Review

Preanalytical investigations of phlebotomy: methodological aspects, pitfalls and recommendations

Cristiano Ialongo*[1,2], Sergio Bernardini[2,3]

[1]Department of Human Physiology and Pharmacology, University of Rome Sapienza, Rome, Italy
[2]Laboratory Medicine Department, "Tor Vergata" University Hospital, Rome, Italy
[3]Experimental Medicine and Surgery Department, "Tor Vergata" University, Rome, Italy

*Corresponding author: cristiano.ialongo@gmail.com

Abstract

Phlebotomy is often addressed as a crucial process in the pre-analytical phase, in which a large part of laboratory errors take place, but to date there is not yet a consolidated methodological paradigm. Seeking literature, we found 36 suitable investigations issued between 1996 and 2016 (April) dealing with the investigation of pre-analytical factors related to phlebotomy. We found that the largest part of studies had a cohort of healthy volunteers (22/36) or outpatients (11/36), with the former group showing a significantly smaller median sample size (N = 20, IQR: 17.5-30 and N = 88, IQR: 54.5-220.5 respectively, P < 0.001). Moreover, the largest part investigated one pre-analytical factor (26/36) and regarded more than one laboratory test (29/36), and authors preferably used paired Student's t-test (17/36) or Wilcoxon's test (11/36), but calibration (i.e. sample size calculation for a detectable effect) was addressed only in one manuscript. The Bland-Altman plot was often the preferred method used to estimate bias (12/36), as well as the Passing-Bablok regression for agreement (8/36). However, often papers did assess neither bias (12/36) nor agreement (24/36). Clinical significance of bias was preferably assessed comparing to a database value (16/36), and it resulted uncorrelated with the size of the effect produced by the factor (P = 0.142). However, the median effect size (ES) resulted significantly larger if the associated factor was clinically significant instead of non-significant (ES = 1.140, IQR: 0.815-1.700 and ES = 0.349, IQR: 0.228-0.531 respectively, P < 0.001). On these evidences, we discussed some recommendations for improving methodological consistency, delivering reliable results, as well as ensuring accessibility to practical evidences.

Key words: phlebotomy; preanalytical phase; statistical data analysis; methods

Received: April 27, 2016 *Accepted: November 25, 2016*

Introduction

The investigation of pre-analytical factors in laboratory medicine is pivotal to improve the overall clinical laboratory quality, and in turn to ensure the patient safety (1). In this regard, phlebotomy is addressed as a crucial process in the pre-analytical phase, in which a large part of laboratory errors is thought to arise, having the potentialities to affect largest part of medical decisions (2,3). Indeed, apart from the provision with qualitatively appropriate supplies that depends on the healthcare service's choice, there are no other means than the operator's skills and compliance with standard procedures to ensure the adequate sample quality (4). As far as pre-analytics is a major concern in current laboratory medicine and an issue for practitioners and researchers, this field of investigation should be fostered in order to produce evidences for best practice (5,6). Indeed, unnecessarily complicating the patient management without any actual improvement, or even oversimplifying and then flawing its safety, might undermine the operator's awareness of a mandatory and careful pre-analytics. Carrying out a pre-analytical investigation poses some methodological concerns regarding the statistical framework used to assess the investigated factor. Furthermore, with respect to phlebotomy, there are some more specific issues arising on the choice of the appropriate cohort,

the standardization of procedures and the deliverability of results to non-academic readers. Thus, such studies should grant the highest reliability, whereby dispelling any doubt of misleadingness.

Scope of the present paper is assessing the appropriateness of the methodology used to carry out a pre-analytical investigation of phlebotomy, fitting researchers with specific recommendations on study set up and delivery. Thereby, in this two-part paper, part I aimed to gather evidences through a review of available literature and summarizing evidence, and part II is concerned with methodological appropriateness and the choice of suitable procedures.

PART I – Evidences

Literature search and data analysis

The literature database MEDLINE was searched using PubMed for papers issued in the last twenty years (January 1996 to April 2016). In order to structure the search, we first set a combination of keywords that targeted the general topic of pre-analytics in laboratory (e.g. "laboratory test", "pre-analytic", "biological OR individual"). Afterwards, we refined the search focusing on ten topics each of which addressed a pre-analytical aspect of phlebotomy strictly related to the operator's choice, and setting appropriate keywords (Table 1).

Table 1. Keywords used for search refinement

Topic	Keywords
Compliance to preparatory fasting	Fasting, meal, diet
Body position	Position, posture, postural changes
Tourniquet application	Tourniquet, venous stasis
Needle gauge	Needle gauge, bore size
Needle type (regular/ butterfly)	Butterfly needle, regular needle
Order of draw	Order of draw, tube order, sample order
Antiseptic swabbing	Disinfectant, alcohol use avoidance, swabbing
Mode of aspiration	Mode of aspiration, vacuum
Discard tube	Discard tube, first tube
Specimen handling	Tube mixing, sample mixing

For each topic searched, suitable papers were extracted by deciding on the bases of the abstract content. Papers were considered suitable only if compliant with all the following requirements: a) were based on an experimental set up aimed to investigate one or more pre-analytical factors related to the procedure of blood drawing, b) reported a quantitative effect (mean change and/or bias) with respect to clinical chemistry, haematology and coagulation tests, c) assessed no other procedure for vein accessioning except venepuncture. Thus, we excluded retrospective studies relying on mathematical models like regression, studies dealing with pre-analytics in metabolomics or biobanking, investigations comparing phlebotomy to other drawing techniques (e.g. saline lock devices or intravenous catheters) as well as those assessing the effect of particular devices or materials (e.g. kind of tube preservatives or infra-red vein finders).

