COST STSM: Quality of reported data on inter-individual variations in the effects and bioavailability of plant bioactives: introducing a quality index and proposing guidelines for complete and accurate reporting – POSITIVe FA 1403 – 41167

Scientific report

1. Location and duration of work

- Laboratory of Phytochemicals in Physiology (Human Nutrition Unit), University of Parma, Italy
- 1st June – 2nd July

2. Objective

The aim of this Short Term Scientific Mission was to prepare the preliminary version of manuscript initially entitled: "**Data reporting on inter-individual variability in clinical intervention studies dealing with effects of plant bioactives"** based on the initiative of the Think Tank Group (TTG). The aim of the initiative was to evaluate the quality of reported data on inter-individual variations in the effect of plant bioactives, including both the effects on cardiometabolic endpoints (topic of WG2) and bioavailability (topic of WG1) and to provide guidelines for design, analysis and reporting of studies on inter-individual variations in the effect of plant bioactives.

3. Material and methods

Ten experts from the Think Tank group of the POSITIVe COST Action created an extensive list of statistical and other parameters to be taken into consideration for the guidelines. The created list of parameters was implemented in the questionnaire for a wider scientific community (315 members of the Action) to examine their interest in the development of the guidelines and opinions on importance of each listed parameter. After evaluation of the questionnaires, the Think Tank group developed the Quality Index Score to be used for evaluation of the quality of data reporting on inter-individual variability (IIV). The score was created based on the parameters considered important by experts who answered the questionnaire. Furthermore, the score was tested by evaluation of 35 collected studies dealing with IIV in response to plant bioactives

During this STSM, the results of the questionnaire were systematically presented, and Quality Index Score was revised and restructured. All steps of this Think Tank group initiative, development and evaluation of the Quality Index Score are described within the methods section in the drafted manuscript. During this STSM, dataset created during the evaluation of 35 studies dealing with plant bioactives was analysed and, based on the results, guidelines on how to report data related to IIV in response to plant bioactives have been created.

4. Summary of the results

The number of identified studies designed to detect specifically the interindividual variations in response to plant bioactives is generally low. A limited number of studies reported the post-hoc analyses of inter-individual variation. However, the quality of reporting results related to inter-individual variability is low, and more specific guidance addressing this issue is required for maximising the impact of scientific research, as well as directing the future actions. The Quality Index was used to assess the general reporting quality but also to identify the list of the most critical parameters that the guidelines should be focused on. This includes, among other parameters, the power of the study addressing the number of between-group comparisons, graphical presentation, measures of variability, reporting on effect size, etc.

The first draft of the manuscript, prepared during the STSM, is included in the Appendix. Some parts of the manuscript still require revisions by Aleksandra Konic-Ristic and Pedro Mena, after which the manuscript will be sent for limited circulation to the main authors and then to all the co-authors. The deadline for the submission of the paper is August 2018.

Appendix

Data reporting on inter-individual variability in clinical intervention studies dealing with effects of plant bioactives

1. Introduction

Plant bioactives, or phytochemicals, represent a large number of diverse, non-nutritive dietary compounds with shown biological activities. They are present in human diet either as constituents of vegetables, fruits or grains, as their natural sources, or in isolated form in supplements or added to fortified foods. Once consumed, phytochemicals are metabolized by host and gut microbial enzymes, deriving a complex mixture of bioactive metabolites present in circulation that are available to interact with different targets in the body and ultimately exert beneficial effects on human health.

There is a large body of evidence that plant bioactives exert an array of beneficial effects highly relevant for the promotion of cardiometabolic health and the prevention of cardiovascular diseases, diabetes type II and their associated risk factors. Cardiometabolic diseases are the major public health concern in westernised world. Although epidemiological data provide sufficient evidence that However, the results obtained in vivo within the clinical trials are often blurred by a noted considerable heterogeneity in responsiveness to dietary interventions. This inter-individual variation may conceal dietary bioactive-health associations and hinder the identification of tailored recommendations for specific subpopulations. The main determinants of the between-subject variation identified are age, sex, genetic variation or microbiota composition, among others However, inter-individual variation in response to the effects of plant bioactives intake on CVD risk factors has not been explored extensively to date.

2. Material and methods (roadmap)

2.1.Brainstorming by 10 experts and questionnaire for all member of the POSITIVe cost action

Two sessions of brainstorming, related to the quality of data reporting on the inter-individual variability (IIV) in dietary intervention studies, were organized by 10 experts from the POSITIVe cost action consortium. During the first session, discussion was focused on evaluation of data reporting and strategies for its systematic improvement and on possibilities for development a score for assessment of quality data reporting. Initial step was a creation of an extensive list of criteria for the evaluation of important parameters (statistical and others) specific for data reporting of studies assessing inter-individual variation in response to plant bioactives. During the second session, experts involved in this activity (in the following text called score developers) decided to consult a wider scientific community and create the questionnaire to be answered by all 315 experts, members of the POSITIVe cost action.

