal inference or prediction

9) **Choose the number and timing of repeated measures of outcomes, and also of predictors** if these are measured not just at baseline. Depending on the model to be used for data analysis and the anticipated drop-out rate, 2 up to 5 repeated measures is often a good choice, but more than 5 is rarely so, unless there are sound substantive reasons for the inclusion of each chosen time point of measurement. Note: We here assume a prospective study. Retrospective cohort studies and case-control studies may be less flexible in terms of time points, depending on the availability of data on past exposures and health outcomes.
10) **Choose the dependent variables or outcomes and their method of measurement and scale type** (quantitative, binary, ordinal, nominal, count, survival time). Generally speaking, quantitative outcomes provide more information and require smaller sample sizes than categorical outcomes.
11) **Choose the independent variables or predictors and their method of measurement and scale type.**

12) **In case of covariates (i.e. predictors that are not of primary interest): state their purpose and the implication for the planned analysis.** In general, covariates (i.e. predictors that are not of primary interest) serve one or more of four purposes: (1) to **increase the power of effect testing and precision of effect estimation** (narrowing the confidence interval width) by reducing unexplained outcome variance, (2) to **correct for confounding**, (3) to **test moderation or effect modification** (interaction between the predictor of primary interest, e.g. treatment or exposure, and the covariate), and (4) to **test mediation** (the covariate as a mediator of the effect of treatment or exposure on the outcome). In randomised studies, the first purpose is usually the most important one and the best covariate for that is almost always a baseline measurement of the outcome variable of interest. In nonrandomised studies, the second purpose is the most important one, but it can also become relevant in an RCT, i.e. if there is substantial and covariate-related drop-out (e.g. if drop-out increases with age), in which case the analysis needs to include the covariate to prevent bias. Finally, the role of a covariate as moderator or as mediator depends strongly on the causal model assumed by the researcher. We get back to this in the section "Choice of method of data analysis".
13) **Multiple testing:** This occurs when testing multiple outcomes, or multiple predictors, or both, and it inflates the risk of a type I error (finding a significant effect or relation where none is). The study protocol must specify how this risk is controlled, e.g. by a Bonferroni correction (which leads to a larger sample size needed), or by specifying a primary outcome and primary predictor, the relation between which is the main research question of the study.
14) **Handling non-compliance with treatment in an intervention study**: choose any of the following (but state the choices in the protocol and justify them): intention to treat analysis, per protocol analysis, sensitivity analyses, path (mediation) analysis, causal inference (Complier Average Causal Effect, CACE, principal stratification).
15) **Replacing drop-outs in RCTs**: while it may be tempting to replace patients in an RCT who drop out after randomisation for the sake of sufficient sample size and power, this can introduce bias in at least two ways: First, by obscuring a possible relation between treatment and drop-out. Secondly, by disrupting the randomisation process. A safe solution to the sample size problem arising from drop-out is to multiply the computed sample size with a factor 100/(100-k), where k is the expected % drop-out. Note: drop-outs must always be included into the analyses to prevent bias (intention to treat), but their contribution to the study power is low if drop-out occurs early, e.g. if it occurs between baseline and first follow-up.

1.4. Design of any quantitative empirical study (including a cross-sectional one)

16) **Sampling design**: given a prespecified population from which to sample (e.g. the residents in all nursing homes in Limburg, or the patients of all family practices in Limburg and Brabant), three popular sampling designs allowing inference on the population are: **simple random sampling, stratified sampling, two-stage (cluster) sampling**. The choice between these has important consequences for the analysis and for the sample size. Depending on the research question and the sampling design, complex statistical data analyses may be needed (e.g. inverse probability

weighting). Another popular sampling design, **convenience sampling**, does not (easily) allow for inference on the population, even though most researchers are unaware of this or choose to ignore the problem.

17) **Preventing missing values**: missing values occur in every empirical study outside the laboratory and can lead to substantial bias and/or a loss of power and precision, depending on the missingness rate and pattern. Although there are several methods to deal with missingness in the analysis stage, many of these are either flawed (e.g. last value carried forward), or else require advanced statistical expertise and software for a correct implementation (e.g. multiple imputation). As in health care, prevention is the better cure. So, try to include into the study design strategies for reducing the impact of missing values, for instance incentives to study participants to fill in a questionnaire completely, or attempts to record the reason for dropping out and to do one last outcome measurement at the time of drop out. Last but not least, **minimize the burden for participants** (number of repeated measures, time needed per measurement), for instance by shortening questionnaires and not administering questionnaires that do not serve a clear purpose in the study. See also RJ Little et al., New England J Med, 2012, 367; 14 https://www.nejm.org/doi/full/10.1056/NEJMsr1203730).