Finally, selected papers were evaluated with respect to the experimental set up, sample size (N), kind of population used for the study (volunteers, donors, inpatients, outpatients), number of individual laboratory tests evaluated, number of factors assessed, testing of data normality, descriptive measures provided (central tendency, dispersion), measure of association between paired observations, methodology used for agreement and bias estimation, clinical significance assessment.

Data were analysed with Microsoft Excel (Microsoft Corporation, USA) spreadsheet and StatsDirect 2.7.2 (StatsDirect Ltd., UK) statistical package, representing relative frequencies as proportions according to the author guidelines (7). The normality of the data was tested by means of Shapiro-Wilk's test. Data dispersion was assessed using a dot-plot and represented by median and interquartile range (IQR) accordingly. The statistical association between qualitative variables was assessed by means of the Fisher's exact test or Fisher-Freeman-Halton test, or by the Spearman's ρ between quantitative continuous variables. In-

stead, the Mann-Whitney U test was used to assess the association between a continuous and a qualitative variable (i.e. the effect of a factor on a median value). The effect size was estimated according to Spearman's ρ for U-test, while the pairwise Cohen's d was used for retrospective estimation of data available in the reviewed studies (8-10). Statistical significance level was set at $P < 0.05$.

Search results

The search provided a total of 136 articles, of which 36 resulted suitable and available for this study according to the established criteria. Three more papers, which resulted potentially suitable, were not available through our local library service, and thus were excluded from this study. Some papers dealt with more than one topic in the same experimental set up, so with respect to each single topic we found:

- 4/36 (0.11) on "compliance to preparatory fasting" (11-14)
- 4/36 (0.11) on "body position" (15-18)
- 6/36 (0.17) on "tourniquet application" (19-24)
- 2/36 (0.06) on "needle gauge" (25,26)
- 3/36 (0.08) on "needle type (regular or butterfly)" (27-29)
- 5/36 (0.14) on "order of draw" (15,19,30-32)
- 2/36 (0.06) on "antiseptic swabbing" (33,34)
- 2/36 (0.06) on "mode of aspiration" (35,36)
- 7/36 (0.19) on "discard tube" (37-43)
- 4/36 (0.11) on "specimen handing (inversion)" (28,44-46)

Thus, it resulted 6/36 (0.17) papers issued by 1996-2000, 5/36 (0.14) by 2001-2005, 7/36 (0.19) by 2006-2010 and 18/36 (0.50) by 2011-2016 (up to March), issued in 17 different journals (Figure 1).

Sample size and study population

With respect to the sample size, 14/36 papers (0.39) had $N \leq 20$, 6/36 (0.17) had $N \leq 30$, and 16/36 (0.44) had $N > 30$. In some studies where $N > 30$, the sample was partitioned into two or more subgroups on which the analysis was repeated independently, so that the actual sample size varied according to the stratification (28,32,33,38,43,45).

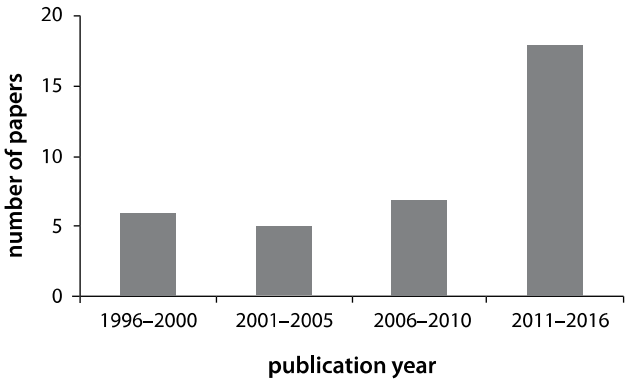

Figure 1. Number of issued papers by publication year.

In only a single case (1/36, 0.03) the authors reported the sample size was chosen basing on a preliminary power analysis, otherwise no information regarding preliminary calculations was given (42).

In 22/36 papers (0.61) the study population was represented by healthy volunteers, in 11/36 (0.30) by outpatients, in 1/36 by inpatients (0.03), in 1/36 by blood donors (0.03) and in 1/36 it was not specified (0.03). Notably, with respect to the median sample size, it was N = 88 (IQR: 54.5 - 220.5) for studies using outpatients and N = 20 (IQR: 17.5 - 30.0) for studies using volunteers, with the difference being statistically significant ($P < 0.001$, effect size $\rho = 0.69$).

Study design and data summarization

All the studies relied on the within-subjects or single-group repeated-measures design. Particularly, 26/36 (0.72) assessed 1 pre-analytical factor, while the remaining 10/36 (0.28) assessed 2 factors (e.g. tourniquet pressure and time). Besides, in 2 papers it was also assessed a third factor which was not related to phlebotomy (sample storage and data transportation respectively) (19,28). In 4/36 papers (0.11) the investigation regarded 1 laboratory test, in 15/36 (0.42) from 2 to 5 tests, in 12/36 (0.33) from 6 to 24, while in 5/36 (0.14) 25 or more tests.