The intention of the questionnaire "How to Report Inter-Individual Variability in Publications" was to assess experts general interest in evaluation of data reporting on IIV in publications, familiarity with the Jadad scale [1] for reporting randomized controlled trials, and interest in integration of the Jadad scale with a new quality index score focused on IIV? The crucial part of the questionnaire was the selection of one or more parameters that they consider important for a high quality data reporting on IIV. The extensive list of 23 parameters was submitted for a review, within the questionnaire. Additionally, the survey assessed experts experience as reviewers or journal editors and their interest of employing quality index score in the review process. The final version of the questionnaire is available in supplement.

2.2.Parameters to be considered and development of the quality index score. Definition of categories.

After collection and evaluation of the questionnaires score developers approach to the selection of parameters to be included in the score in a following way. Parameters selected by 50% or more experts were directly taken into consideration in the process of development of the quality index score. Parameters that were considered important by 40-50% of experts were additionally discussed and evaluated by score developers, while parameters selected by less than 40% of participants were excluded from the score. As a result, the list of 11 parameters grouped in 4 categories was defined for design of the quality index score and development of guidelines for data reporting on IIV. The next step was a creation of the first version of dictionary, with detailed definitions of conditions related to data reporting and assigned marks, as a base for the calculation of the quality index score. Marks were assigned to each parameter and its related condition from these four categories in a following way: if particular parameter is not reported in the study at all its score is 0; if it is reported but not informative enough to completely describe IIV its score is 0.5; and if it is reported and completely illustrate IIV, its score is 1. The exception to this scale is the last category, related to the individual data availability. Considering this category as the most important for the assessment of IIV developers extended its scale to 0-1-2. The detailed dictionary of the quality index score and parameter's marks is presented and described in the results section.

2.3.Collection and evaluation of papers reporting IIV in plant bioactives (n=19 expert).

Score developers decided to test and validate the quality index score by evaluating existing studies based on the comprehensiveness of data reporting on IIV. For this reason, 35 relevant papers were collected by 19 experts, members of POSITIVe Think-Tank group (in the following text called evaluators). Before applying the quality index score on all 35 studies, the pilot testing of the dictionary comprehensiveness was done by evaluation of five studies. Each study was evaluated independently by two or three different evaluators and the final scoring was done based on their consensus. After the testing, evaluators provided their critical opinion to the developers and helped in revising the dictionary to be clearer and user friendly. Revision was related to the definitions of conditions, adding certain parameters that were found to be important after the testing and regrouping parameters between the categories. The final version of the dictionary for quality index score was then created and all 35 studies were evaluated and scored in the same way as in the testing phase. After the evaluation, data reporting of each study was assessed by the total score and 4 sub-scores related to the categories. Regardless the evaluation done by employing quality index score, evaluators assessed the overall quality of data reporting on IIV for each study as bad, mild, good, OK based on their personal opinion. Quality index score was then validated by comparing these two methods of evaluation.

2.4.Statistics

Accuracy of developed quality index score was examined by analysing relations between quality of the studies assessed by personal opinion of the experts (weak, mild, good) and by quality index score. Ordinal variable was created for the quality level assessed by experts (weak=1, mild=2, good=3). Overall quality score was calculated for each study as sum of all marks given to each study divided by 11 (number of selected parameters). Completeness of reporting within each defined category was calculated for each study as sum of all marks assigned to the parameters from the category divided by the number of parameters for that particular category. In this way completeness of reporting was standardized for all categories. Spearman correlation coefficient was calculated to assess relation between overall quality index score and quality level's assessed by experts. Cohen's kappa coefficient was calculated to test the agreement between tertiles of the overall quality index score and quality levels assessed by the experts. Impact of completeness of each defined category within the quality index score on the quality levels assessed by experts was also assessed by Spearman correlation coefficient. A value of p< 0,05 was taken to indicate a significant result. All analyses were performend using SPSS statistical sofware (IBM SPSS statistics 20, SPSS Inc., Chicago, IL USA)

2. **Results**

3.1. Questionnaire results

Questionnaire 'How to Report Inter-Individual Variability in Publications' was answered by 21 % (n=66) of experts. Majority of them (96%) considered development of the quality index score important for assessment of quality of data reporting on IIV. Experts had different approach to the selection of parameters. More critical approach, and selection of less than 10 parameters, was noticed in 44% (n=29) of those who answered the questionnaire. This subgroup of experts was focused on statistical parameters (sample size, normality, p-value related on IIV), parameters related to measures of central tendencies and dispersion as well as parameters related to the population characteristics and stratification by different factors that could affect IIV. Other parameters were represented in less than 35 % of their questionnaire. The rest of experts were more likely to choose more than 10 parameters, 24 % of them selected between 11 and 15 parameters, while 32 % of them considered more than 15 parameters as important. However, answers of the total number of experts that answered the questionnaire were taken into consideration for development of the quality index score. Percentage of experts, who considered listed parameters important to be reported when assessing IIV in response to plant food bioactives, are presented in figure

1. Sample size calculation, dispersion parameters and population characteristics, as the most important parameters related to IIV, were selected by more than 70% of experts, as expected. Data distribution, p value, mean, outliers, data before and after intervention by subgroups and stratification by different factors were considered important by more than 50% of experts. However, stratification by ethnicity, hormonal status, polymorphism and gut microbiota composition were selected by 40-50% of experts as well as median, full data presented on individual level provided in the supplementary material and data depicted in scatter or box plots. Graph of data distribution, individual presentation of non-normally distributed data and individual data for study end point were assessed as important by less than 35% of experts. Additional parameters under the option "other" were suggested by 12 experts. Only one of them (coefficient of variation) was related to the data reporting from the statistical point of view while all other parameters were more general and not directly related to the data reporting but to factors important for IIV like dietary habits, physical activity etc. The experts were asked about the Jadad scale for reporting randomized controlled trials and 36.4 % of them stated that they are familiar with this scale developed in 1996 by Jadad et al. [1] while 59 % of experts said that it would be important to supplement the Jadad scale with the quality index score related to IIV.