2. Choice of method of data analysis

2.1. Analysis of intervention and longitudinal studies for causal inference or prediction

17) **Treatment (exposure) effect evaluation**: **compare between groups, not within groups.** Treatment and exposure effects cannot be evaluated by within-group comparisons, i.e. by testing whether significant outcome change occurs in a treated (or exposed) group. Within-group changes can have multiple causes, among others seasonal effects, spontaneous recovery, plabeco effects, regression to the mean. The relevant comparison is between-groups after treatment (or exposure), that is, between a treated (exposed) and untreated (non-exposed) group.
18) **Adjusting for baseline in an intervention study**: usually, the outcome of interest in an intervention study is measured before (baseline, pre-test) and after (outcome, post-test) treatment. The baseline can be included into the analysis in either of two ways, and at least for quantitative outcomes most other ways are equivalent to either of these two: We can either include the baseline as covariate and analyse the post-test as dependent variable, or analyse as outcome the change from baseline (post minus pre). In a randomised trial or a cluster randomised trial, both methods are unbiased, but the covariate method has more power in most circumstances. In a nonrandomised group comparison where there is a baseline difference between the two groups, use both methods and compare their results. If the methods give contradictory conclusions (Lord's paradox), no final conclusion can be drawn without strong assumptions and/or further analyses (see e.g. Van Breukelen, J Clin Epid 2006, Multiv Behav Res 2013). Note: testing for baseline differences between groups is useful in nonrandomised studies, but not in randomised studies. In the latter, a significant baseline group difference can only occur by chance (type I error) if the randomisation procedure was sound and implemented properly and the covariates were measured before the randomisation (which they should always be).
19) **In prediction and effect evaluation using regression analysis (or ANOVA**): Specify the initial model (predictors, interactions, outcomes) and the strategy for model changes during data analysis. Will the model be adapted (pruned) during data analysis ? If so, by which steps and decision rules, and why ? In this respect, the different roles of covariates need to be kept in mind as discussed in the section on the choice of the design, that is, to increase power and precision in RCTs, to adjust for confounding in nonrandomised studies, or to test moderation/effect modification or mediation in all studies. **To increase power or to adjust for confounding**, include the covariate into the regression (or ANOVA) model as a so-called main effect. If the number of confounders in a nonrandomised study is large, consider propensity scoring. **To test moderation (effect modification)** of a treatment (or exposure) effect by a covariate, include treatment, covariate and treatment*covariate into the model, using dummy coding for categorical predictors. **To test mediation**, a classical approach due to Baron and Kenny (J Personality and Social Psychology, 1986) is to do three regression analyses: (1) regression of outcome Y on cause X (adjusted for confounders) to obtain the total effect of X on Y, (2) regression of mediator M on cause X (adjusted for confounders), and (3) regression of Y on X and M (adjusted for

confounders). This last analysis provides the so-called direct effect of X on Y, and the difference with the total effect in analysis (1) is the mediated (by M) effect of X on Y, which, at least in linear regression, equals the effect of X on M (analysis 2) times the effect of M on Y (analysis 3). This classical mediation analysis rests on several assumptions, notably absence of unmeasured confounders. Modern literature on **causal inference** presents advanced methods to cope with the limitations of the classical method, but at the price of requiring special statistical expertise and special software (some keywords: principal stratification, structural means models). Further, these advanced methods are not assumption-free themselves.

20) **Multiple testing:** This occurs when testing multiple outcomes, or multiple predictors, or both, and it inflates the risk of a type I error (finding a significant effect or relation where none is). The study protocol must specify how this risk is controlled, e.g. by a Bonferroni correction or by specifying a primary outcome and primary predictor, the relation between which is the main research question of the study.
21) **Subgroup analyses**, for instance treatment effect testing for males and for females separately. Such analyses increase the risk of type I errors (due to multiple testing) and of type II errors (due to reduced sample size). Therefore, subgroup analyses should only be done **after finding a significant and 'clinically relevant' interaction** (moderation, effect modification) between the predictor of interest (e.g. treatment) and the predictor used to construct subgroups (e.g. gender). Exception: if the study is known to have too low power to detect interaction, and there are good reasons for expecting interaction, subgroup analyses can be done provided these are presented as exploratory pending confirmation by a more powerful study showing interaction.
22) **Handling non-compliance with treatment in an intervention study**: choose any of the following (but make the choice in the protocol already and justify it): intention to treat analysis, per protocol analysis, sensitivity analyses, path (mediation) analysis, causal inference (Complier Average Causal Effect, CACE, principal stratification).