In 28/36 papers (0.78) no normality test was mentioned or reported, while in 5/36 (0.14) the Kolmogorov-Smirnov or D'Agostino-Pearson's test was used (without specifying which kind in one case), and in 3/36 (0.08) the paper was unclear re-

garding whether the test was performed and which one was adopted (15,19,25). Noteworthy, the use of normality test was not associated with the sample size, in that the frequency with which it was used in studies with N ≤ 20 and N > 20 did not differ statistically (P = 0.328).

The paired Student's t-test was the most used statistical test for assessing the effect produced by the pre-analytical factor and it appeared in 17/36 papers (0.47), while the non-parametric equivalent Wilcoxon's paired-ranks test was used in 11/36 (0.31). In this regard, in 2/36 cases (0.06) the authors stated that Student's or Wilcoxon's test was chosen after the result of a normality test (12,46). In 4 papers (0.11) the authors used linear models to analyse their data, which were represented by parametric or non-parametric (Friedman's) 1-way ANOVA, or in 1 single case (0.03) by a linear mixed effect model (LMEM). Except when it was explicitly reported, the choice of a non-parametric instead of a parametric statistical test was made independently from a sample size N ≤ 20 (P = 0.720), as well as a prior execution of a normality test (P = 0.811). Noteworthy, although 29/32 papers (0.91) assessing more than 1 laboratory test used a 2-sample location test (Student's or Wilcoxon's test), just 1/29 (0.04) corrected the α inflation by means of the Bonferroni method (41).

The mean was the central tendency measure most frequently used (26/36 papers, 0.75) to summarize data, and in 5/36 cases (0.14) it was used even when the statistic assessment was achieved by means of a non-parametric test (16,19,34,41,44). With respect to variability, the standard deviation was the measure most frequently used (19/36, 0.53) along with the interquartile range (8/36, 0.22), while just 4/36 (0.11) papers used the 95% confidence interval alongside the mean (28,36,41,42). Just 9/36 papers (0.25) provided the value of correlation between paired data, thereby allowing the retrospective estimation of the observed effect size (21,22,25-27,29,39,40,43). With respect to the appropriateness of summarization, considered as the kind of measure of central tendency and dispersion adopted with a parametric or non-parametric test, it resulted independent from the journal that issued the research (P = 0.676).

Bias and clinical significance

The Bland-Altman analysis was used to estimate bias in 12/36 papers (0.33), followed by the percentage mean difference (8/36, 0.22) which represented the difference between baseline and treatment values divided by the treatment value. Notably, in 12/36 papers (0.33) no bias estimation was reported, while in 3/36 (0.08) cases a Bland-Altman like plot analysis was used although it was not mentioned as such in methods (37,38,43). Lastly, in 2/36 (0.06), along with Bland-Altman analysis, the bias was estimated through Passing-Bablok regression with 95% confidence intervals (21,41).

With respect to agreement between laboratory test results obtained with and without the factor applied, the Passing-Bablok regression was used in 8/36 (0.22) papers, while in 24/36 (0.67) papers no agreement assessment was shown. In 3/36 (0.08) cases it was used the ordinary least-squares regression and in 1 single case (1/36, 0.03) the simple linear correlation (37,39,42,43).

Regarding the clinical significance of the pre-analytical factor, in 16/36 (0.44) papers the authors preferred the direct comparison of the corresponding bias with the value of biological variability reported in databases. Conversely, in 4/36 (0.11) paper it was used a more statistically structured approach based on the total change limit (TCL) or the reference change value (RCV) (13,14,19,41).

Interestingly, in those papers that allowed the retrospective estimation of the effect size (7/36 cases, 0.19), the size of the effect produced on individual laboratory test and the corresponding magnitude of the bias resulted uncorrelated (Spearman's ρ = 0.146, P = 0.142). However, when the clinical significance of bias was used as grouping criterion, the effect size significantly differed in median magnitude (P < 0.001, effect size ρ = 0.59) (Figure 2) (21,22,25-29). Particularly, the median was 0.349 (IQR: 0.228-0.531) and 1.140 (IQR: 0.815-1.700) for clinically significant and non-significant bias respectively.

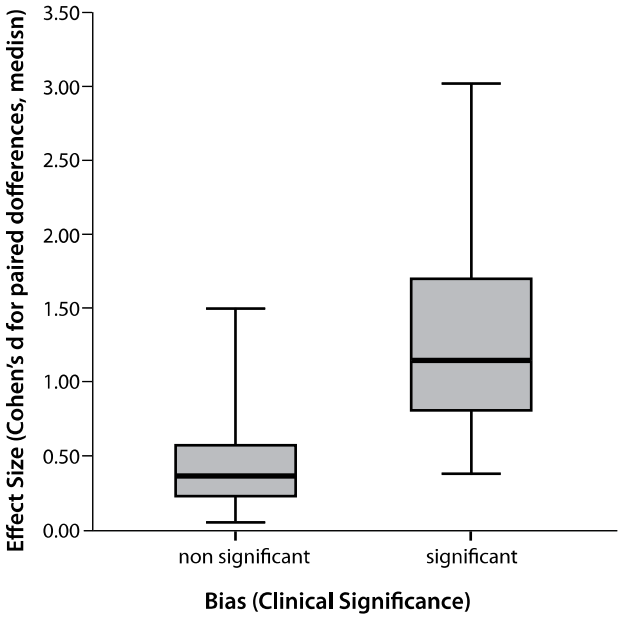

Figure 2. Clinical significance of bias and estimated effect size
The box-plot shows the median and spreading of effect size according to the clinical significance of the associated bias ($P < 0.001$, effect size $\rho = 0.59$).