Among the experts who answered the questionnaire there were 16 (24 %) members of editorial board in one or more highly rated scientific journals (British Journal of Nutrition, Scientific Reports, Molecular Nutrition and Food Research, Food and Nutrition Research, etc.). Majority of experts (80 %) also said that journals' editors might be interested in the quality index score to be used in evaluation of manuscripts dealing with IIV in response to plant bioactives.

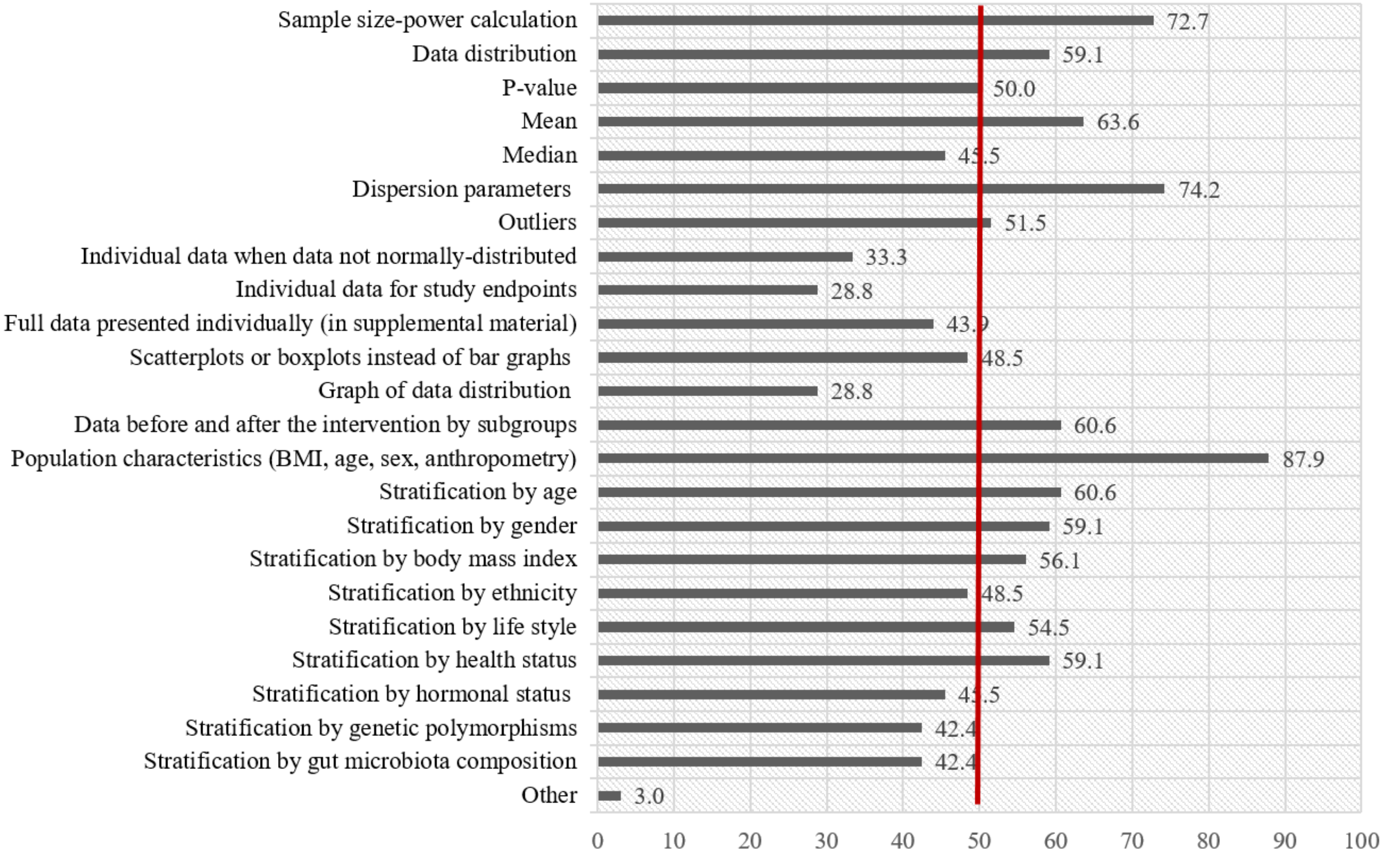

Figure 1 Percentage of experts who considered listed parameters important to be reported when assessing inter-individual variability in response to plant food bioactives

3.2. Development of the quality index score (dictionary)

The final version of the dictionary with defined conditions for each parameter and assigned scores is presented in table 1. Sample size/power calculation, distribution of the data and p-value related to the IIV were grouped in one category as they are all crucial for the high quality data analysis and reporting. Moreover, all of them were selected as important by more than 50% of experts. Additionally, after testing the first version of the dictionary for the quality index score, evaluators stressed the importance of reporting on the effect size of the applied statistical tests as the important parameter for complete understanding of the p value [2]. Finally, the sum of scores based on the first category – *Statistics*, reflects on quality of data reporting with respect to four parameters: sample size/power calculation, distribution of data, p value and effect size. All parameters in this category could be assessed by dichotomous score of 0 or 1.