2.2. Analysis of studies aiming at clustering persons and/or variables

23) **Clustering variables**: the most commonly used method for this is **factor analysis (or the very similar principal components analysis)**, which is used especially to cluster the items in a given health questionnaire into subscales which each measure a certain construct (e.g. attitude, self-efficacy and social norm in health promotion, or physical and mental health of elderly). There are many different versions of factor analysis, either exploratory (in e.g. SPSS) or confirmatory (in e.g. Lisrel, Amos, MPLus). Since all methods are based on correlations or covariances, they presuppose that all variables in the analysis are of the same scale type and range, preferably quantitative, but ordinal scales are also acceptable if they are at least 5-point and do not show strong floor or ceiling effects. For other ordinal and dichotomous variables, special correlations and large samples may be needed. For nominal variables with more than two categories, factor analysis is not suitable.
24) When using factor analysis, specify the software used, the type of correlations analysed, the factor model (in case of confirmatory FA) and the method and criteria used in case of exploratory FA). For exploratory FA: Use as method principal components if the aim is to explain as much observed variance as possible, but use

principal axis factoring (also known as principal factor analysis) to describe the pattern of correlations as well as possible. Choose the number of factors based on multiple criteria: the eigenvalue > 1 rule, the scree plot, the residual correlations (should not exceed 0.10 with a few exceptions), and theory/meaningfulness of course. Rotate the selected nr of factors with oblique methods (Promax, Oblimin), not with orthogonal methods (e.g. Varimax), to allow factor-factor correlation and thereby get a more parsimonious pattern of factor loadings.

25) **Clustering persons**: the most commonly used methods are cluster analysis (K-means, hierarchical), and latent class analysis (including GBTM for repeated measures of a single or a few variables). Here too, the method must be specified as much as possible. For instance, in cluster analysis state the distance measure used as well as the clustering algorithm. In all methods state by which criterion/criteria the number of clusters or latent classes was chosen.
26) **Item missingness in scale (factor, reliability) analysis** : A reasonable and feasible method for handling item missingness if items are about the same construct and have the same response format (e.g. 5-point scales), is the following: First, leave persons and items with too many missings (e.g. more than 20%) out of the scale analysis and do not compute total scores for these persons. Secondly, replace (impute) missing values with the mean of that person on all other items in the same scale, or with the mean of all other persons on that item, or with the mean of these two means. With each of these three methods, first make sure that all items are scored in the same direction, e.g. with score 5 expressing a high level of the construct and score 1 a low level. There are advanced methods for missing item responses, based on item response theory and multiple imputation. However, these require more expertise and special software, raising the question of their cost-effectiveness.

2.3. Analysis of any quantitative empirical study

27) **Nesting must be accounted for in the analysis to prevent errors of significance testing (mostly type I errors) and errors of effect estimation (aggregation bias, ecological fallacy).** Examples of nesting in health research are:
28) **Repeated measures within the same persons.** These measures are correlated, as opposed to independent, observations. If visual acuity is measured on each of both eyes of N patients, the sample size (nr of independent observations) is N, not 2N ! Likewise for measurements at both legs or arms, or EEG measures on different skull locations, or measurements taken at different time points in a follow-up study.
29) **Participants recruited within a large number of organisations** (e.g. students in schools, patients in family practices, nurses and residents in nursing homes, respondents in communities, individuals within their families). If we want to generalise our results to a larger population of organisations, as opposed to only those organisations that participate in the study, we need to treat the organisations in the study as a sample, and treat organisation effects on the outcomes as random effects (not as fixed, as is the case when using dummy coding of organisation). This has two consequences: First, there is now **sampling error at two levels**, organisation and individual. Secondly, the outcomes of individuals in the same organisation are

correlated due to the shared random organisation effect. This in turn has major consequences for the analysis (see next bullet) and the sample size needed (see the section "Sample size (power) calculation"), even if organisational effects on the outcome are small relative to individual differences within the same organisation (keywords: **intraclass correlation, design effect**).