Summarizing evidence

Pre-analytical investigations of phlebotomy resulted constituting a heterogeneous body of investigations based on the general framework of within-subjects repeated measures design, in which the methodological approach showed a certain variability even among papers issued by the same group of authors.

First, we recognized some methodological inaccuracies that could be considered general issues of research articles, and that usually are addressed at the level of author guidelines by journals. For instance, the choice of the statistical test (parametric or not) was carried out independently from any assessment of the dataset structure (e.g. size, normality, dispersion). In this regard, we found a lack of association between the appropriateness of summarization and the journal issuing the paper, suggesting that this kind of flaws probably depends on a scarce attention paid by authors to that kind of guidelines.

Second, we observed certain specific drawbacks, some of which strictly related to the conceptual and statistical framework that characterized this kind of studies, and that can be resumed as follows:

- the choice of the population study affecting the sample size, with cohorts of healthy volunteers ranging below the size of 30
- the investigations of more than one laboratory test within the same experimental framework without an opportune correction, causing an inflation of the statistical significance
- the lack of calibration (i.e. sample size calculation based on the least significant detectable difference), especially for small sized cohorts
- the use of 2-sample location test (e.g. Student's t-test or Wilcoxon's test) as a "screening approach" to a factor in the study
- the prevalent use of Bland-Altman plot, without a regression analysis of trend of individual differences to assess a proportional effect
- the agreement analysis treated as complementary rather than fundamental for a pre-analytical investigation, and therefore often ignored or sometimes carried out with inappropriate methodologies.

Particularly, regarding the last three points, the pre-analytical factor, the bias and the agreement were usually treated as if there was no relationship between them, leading to use multiple tests (often redundant) that resulted in a fragmented statistical framework. Therefore, many studies had potentially non-homogeneous calibration through different statistical methods, and thus were potentially at risk of delivering some unreliable results. For instance, we noticed that factors showing a non-clinically significant bias were associated with a smaller effect size when assessed by means of 2-sample location. Lastly, we also noticed a certain lack of standardization in operative procedures reported in the various investigations, and a general inhomogeneity regarding how presenting data and delivering results to the reader. Thus, based on these evidences, we have developed a set of recommendations presented in part II of this document, aimed to ensure the adequate quality level

to researches dealing with pre-analytical issues related to phlebotomy.

PART II – RECOMMENDATIONS

Setting up the cohort

The nature of subjects within a cohort should be chosen in order to address a specific diagnostic issue, rather than a generic laboratory concern. Indeed, phlebotomy constitutes the essential connection between the clinics and the laboratory diagnostics, with venepuncture prompted by a precise medical question (4). Therefore, the patient-side perspective should be preferred over the laboratorian-side perspective, even if the investigation concerns a technical aspect of laboratory pre-analytics.

The choice of the population in a phlebotomy study can make the difference when the results are generalized to a different population. For instance, evidences on mechanical factors gathered in healthy subjects may not suit oncologic patients having chronic lymphocytic leukaemia or under tamoxifen treatment, in that they show an abnormal cell fragility (47,48). Conversely, the same cohort might suit the investigation on false positives in laboratory testing of general population due to pre-analytical errors in phlebotomy.

Calibrating the study

A study should be meant to detect the meaningful effect size of a factor, avoiding both excessive (over-powered) or scarce (under-powered) sensitivity (49). Study sensitivity depends on the particular statistical test adopted to assess significance, as well as on the size of the cohort that was chosen to carry out experiments (50-52). As the sample size has the larger impact since can be more easily varied by the researcher, its strict management should be meant for achieving the appropriate study calibration and avoiding unreliable results (see Appendix A) (53).

Invasive procedures naturally tend to rely on a small cohort, and in phlebotomy, some of the experiments even require multiple vein accessioning (e.g. the comparison of butterfly *versus* straight needle). Moreover, some medical conditions can further complicate the enrolment of patients. For instance, it could be easy to adequately size the study when the enrolment concerns subjects under oral anticoagulant therapy with INR between 2 and 3, but the situation could markedly change at higher values of INR (38,43). A practical way of properly sizing a study consists in starting from an expected magnitude of the effect, that for instance could be estimated by means of retrospective calculations using previously available data (54). Then, the required sample size can be achieved inputting the value thereby obtained into stand-alone freely-available dedicated software as well as on-line web tools, choosing the statistical test that is going to be used (55-57).

Beside calibration, a study should also rely on an accurate data validation, achieved assessing the dataset shape. A normality test is useful to show any eventual significant distortion produced by erratic observations, for instance like the ones that can arise due to biological variation (outliers) (58). It should be remarked that skewness markedly affects parametric statistics (Student's t-test), so that the choice between parametric and non-parametric tests should be made carefully and not only basing on the sample size (59-61). Indeed, the choice of the inappropriate statistical test is responsible of a deflation of sensitivity, that is already an issue of small-sized studies (51). Therefore, data validation should be mandatorily carried out as strictly as possible.