Another set of six parameters listed in the questionnaire are regrouped in four and integrated in the second category – *Reporting*. Reporting on general characteristics of the subgroups where IIV was evaluated (age, gender, body mass index, smoking status etc.) was the parameter most often selected by experts (87.9 %).

Since they had diverse opinions about stratification according to the different characteristics it was decided to keep this parameter open for any characteristics collected for the subgroups. Reporting on data for study end-points (before and/or after the intervention) was also included in the *Reporting* category since it was considered important by 60 % of experts. Furthermore, measures of central tendencies and dispersion parameters reported for each subgroup, where IIV was evaluated, are merged and included in this category as a single parameter. Reporting on outliers, as an important factor for assessment of IIV selected by more than 50 % of experts, was integrated in the *Reporting* category as the fourth parameter. Data reporting on end-points by subgroups and on outliers could be scored by values 0, 0.5 or 1 depending of the comprehensiveness of the reporting as explained above. On the other hand, reporting on general characteristics and measures of central tendencies & dispersion parameters could be scored by 0 or 1. In conclusion, overall score related to the *Reporting* category, reflects the quality of reporting on four parameters: general characteristics of the subgroups, data reporting for end-points by subgroups, measures of central tendencies & dispersion parameters and outliers.

Third category considered important by developers is *Data presentation***.** Though complexity of the tables was not listed as a parameter in the questionnaire, after the testing, evaluators suggested to include it in this category. The purpose of this parameter is to assess reporting on additional values apart of measures of central tendencies & dispersion parameters, that could reflect IIV (min, max, interquartile range etc.). Presenting data as a scatter plot or box plot instead of bar chart was considered important by 49 % of experts. However, developers decided to extend this parameter and assess a quality of graphical presentation of data making distinction between presenting data on primary and secondary outcome. It means that, in the final version of the dictionary, the scatter plots, box plots or heat maps that depicts data related to the primary outcome are considered as the most useful (score = 1); histograms that depicts primary outcome or scatter plot/box plot that depicts secondary outcome is considered as weak but still useful way of illustrating IIV (score = 0.5); and bar charts, curves etc. for any end-point are considered as not helpful in assessment of IIV (score = 0).

Fourth, and considered as the most useful, category is related to the individual data availability i.e. transparency of analyzed data set. However, developers distinct three different levels within this category. Individual data available for each end-point, reported together with the characteristics of the study participants on individual level, is the most appreciated option (score = 2). Individual data reported for each end-point but without any additional characteristics of study participants are still considered useful but less than the previously described option (score = 1) while studies that did not show any individual data are not scored for this category (i.e. score = 0).

Table 1 Dictionary of conditions related to data reporting on IIV and associated scores for the evaluation of quality of data reporting in intervention studies dealing with plant bioactives

Category	Parameters	Condition	Score
Statistics	Sample size - power calculation	Authors reported on: (1) power calculation focused on assessing inter-individual variation based on primary OR the most limiting outcome AND (2) on all the data used in power calculation (% of statistical power, significance level, expected dropout rate, expected difference between groups of the mean or % with the event) AND (3) the resulting sample size per each group.	1
		Authors did not describe sample size taking into account all the three previous conditions.	0
	Normality distribution of data	Authors specified the test used for normality (OR indicate something related to data normality, for instance, log transformation)	1
		Authors did not report any information related to the normality or distribution of the data in general	0
	p-value	Authors reported p-value that support IIV (e.g. p-value related to the examination of differences between two or more factors affecting IIV such as sex, age, genotypes, etc.).	1
		Authors did not report any p-values related to the IIV	0
	Effect size	Authors reported the magnitude of the IIV for the selected outcome(s) by standardized mean differences as an index of effect size (Cohen's d, % of coefficient of variation, etc.) or any parameters related to the effect size suitable for the conducted statistical tests.	1
		Authors did not indicate any parameters related to the effect size i.e. magnitude of the IIV for the selected outcome(s) (any of standardized mean differences was not reported).	0
Reporting	General characteristics of the subgroups where IIV was evaluated	Authors reported on one or more general characteristics (for instance, ethnicity, BMI, age, gender, smoking status, etc.) for each of the subgroups where IIV was evaluated.	1
		Authors did not report any of general characteristics for the study sample (for instance, ethnicity, BMI, age, gender, smoking status, etc.) on each subgroup where IIV was evaluated.	0
	Data reporting for end-points by subgroups	Both pre- and post-intervention data (or post-intervention data as % change with respect to a provided baseline value) were reported for different subgroups where IIV was evaluated.	1
		Post-intervention data (or % change without a provided baseline value) are provided by each subgroup where IIV was evaluated.	0.5
		Neither pre- nor post-intervention data were reported for different subgroups where IIV was evaluated.	0
	Measures of central tendencies and dispersion parameters	Authors reported on one or more measures of central tendencies (mean, median, etc.) AND one or more dispersion parameters (standard deviations, standard error, inter quartile range, 95% confidence interval, etc.) for EACH subgroup where IIV is evaluated.	1