30) Appropriate methods of analysis for nested designs are mixed (= multilevel) regression (also known as growth curve modeling), GEE, repeated measures ANOVA, and latent class analysis (GBTM, LCGA). Of these, the first two are more general in their assumptions and range of application than the last two. Repeated measures ANOVA presupposes a quantitative outcome variable and cannot easily handle missing data or within-subject (time-varying) covariates. Latent class analysis makes restrictive assumptions about individual differences and the pattern of correlations between repeated measures. There is also an even more general method known as mixture modeling, which combines latent class analysis with mixed regression. However, this method requires special software and easily leads to complex models that give rise to numerical and statistical problems when applied to data.
31) Repeated measures on persons nested within organisations can be analysed with so-called 3-level mixed (multilevel) regression models. For these, always contact a statistician with good expertise in mixed modeling.
32) **Exception: if the number of organisations in the study is small**, say less than 10, then we treat organisation as fixed instead of random, using dummy coding, since the number of organisations is too small for a reliable estimation of the random organisation effect variance and the intraclass correlation. Treating organisation as fixed instead of random comes at the price of not being allowed to generalise to a larger population of organisations, but merely to all individuals within the participating organisations.
33) **Selection of study participants for analysis**, e.g. by leaving out persons with missing data, is a questionable research practice that may lead to serious bias. Any selection should be reported and justified. Example: leaving out participants not meeting the inclusion criteria as laid down in a study protocol published before data collection.
34) **Handling missing data:** Report the missingness rate per variable and any pattern found (e.g. baseline predictors of outcome missingness, pattern of missingness across time points of measurement). **Choose, report and justify the method** for handling missingness. Strike a balance between minimizing power loss and bias on the one hand and simplicity and transparency on the other hand. Depending on the **missingness rate and pattern** (keywords: MCAR, MAR, MNAR), and the variable on which missingness occurs (**predictor or outcome**), and the **study design** (RCT or not, repeated measures or not), one of the following methods is a good choice: multiple imputation (notably for missing predictors in observational studies), or mixed regression without imputation (for missing outcomes in RCTs and observational studies), or a simple form of imputation (for missing predictors in RCTs), or complete cases analysis (if the missingness rate is low). For missingness of the MCAR and MAR types, these methods are reasonable, but for MNAR missingness more advanced methods and sensitivity analysis may be needed. Note: for item missingness in scale analyses, see *Guidelines for studies aiming at clustering persons and/or variables*.
35) **If the results are obtained by a (partly) data driven process, such as stepwise regression or selection of a clustering or a factor model based on repeated analyses**

of the same data, perform a cross-validation of the final model on a new sample. Alternatively, consider splitting the sample into two halves, develop the model on the first half and cross-validate it on the second half. Subsequently, develop the model on the second half and cross-validate it on the first half. Choose the model among these two models which performs best in the cross-validation and apply that model to the total sample for maximum precision and power. There are more sophisticated methods for cross-validation, so this is a quick (and perhaps a bit dirty) advice.

36) **Last but not least, analyses and their reporting should not focus on p-values, but on interpretable effect size measures and on confidence intervals (or credibility intervals** in case of Bayesian analyses). Further, these results should be based upon **careful data modeling** which balances between completeness (including all confounders and moderators/effect modifiers) and parsimony, and which checks model assumptions (e.g. normality of random effects and residuals, linearity of relations). See also https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
37) **Note for those in favor of Bayesian analyses**: A general misconception is that Bayesian analysis leads to a higher probability of finding significant results. Classical and Bayesian approaches lead to the same results asymptotically (i.e. when the sample size is large). Results may be different in case of a small sample size if the prior is informative. However, in that case the prior should be constructed critically and justified strongly. Choose and justify a prior based on prior research and choose it prior to collecting your data. Choosing a prior after having collected and inspected your data is bad practice, similar to testing a treatment effect on several outcomes and then selecting the most significant one for reporting.