Setting the procedures

Procedures used to investigate pre-analytical factors should be standardized, as the reliability of such a study strictly relies on their correct application and execution. Indeed, the lack of standardization could introduce uncontrolled confounding factors that might lead to contradictory findings, as it was shown happening for the "fasting" condition or the venous stasis induction (24,62). Thus, if a referenced protocol is currently unavailable, the author should detail what was performed instead of using general terms or descriptions (e.g. "ve-

nous stasis was induced applying an elastic tourniquet at 5 cm above the site of insertion, inducing an equivalent pressure to 60 mmHg, holding in place for 1 minute after 21G needle insertion" instead of "samples were collected after venous stasis was induced").

With respect to laboratory tests used as part of the experimental procedures, the recommendation concerns the way they should be arranged when the study deals with a panel of multiple analytes. Actually, this implies that the same cohort is independently tested several times (once for each analyte) within the same experimental framework, a fact that rises some concerns on reliability due to the probabilistic nature of the statistical assessment (53). In fact, it causes an inflation of the rate of falsely significant results, requiring an opportune correction like an upward adjustment of the P-value through appropriate statistical procedures (see Appendix B for details) (63,64).

Maximizing the design

The methodological framework of pre-analytical investigations should maximize the reliability and consistency of the achieved information. Thus, the approach based on assessing the effect of factor, the bias and the agreement within the same design as separate entities through distinct statistical methods should be discouraged since inappropriate and unnecessary.

The two major concerns arising from the use of multiple methods are homogeneity of calibration and robustness. Recalling what stated earlier on statistical power, it is virtually impossible to achieve a homogeneous sensitivity for different statistical procedures at the same sample size (50-52). Furthermore, different methods show unlike robustness toward the same shape (i.e. outliers) and variability (i.e. inhomogeneity of variance) of data, that may arise due to an underlying heterogeneity of the cohort (65). Thus, combining these two factors, a research might show redundant tests producing even discordant evidences (see further in this section). Instead, the statistical framework should avoid any ambiguity, maximizing the advantage of within-subjects designs that allow controlling the intra-individual variability increasing the precision of estimates and in turn the study sensitivity (50,52). In this regard, linear models like regression and especially linear mixed-effects model (LMEM) should be preferred.

The LMEM (or multilevel model) is a general case of multiple regression (i.e. a regression with more than one predictor) suitable to handle the contribution of individual variability in the analysis of multiple effects (66,67). It can handle both effects that can be experimentally replicated and have the same size for all tested subjects (namely "fixed", like two different bore sizes or different stasis duration), and effects that lay outside the experimental control and have a certain variability (namely "random", like the homeostatic point of each subject in the study) (66,68). In pre-analytical investigations, the two kinds of effects are always combined, because planned factors are applied to a random set of individuals (28,69). Thereby, LMEM can decompose total variability (i.e. variance) into components, showing the contribution of within-subject, analytical (i.e. method imprecision), and factor effect (bias) separately. For instance, one may plan to investigate the rate of pseudohyperkalemia due to needle bore size, and simultaneously investigating the effect of age (random effect), MCV (random effect) and gender (fixed effect) of the subjects adding the appropriate terms. The LMEM is an observation-centred rather than a factor-centred framework like ANOVA and repeated-measures ANOVA (70,71). Thereby, it has other two points of strength: a) it can handle missing data produced by outliers removal or eventual drop outs, and b) it can account for correlation between observations like the effect of baseline value on response within the same individual. A major (and technically the only one) limitation to the use of LMEM is the methodological complexity, that demands the appropriate level of statistical knowledge to properly set-up the experimental design, transferring data into the statistical frame and interpreting the results (66).

The Passing-Bablok regression is an in-error variable method that relies on a non-parametric estimation of coefficients to gain robustness (72-74). With respect to Deming model that relies on the

least-squares estimates, it is fairly insensitive to outliers, and allows to handle single measurement for each observation pair ignoring the analytical imprecision (75,76). As well as any other regression method it shows the agreement between paired observations, that is the way they scatter around a line with no prevailing effect of one procedure over the other. However, being a linear model of relationship, it decomposes the observed effect into a constant (intercept c) and a proportional (slope b) bias (77,78). What is more it relies on confidence interval for assessing significance, accustoming researchers and readers to give up the use of P-value (79). Limitations of the Passing-Bablok model are that it cannot handle multiple factors and missing data, as well as it necessitates a high correlation between paired observations to hold.

It should be remarked that the use of 2-sample location tests (e.g. t-test and non-parametric equivalents) as a means to assess statistical significance of a factor, alone or beside regression analysis, should be discouraged. Actually, they investigate only systematic difference, that is systematic bias at an agreement analysis, and can be considered reliable just when observations cover a narrow range and no significant trend is supposed to arise (80,81). Therefore, a paper could report a non-significant factor at t-test producing a proportional bias instead, confusing the reader.

Assessing clinical significance

The clinical significance of any procedure should be always assessed within the statistical framework adopted to test the factor, and reported alongside the statistical significance. If the assessment is carried out with linear models (Passing-Bablok regression, LMEM) it is suitable to use the RCV, popularized in laboratory medicine by Fraser, to get the actual threshold of clinical significance (82). If the observed bias is larger than the expected combined effect of analytical and biological variability, then clinical significance is achieved. The RCV can be obtained at different levels of the laboratory assay knowing the corresponding actual imprecision of the analytical method by means of quality control samples (an example applied to regression analysis in shown in Appendix C). Alternatively, the technically equivalent total change limit (TCL) can be used (83). As they both depend on the underlying assumption of statistical normality (same probability of getting an equally large positive or negative variation), they can be reformulated using a robust non parametric model in order to better resemble the structure of data and gain the appropriate sensitivity (84,85). Lastly, the comparison of achieved bias with desirable values obtained from databases is another alternative to the use of statistically derived boundaries, but it does not take into consideration the actual imprecision of the methods used to perform experiments (86).