		Authors did not report any measures of central tendencies or dispersion measures for subgroups where IIV was evaluated, regardless of reporting these parameters for the total sample.	0
	Outliers	Authors indicated outliers AND described them (explained the reason for treating them as outliers).	1
		Authors indicated that outliers existed and that were excluded from the analysis but without describing them.	0.5
		Authors did not indicate any information related to outliers.	0
Data presentation	Tables	Tables contain additional measures of variability (min-max, inter quartile range, outliers values, etc.) or individual measures (responders/non-responders, etc.).	1
		Tables did not contain any extra measures of variability (min-max, inter quartile range, outliers values, etc.) nor individual measures (responders/non-responders, etc.).	0
	Graphs	Authors presented data for the primary outcome by scatterplots, boxplots or heat maps.	1
		Authors presented data by histograms for a primary outcome OR as scatterplots and boxplots for secondary outcomes.	0.5
		Data are graphically presented as bar chart, curves, etc. (for any of the study point-before or after the intervention) but not as scatterplots, boxplots, heat maps or histograms.	0
Individual data availability	Presentation of full data & population characteristics	Authors provided individual data for each end-point, together with the characteristics of the samples on the individual level, in the paper or in the supplemental material.	2
		Authors provided individual data at each end-point (even presented in the figures) but without any additional characteristics of the sample on the individual level.	1
		Authors did not provide individual data.	0

3.3. Validation of the quality index score - evaluation of collected studies

Quality index score and sub scores related to defined categories have been calculated for all collected studies (n = 35). Quality of data reporting on IIV was additionally assessed by expert's personal opinion for 30 studies. The weak but significant agreement was found between tertiles of overall quality index score and three levels of quality (weak, mild and good) assessed by experts (Cohen's k=0.216, p=0.054). However, the same agreement was confirmed by Spearman rank correlation coefficient (r = 0.697, p < 0,001). According to the expert's personal opinion, independent from the quality index score, 2 studies were assessed as weak regarding the data reporting on IIV, 12 studies were assessed as mild and 16 studies as good. Numbers and percentages of studies that reported any data (scored either with 1 or 0.5 according to the dictionary) on selected parameters, within these three groups of studies, are presented in table 1. Significant correlation was found between completeness of particular score categories (*Statistics and*

Reporting) and quality levels assessed by experts (Spearman's r =0.566, p = 0.001; r = 0.509, p = 0.004 respectively). On contrary, *Data presentation* category was in inverse correlation with the expert's opinions independent of quality index score (r = - 0.365, p = 0.047). Individual data availability, as an independent parameter, was not analysed in this way since only four studies provided data but without additional characteristics of the groups where IIV was evaluated. Two of them were assessed as mild and another two as good. Percentages of weak, mild and good studies that reported on selected parameters are presented in figure 2. As confirmed with indirect correlation, graphs and tables as defined in dictionary were not key parameters for comprehensive explanation on IIV. Out of 14 studies that presented graphs as defined in the dictionary 71% was assessed as weak or mild. Only two studies reported on additional measures of variability (min-max, inter quartile range, etc.) or individual measures (responders/non-responders, etc.) within the tables, and they were assessed as mild. The same is for outliers. On the other hand, more than 60% of studies that reported on other parameters from the *Reporting* category were assessed as good by experts with respect to the quality of explanations related to the IIV (figure 2.). Out of 21 studies that reported on data for end-points by subgroups, 63% was assessed by experts as good and there were no studies assessed as weak. The result was similar for studies that reported on measures of central tendencies & dispersion parameters by subgroups, 68% of them were assessed as good. Out of 20 studies that reported on general characteristics of the sample where IIV was evaluated, 60% were assessed as good. However, *Statistics* category was found as the most important for high quality data reporting on IIV. All studies, that were assessed as good by experts, reported on, at least one, parameter from the *Statistics* category, while 75% of them reported on two or more parameters from this category. Moreover, 91% of those that reported on effect size were assessed as good by expert's independent opinion. Only one study reported on sample size as described in the dictionary, as expected. Studies that reported on data distribution and p value related to the IIV were assessed as good in 67% and 61% of cases respectively (figure 2).