3. Sample size (power) calculation

38) Most formulae and software for sample size (power) calculations concern the test of the difference between means (t-test, ANOVA), or between proportions (odds ratio), or correlation, with some extensions to multiple regression. **Sample size calculations for multivariate methods for clustering persons or variables** are either not available or require the input of too many unknown parameters to be of much practical use. **Popular rules of thumb such as "Take 10 times as many persons as variables",** lack statistical justification and can easily lead to underpowered studies. The advice below concerns studies aimed at testing outcome differences between groups, or correlations between a small number of variables, as in studies of treatment and exposure effects.
39) **Sample size (power) calculation must be done before the study is run, not post-hoc**. Post-hoc power calculations based on the effect size or correlation as found in the sample are tautological: they give an estimated power above 50% if, and only if, the effect was significant. So, computing post-hoc power for a non-significant effect always leads to the conclusion that the power was too low (less than 50%). Power calculations must be based on what is considered the smallest possible effect size or correlation that is worthwhile detecting. There can be debate about that size and a researcher can choose to adapt the effect size to the feasible sample size, which is certainly not a recommended strategy. Still, stating the effect size that is assumed for the sample size calculation at least ensures transparency about the study power. As such, it is always better than no power calculation at all.
40) **There is not a single sample size/power formula**. There are many formulae and the correct formula for a given study depends on the method of data analysis to be used and thereby on the design, in particular on the nature of the outcome variable (continuous, binary etc.), the predictor of interest (continuous, binary etc.), randomisation yes/no, repeated measures y/n, covariates y/n, nesting y/n etc.
41) **Popular free software for power calculation,** such as e.g. GPower, OpenEpi, or Sealedenvelope, **can only handle the most simple designs**. Using such software for designs that require more advanced statistical methods of analysis than a simple t-test, or 2*2 contingency table, or correlation, leads to an incorrect (usually too small) sample size. Examples:
42) **The sample size for a binary outcome must be at least 60% higher than for a continuous outcome, sometimes even much more,** depending on the distribution of the outcome. Dichotomizing continuous outcomes is therefore usually ill advised.
43) **Nested designs require larger samples than non-nested designs due to the design effect**. In particular, a **cluster randomised trial** (CRT) may easily require 3 times as many participants as a trial with individual randomisation (RCT), and the effective sample size in a CRT is the nr of clusters rather than the nr of individuals (keyword: **design effect**). See e.g. Van Breukelen & Candel, J Clin Epid 2012.
44) **Observational studies need to adjust for confounders. The price of this is a larger sample size than for an RCT.** A reasonable default is to take twice the sample size of an RCT with the same predictor and outcome of interest. This default is based on the assumption that the squared multiple correlation of the predictor with all confounders does not exceed 0.50 (keyword: **variance inflation factor**).

45) **Formulae and software for power calculations can also be used to compute the sample size needed for a given precision, i.e. confidence interval width.** For estimating a mean difference or a difference between two proportions, simply choose a power of 50% and choose as true mean difference half the maximum admissible confidence interval width (i.e. the maximum admissible error of estimation). For correlations, odds ratios and relative risks, the method is a bit more complicated due to the non-normal sampling distribution of those statistics.
46) **Sample size adaptation (e.g. an increase) while the study is running**, based on interim analysis of available results, calls for special methods of analysis after study completion to prevent an increased risk of type I errors (false positives) in significance testing, or undercoverage of confidence intervals. See e.g. Proschan, Biometrical Journal, 2009, 51(2), 348-357). This holds especially if the sample size adaptation is based on the estimated effect size and/or on unblinded data analysis. Safe use of such methods requires a high level of statistical expertise in so-called adaptive trial design.

4. Reporting of analyses

A separate chapter on the QA page will be devoted to this. This section only states a few basic rules and gives references to some useful websites on reporting.

From: Moher et al. CONSORT elaboration and explanation: updated guidelines for reporting parallel group randomised trials. J Clin Epid 2010, 63, e1-e37 (citation from page e16):

Data can be analysed in many ways, some of which may not be strictly appropriate in a particular situation. It is essential to specify which statistical procedure was used for each analysis, and further clarification may be necessary in the results section of the report. The principle to follow is to,
"Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results" *(www.icmje.org). It is also important to describe details of the statistical analysis such as intention-to-treat (see Box 6).*

In this respect, proper data management and archiving of all syntaxes used for statistical analyses deserve attention (see also *NFU richlijnen kwaliteitsborging mensgebonden onderzoek, versie 2019, sectie 9.4).*

Further, reporting should be honest and complete. Two examples of bad practice that both reduce the reproducibility of empirical results are: (1) selective reporting of positive results and leaving out negative results, and (2) presenting exploratory analyses as confirmatory.

For specific guidelines on reporting a good starting point may be the following website:

https://www.equator-network.org/reporting-guidelines/

which provides useful information (documents and keywords) per study type, for instance on clinical trials (CONSORT) and observational studies (STROBE). But please note that this information may not be as complete and up-to-date as that on the dedicated website on which the equator network draws. For instance, for clinical trials check
http://www.consort-statement.org/extensions
which, a.o. others, gives guidelines for adaptive designs that are not on the equator website.

Further, the Equator network does not cover all relevant dedicated websites. For instance, for observational studies it refers to the STROBE initiative, but not to the STRATOS initiative, which is of more recent date and found on https://stratos-initiative.org/publications
Similarly, for systematic reviews it refers to PRISMA, but not to ROBIS.

Finally, useful documents on clinical trials design and analysis may also be found on the website of the European Medicines Agency (EMA), see
https://www.ema.europa.eu/en
and that of the European Food Safety Authority (EFSA), see
http://www.efsa.europa.eu/
and that of the US Food and Drug Administration, see
https://www.fda.gov/
In all three cases, this may require some searching, however.