A concluding remark on statistical methodology regards the recommendation to use the difference plot (better known as Bland-Altman plot) for bias assessment and clinical significance in these studies. The method was devised to estimate at-a-glance, by the scatter plot of individual differences between paired observations, the 95% limits of agreement using the ± 1.96 standard deviation interval around the average bias (87-89). The procedure has the major advantage of computational simplicity and visual immediacy, but in order to emphasize the random component of bias it constrains the modelling of the systematic and proportional components (90). It should be also not considered complementary to regression analysis, also because procedures based on least squares estimate do not return independently distributed residuals while the Bland-Altman plot assumes differences to behave otherwise (91). Therefore, use and interpretation of this kind of plot within a framework based upon linear modelling should be carefully undertaken.

Delivering the evidences

A pre-analytical investigation of phlebotomy should aim to deliver information of practical relevance, and thus it should be meant to reach also non-academic recipients. This makes accessibility a major objective, and the author should take into consideration the impact in the decision-making of the phlebotomist accessing his research. In this

regard, it should be advisable to use P-value beside the confidence interval plus the level of clinical significance, as that was shown to produce the highest rate of correct interpretation of results (92,93).

General recommendations of scientific writing are considered mandatorily applied to these studies (7). However, two special recommendations concern the section reporting the study discussion. First, the authors should take care of emphasizing the supposed mechanisms behind the results, especially with respect their relevance for the actual practice of phlebotomy and the current operative procedures. Second, they should avoid mentioning statistical aspects related to results for not distracting the reader, leaving such aspects to footnotes or appendices to the main text.

Table 2. Recommendations for pre-analytical studies in phlebotomy

- Choose the subjects of your study to address laboratory pre-analytics with respect to a precise diagnostic issue, considering the limitations and pitfall of results when generalized.
- Adopt a standard procedure if available to investigate a pre-analytical factor, or detail it if there is none available.
- Use a statistical framework in order to give consistency and unity to the analysis of data, avoiding multiplication of methods and inhomogeneous calibration (i.e. minimum effect size detectable).
- Assess the clinical significance of bias at different levels, preferably the same of the internal quality control performed on the laboratory tests used for the study.
- Ensure accessibility to results and discussion focusing on mechanisms, relating to current procedures or guidelines, and avoid including any statistical consideration.

Conclusion

In phlebotomy, operative procedures are fundamental for an appropriate patient management and clinical testing reliability, and on their simplicity and effectiveness depends the level of compliance they can reach (94,95). The academic research has a pivotal role in this, and pursuing standardization should be considered part of the consolidation process undertaken by any research field aware of its scope. Actually, this means attaining the unity in the methodological paradigm to achieve effectiveness, with a concise, consistent and efficient production of cumulative knowledge (96). Thus, issuing recommendations (a summarization of which is displayed in Table 2) should be regarded as the first step in such a cultural growth.

In this work, we mainly discussed the statistical methodology, trying to recognize the specific concerns of pre-analytical investigations of phlebotomy. When we completed to review the papers, that body of publications looked like highly heterogeneous with some redundancies within the statistical framework. Thereby, we considered a suitable approach pruning the existing framework inherited from the method comparison studies, basing on the evidences that it was not always properly or sufficiently replicated in all its fundamental parts (97). Maybe, since laying in between clinics and laboratory, phlebotomy has long struggled to gain in scientific literature its own identity and the same reputation as laboratory assays. For instance, it's symptomatic that just two out of the six papers issued by 1996-2000 used a kind of difference plot, and both of them neither mentioned the Bland-Altman eponymous nor cited the original paper (by the way, Scopus showed 3914 citations yet by that time) (19,20). Conversely, in the past years, mostly the last five (see Figure 1), we observed a change in trend and a growing attention payed toward this kind of research. Probably, we owe that to the efforts spent for addressing the cardinal role and the pre-analytical relevance of phlebotomy in modern laboratory medicine, making of it a major concern (1,98).

There are some aspects of this work that should be addressed as possible limitations, and for which we would provide a justification. Actually, we did not structure this work as a systematic review, relying on PubMed MEDLINE alone, and it could be argued that a certain bias of partiality arose. However, we were concerned with the way the scientific information was produced and delivered, and not with its use for generating meta-analytical results. PubMed represented our objective being a comprehensive health information resource that is

preferably queried by academic readers over other databases (99,100). Thus, merging the search results of different sources would have meant deviating from the perspective of largest part of potential readers, introducing a bias of liberality instead.

Second, it could be objected that the pruning was rather an arbitrary choice of the suitable techniques not based on a consensus, which tended to privilege more complicated statistical procedures. Actually, the logic was contrasting the unnecessary multiplication of methods within the statistical framework mostly caused by their customarily use. For instance, we proposed to carefully handle the Bland-Altman plot, reputed a mainstay of the comparative paradigm (97). Interestingly, it should be noticed that the celebrated simplicity was already recognized not a guarantee of appropriateness and homogeneity regarding its use and diffusion (101).