Table 2 Reporting on parameters within defined categories (n(%)) and completeness of reporting (%) within each category by quality of studies assessed by experts opinion

		Weak (n = 2)		Mild (n = 12)		Good (n = 16)		Total (n = 30)		Completeness vs Quality levels assessed by experts	
Category	Parameter	n (%)	Completeness	n (%)	Completeness	n (%)	Completeness	n (%)	Completeness	Spearman correlation coeff.	p value
Statistics	Sample size - power calculation	0 (0 %)	37.5%	0 (0%)	20.8%	1 (6%)	48.4%	1 (3%)	36.7%	0.519	0.003
	Normality/distribution of data	1 (50 %)		2 (17%)		6 (38%)		9 (30%)			
	p-value related to IIV	1 (50 %)		8 (67%)		14 (88%)		23 (77%)			
	Effect size	1 (50 %)		0 (0%)		10 (63%)		11 (37%)			
Reporting	General characteristics of the subgroups where IIV was evaluated	1 (50 %)	25%	7 (58%)	44.8%	12 (75%)	59.4%	20 (67%)	51.3%	0.509	0.004
	Data reporting for end-points by subgroups	0 (0%)		8 (67%)		13 (81%)		21 (70%)			
	Measures of central tendencies and dispersion parameters	1 (50 %)		6 (50%)		14 (88%)		21 (70%)			
	Outliers	0 (0%)		2 (17%)		0 (0%)		2 (7%)			
Data presentation	Tables	0 (0%)	25%	2 (17%)	33.3%	0 (0%)	14.1%	2 (7%)	22.5%	-0.365	0.047
	Graphs	1 (50%)		7 (58%)		6 (38%)		14 (47%)			
Individual data availability		0 (0%)		2 (17%)		2 (13%)		4 (13%)			

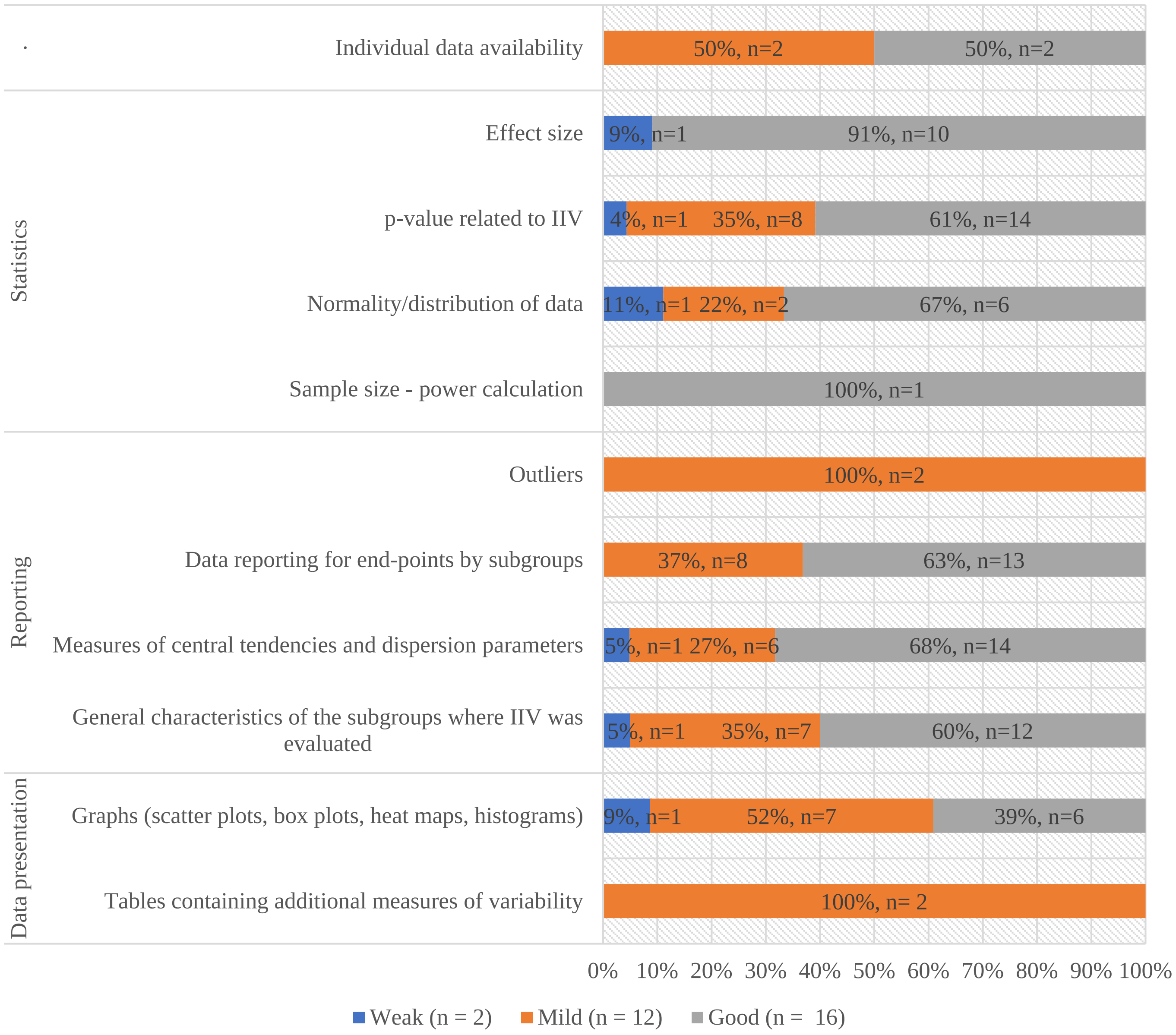

Figure 2 Percentage of good, mild and weak studies with respect to the reported parameters within defined categories

3. **Discussion**:

4.1. General discussion about reporting IIV in plant bioactives and the QI

To be added

4.2. Discussion by category and parameter (some figures??)

Individual data availability was considered as the most useful parameter for understanding the IIV by readers. Moreover, as defined in the dictionary, individual data followed by sample characteristics (age, gender, ethnicity etc.) provided also on individual level are of the greatest help, not only for understanding IIV on the one trial level but principally for the meta analyses [3]. Although data sharing increase in clinical research community, we found only four out of 35 studies, dealing with effects of plant bioactives, that provided individual data but without additional characteristics of the sample on individual level [4–7]. Despite individual data availability, two of them were assessed as mild regarding data reporting on IIV because sample characterises on individual level were not available and reporting on IIV was mild. However, in order to have a better understanding of IIV in such trials we highly recommend authors to prepare their data for sharing either in publication or in one of the existing data bases of clinical studies such as ClinicalTrials.gov. Ohmann et al. summarized all principles and recommendations for Individual Participant Data (IPD) sharing developed by different authorities that should be followed in data sharing process [8]. In case that authors decide not to share IPD, comprehensive approach to the data reporting on parameters from the *Statistics* category is necessary, starting from the sample size calculation that takes into account IIV. As a part of study design procedure, an adequate calculation of sample size is the essential element of the high quality data reporting on IIV. After the evaluation of 35 studies, regarding the sample size calculation, we found that conditions given in the dictionary are too demanding for this research area. There are still not enough studies, dealing with effects of plant bioactives that reported on IIV between different groups, to find the population standard deviation by particular groups and related intervention. That is the reason why we found only one study that took into consideration IIV for the sample size calculation [6]. However, it is highly advisable to look for all studies that reported on similar results and, if they exist, to take into account reported IIV to calculate a sample size.

Reporting on the distribution of data, when dealing with IIV, is as important as for all other data reporting but we want to emphasize importance on checking the data distribution and other assumptions that has to be met in order to get accurate results. Misunderstanding of assumptions that have to be satisfied before employing parametric tests is often. For example, the assumption for dependent t-test, that the sampling distribution of the differences between scores of two measures should be normal is usually misinterpreted by normal distribution of scores themselves; the assumption of normal distribution within the groups important for employing one-way anova is misinterpreted as normal distribution of the total sample etc. Assumptions for the extended list of parametric tests are explained in details by Field et al. [9]. Visual methods for checking normality like histograms, box plots, stem-and-leaf plots etc. could be helpful for large data sets since statistical tests (e.g. Kolmogorov-Smirnov test and Shapiro-Wilk test) could be significant i.e. reject the hypothesis of normal distribution even if deviations from normality are small. On the other hand, for small samples (< 30), statistical tests are necessary [10]. Shapiro-Wilk test is recommended as as the best choice for testing the normality of data [11].

Despite the fact that p value, related to the IIV, was the most often reported parameter among evaluated studies (77%) there were still 39% of them assessed by experts as weak or mild. Reporting on p value, without reporting on measures of central tendencies and SD neither the change by subgroups, is not as informative as p value reported together with these parameters. For example, reporting on p value, related to different effects between men and women, as explanation of the scatter plot, without reporting on mean and SD for each subgroup at each end-point, is not as informative as it would be if authors provide these data numerically, especially if differences are small. Such studies cannot be used in meta-analysis related to the IIV, unnecessarily limit the understanding of IIV by not reporting on data that definitely exist, and, as mentioned above, cannot be helpful for sample size calculation in future studies. In conclusion, reporting on p value should always be followed by reporting on mean and SD for compared groups and preferably by reporting on median, min-max etc. at each end-point.

Additional parameter important for understanding statistical significance of the effects of the intervention (p value) is the effect size that actually reviled substantive significance. Despite the fact that the p value provides the information that effect of the intervention exists or not it doesn’t say anything about the size of the effect (magnitude of the difference between groups/treatments) [12]. As shown in the results section, 91% of studies that reported on effect size were assessed as good by experts regarding the quality of data reporting on IIV. Thus, it is highly recommended to report on effect size together with the p value. Guidance on how to calculate and interpret an effect size for different types of analysis are summarized by Durlak and Field et al.[9,13]

Outliers, End-points by subgroups, General characteristics by subgroups/any that were examined in the survey: to be added

Though 61% of studies, that presented data graphically as defined in dictionary, were assessed by experts as weak or mild, we still recommend graphical representation of data but as addition to the, previously described, crucial statistical parameters. Moreover, as effect size additionally explain p value, in the same way appropriate graphs could disclose statistics reported in tables. This is especially important for small sample sizes which is usually the case in nutritional intervention studies. Scatter plots of raw data are the most useful graphs regarding transparency of the results and especially regarding the IIV when we are dealing with small sample sizes. Such graphs could clearly show where standard deviations come from, particularly if characteristics of the subgroups are reported. Box plots are also very welcomed way of data presentation since it shows outliers and variation. Nevertheless, box plot still summarises data and are more meaningful for the larger sample sizes. The same is for histograms, they are helpful in depicting distribution of large samples but for the small samples it is hard to understand it clearly [9,14]. Bar charts are not recommended way of presenting data since they cannot say much more than table, moreover, they are not helpful in revealing distribution of data at all since the same bar chart could be created based on differently distributed data sets [14]. Another way of favourable data reporting are tables with more parameters that could uncover the distribution like min-max, median, inter-quartile range, coefficient of variation etc. In this way readers get clearer picture about data distribution and direction of variation than from mean and standard deviation. Brindani et al. reported on central tendencies and described the data distribution from their sample by additional parameters... [15]