In a future perspective, the critical process initiated issuing these recommendations should culminate in the development of a complete chart dedicated to pre-analytical investigations (and not only strictly concerning phlebotomy) similar to the Standards for Reporting Diagnostic Accuracy's (STARD) chart. Potentially, that would consolidate the contribution of this research field to both laboratory quality and patient safety (102). However, the adherence to strict requirements represents an additional effort in managing a study, that can be experienced as impractical if the peer-reviewing process does not encourage to comply with it and the research quality is not exalted by an increased citation rate (103-105). The experience maturated with the STARD has shown how much all such factors hindered the consolidation of such a new paradigm, despite the wide resonance it had in the scientific literature (106-109). Obviously, there must be correspondence between authors, peer-reviewers and journals to let any new concept reaching acceptance and spreading (110-113).

What outlined above can be nothing but a slow process of growth that demands collective awareness and positive disposition to achieve maturity. Actually, we need to challenge the safe zone of customaries to follow that growth.

Potential conflict of interest

None declared.

APPENDIX A – Calibrating the study for a robust regression analysis of bias

Any regression model requires to be calibrated on the applied range of values, the size of slope change to detect and the variability in observations, and this holds for both parametric and non-parametric methods (72,73,112). However, the latter requires simulation studies, so that for Passing-Bablok method it is suitable to use summarized results in tables available through original papers. To use the aforementioned tables, it is necessary to input the range of any two series of observation (since the method is invariant and the two must be highly linearly correlated), their dispersion (as coefficient of variation, CV%) and the expected minimum slope change (as ratio).

Thus, if a series of observations ranges from c_{min} to c_{max}, then:

$$\textit{Eq.1.1} \quad C_{range\ size} = \frac{C_{max}}{C_{min}}$$

For studies involving a homogeneous population it can be suitably used c = 2, specifying that Passing and Bablok modelled any value $2 \leq c < 4$ as c = 2 and $4 \leq c < 8$ as c = 4 (73). Then, for a 10% proportional bias, the slope change becomes b = 0.9-1.1, so that if CV = 2% holds for both series of observations, a cohort of 30 individuals would be suitable to achieve the customarily 80% sensitivity

at α = 0.05. However, with a CV as high as 5% with all other terms being equal, the same sensitivity would be achieved basing on 90 subjects. It should be remarked that, Passing and Bablok themselves recommended to use always a sample size of 30 at least when applying their model (113).

APPENDIX B - Deflating the level of statistical significance in multiple independent testing

The α level represents the testwise probability of a false positive result, that is obtaining a P-value < α just by chance when the null hypothesis is really true. Therefore, the test wise probability of a true positive result equals 1 - α, that is 0.95 if α = 0.05 as usual. Sometimes, the same experimental framework (same cohort) is used to test a factor on several variables independently (A *vs.* A', B *vs.* B', *etc.*, to not confound with multiple post-hoc comparisons A *vs.* A', A' *vs.* A", A *vs.* A"), which implies replicating C times the hypothesis testing procedure. In plain words, it corresponds to asking whether the population has at least one statistically significant characteristic among those tested with respect to the same factor. In this case, the experimentwise probability of a false positive outcome becomes 1 - (testwise probability of false positive result)^C = 1 - (1 - testwise probability of true positive result)^C that is:

$$\textit{Eq.2.1} \quad \alpha_{experimentwise} = 1 - (1 - \alpha_{testwise})^C$$

Hence, for 2 tests the experimentwise α = 0.10, for 4 α = 0.19, but for 10 test it gets as high as α = 0.40.

The Bonferroni correction for deflating α is achieved through:

$$\textit{Eq.2.2} \quad \alpha_{Bonferroni\ testwise} = \frac{\alpha_{testwise}}{C}$$

Instead, the Šidák correction is achieved through:

$$\textit{Eq.2.3} \quad \alpha_{Šidák\ testwise} = 1 - (1 - \alpha_{testwise})^{1/C}$$

Thus, for 10 independent tests the adjusted testwise α returned by both methods would be α = 0.005, but in case of 20 tests it would get as low as α = 0.0025 for Bonferroni and α = 0.003 for Šidák. Then, substituting for the adjusted testwise α in Eq.2.1, it is possible to obtain the deflated experimentwise α. Hence, applying the method of Šidák, the experimentwise significance for 20 tests would return α = 1 – (1 – 0.003)^{20} = 0.058 instead of the uncorrected α = 1 – (1 – 0.05)^{20}= 0.642).

APPENDIX C – Using RCV for clinical assessment of linear regression bias

If CV_I is the within-subject biological variability and CV_A the method imprecision, then the RCV can be computed according to the formula (82):

$$\textit{Eq.3.1} \quad \mathbf{RCV} = Z \times \sqrt{2 \times (CV_I^2 + CV_A^2)}$$

Where z is a constant for the level of statistical confidence (1.96 for α = 0.05). For a hypothetical analyte X (expressed in arbitrary units, au), the literature reports CV_I = 0.13 (13%), and by the internal quality control it is known that $CV_{A\text{-}LOW}$ = 0.21 (21%) around the 25 au level, and $CV_{A\text{-}HIGH}$ = 0.14 (14%) around the 85 au. Hence, appropriately substituting in Eq.3.1, the RCV at α = 0.05 statistical significance results ± 0.68 (68%) and ± 0.53 (53%) at low and high level respectively.