Conclusion:

To be added

References:

1. Jadad, A. R.; Moore, R. A.; Carroll, D.; Jenkinson, C.; Reynolds, D. J. M.; Gavaghan, D. J.; McQuay, H. J. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control. Clin. Trials* **1996**, *17*, 1–12, doi:10.1016/0197-2456(95)00134-4.

2. Durlak, J. A. How to Select, Calculate, and Interpret Effect Sizes. *J. Pediatr. Psychol.* **2009**, *34*, 917–928, doi:10.1093/jpepsy/jsp004.

3. Doshi, P.; Goodman, S. N.; Ioannidis, J. P. A. Raw data from clinical trials: Within reach? *Trends Pharmacol. Sci.* **2013**, *34*, 645–647, doi:10.1016/j.tips.2013.10.006.

4. Borel, P.; Desmarchelier, C.; Nowicki, M.; Bott, R.; Morange, S.; Lesavre, N. Interindividual variability of lutein bioavailability in healthy men: Characterization, genetic variants involved, and relation with fasting plasma lutein concentration. *Am. J. Clin. Nutr.* **2014**, *100*, 168–175, doi:10.3945/ajcn.114.085720.

5. Setchell KD、ubi JE, Cole S, Guy T, H. B. Dietary Factors Influence Production of the Soy IsoflavoneMetaboliteS-(-)EquolinHealthyAdults. *J. Nutr.* **2013**, *143*, : 1950–1958, doi:10.3945/jn.113.179564.contributes.

6. Mackay, D. S.; Gebauer, S. K.; Eck, P. K.; Baer, D. J.; Jones, P. J. H. Lathosterol-to-

cholesterol ratio in serum predicts cholesterol-lowering response to plant sterol consumption in a dual-center, randomized, single-blind placebo-controlled trial. *Am. J. Clin. Nutr.* **2015**, *101*, 432–439, doi:10.3945/ajcn.114.095356.

7. Setchell, K. D. R.; Cole, S. J. Method of Defining Equol-Producer Status and Its Frequency among Vegetarians. *J. Nutr.* **2006**, *136*, 2188–2193, doi:10.1093/jn/136.8.2188.

8. Ohmann, C.; Banzi, R.; Canham, S.; Battaglia, S.; Matei, M.; Ariyo, C.; Becnel, L.; Bierer, B.; Bowers, S.; Clivio, L.; Dias, M.; Druml, C.; Faure, H.; Fenner, M.; Galvez, J.; Ghersi, D.; Gluud, C.; Groves, T.; Houston, P.; Karam, G.; Kalra, D.; Knowles, R. L.; Krleža-Jerić, K.; Kubiak, C.; Kuchinke, W.; Kush, R.; Lukkarinen, A.; Marques, P. S.; Newbigging, A.; O'Callaghan, J.; Ravaud, P.; Schlünder, I.; Shanahan, D.; Sitter, H.; Spalding, D.; Tudur-Smith, C.; van Reusel, P.; van Veen, E.-B.; Visser, G. R.; Wilson, J.; Demotes-Mainard, J. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open* **2017**, *7*, e018647, doi:10.1136/bmjopen-2017-018647.

9. Field, A. P.; Miles, J.; Field, Z. *Discovering statistics using R*; 2012;

10. Ghasemi, A.; Zahediasl, S. Normality tests for statistical analysis: A guide for non-statisticians. *Int. J. Endocrinol. Metab.* **2012**, *10*, 486–489, doi:10.5812/ijem.3505.

11. Thode, H. Testing For Normality. **2002**, doi:10.1201/9780203910894.

12. Sullivan, G. M.; Feinn, R. Using Effect Size—or Why the P Value Is Not Enough., doi:10.4300/JGME-D-12-00156.1.

13. Durlak, J. A. How to select, calculate, and interperet effect sizes. *J. Pediatr. Psychol.* **2009**, *34*, 917–928, doi:10.1093/jpepsy/jsp004.

14. Weissgerber, T. L.; Milic, N. M.; Winham, S. J.; Garovic, V. D. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm., doi:10.1371/journal.pbio.1002128.

15. Brindani, N.; Mena, P.; Calani, L.; Benzie, I.; Choi, S. W.; Brighenti, F.; Zanardi, F.; Curti, C.; Del Rio, D. Synthetic and analytical strategies for the quantification of phenyl-γ-valerolactone conjugated metabolites in human urine. *Mol. Nutr. Food Res.* **2017**, *61*, 6–10, doi:10.1002/mnfr.201700077.