Through the Passing-Bablok regression, it is shown that an alternative procedure compared to the standard for the collection of blood samples caused both systematic (c = -6.46 au; 95% CI: -9.22 to -3.53 au) and proportional (b = 1.16 au; 95% CI: 1.12 to 1.22 au) bias in the analysis of the analyte X. Therefore the final equation to find out the corresponding biased values in the alternative procedure is:

Eq.3.2 **biased value** = – 6.46 + 1.16 × (*reference value*)

Using Eq.3.2, at low quality control it was found 22.5 au with bias (25 - 22.5) / 25 = 0.10 (10%), and at high quality control it was 92.1 au with bias (85 - 92.1) / 85 = - 0.08 (- 8%). Therefore, although statistically significant, no clinical significance was produced by the alternative collection procedure.

References

1. *Lippi G, Guidi GC, Mattiuzzi C, Plebani M. Preanalytical variability: the dark side of the moon in laboratory testing. Clin Chem Lab Med 2006;44:358-65. https://doi.org/10.1515/CCLM.2006.073.*
2. *Lima-Oliveira G, Guidi GC, Salvagno GL, Montagnana M, Rego FG, Lippi G, et al. Is phlebotomy part of the dark side in the clinical laboratory struggle for quality? Lab Med 2012;43:172-6. https://doi.org/10.1309/LMZ7YARD6ZSDIID.*
3. *Hallworth MJ. The '70% claim': what is the evidence base? Ann Clin Biochem 2011;48:487-8. https://doi.org/10.1258/acb.2011.011177.*
4. *Ialongo C, Bernardini S. Phlebotomy, a bridge between laboratory and patient. Biochem Med (Zagreb) 2016;26:17-33. https://doi.org/10.11613/BM.2016.002.*
5. *Simundic AM. Preanalytical phase - an updated review of the current evidence. Biochem Med (Zagreb) 2014;24:6. https://doi.org/10.11613/BM.2014.001.*
6. *Lippi G, Banfi G, Church S, Cornes M, De Carli G, Grankvist K, et al. Preanalytical quality improvement. In pursuit of harmony, on behalf of European Federation for Clinical Chemistry and Laboratory Medicine (EFLM) Working group for Preanalytical Phase (WG-PRE). Clin Chem Lab Med 2015;53:357-70. https://doi.org/10.1515/cclm-2014-1051.*
7. *Simundic AM. Practical recommendations for statistical analysis and data presentation in Biochemia Medica journal. Biochem Med (Zagreb) 2012;22:15-23. https://doi.org/10.11613/BM.2012.003.*
8. *Fritz CO, Morris PE, Richler JJ. Effect size estimates: current use, calculations, and interpretation. J Exp Psychol Gen 2012;141:2-18. https://doi.org/10.1037/a0024338.*
9. *Gravetter FJ, Wallnau LB, eds. Statistics for the behavioral sciences. 9th ed. Belmont, CA: Wadsworth Cengage Learning; 2013. p. 767.*
10. *Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863. https://doi.org/10.3389/fpsyg.2013.00863.*
11. *Lippi G, Lima-Oliveira G, Salvagno GL, Montagnana M, Gelati M, Picheth G, et al. Influence of a light meal on routine haematological tests. Blood Transfus 2010;8:94-9.*
12. *Lima-Oliveira G, Salvagno GL, Lippi G, Gelati M, Montagnana M, Danese E, et al. Influence of a regular, standardized meal on clinical chemistry analytes. Ann Lab Med 2012;32:250-6. https://doi.org/10.3343/alm.2012.32.4.250.*
13. *Lima-Oliveira G, Salvagno GL, Lippi G, Danese E, Gelati M, Montagnana M, et al. Could light meal jeopardize laboratory coagulation tests? Biochem Med (Zagreb) 2014;24:343-9. https://doi.org/10.11613/BM.2014.036.*
14. *Plumelle D, Lombard E, Nicolay A, Portugal H. Influence of diet and sample collection time on 77 laboratory tests on healthy adults. Clin Biochem 2014;47:31-7. https://doi.org/10.1016/j.clinbiochem.2013.11.002.*
15. *Sulaiman RA, Cornes MP, Whitehead SJ, Othonos N, Ford C, Gama R. Effect of order of draw of blood samples during phlebotomy on routine biochemistry results. J Clin Pathol 2011;64:1019-20. https://doi.org/10.1136/jclinpath-2011-200206.*
16. *Lippi G, Salvagno GL, Lima-Oliveira G, Montagnana M, Danese E, Guidi GC. Circulating cardiac troponin T is not influenced by postural changes during venous blood collection. Int J Cardiol 2014;177:1076-7. https://doi.org/10.1016/j.ijcard.2014.10.018.*
17. *Lippi G, Aloe R, Avanzini P, Banfi G. Measurement of iron in serum and EDTA plasma for screening of blood transfusion in sports. Drug Test Anal 2015;7:253-4. https://doi.org/10.1002/dta.1696.*
18. *Lippi G, Salvagno GL, Lima-Oliveira G, Danese E, Favaloro EJ, Guidi GC. Influence of posture on routine hemostasis testing. Blood Coagul Fibrinolysis 2015;26:716-9. https://doi.org/10.1097/MBC.0000000000000326.*
19. *Rosenson RS, Staffileno BA, Tangney CC. Effects of tourniquet technique, order of draw, and sample storage on plasma fibrinogen. Clin Chem 1998;44:688-90.*
20. *Ritchie JL, Crawford VL, McNulty M, Alexander HD, Stout RW. Effect of tourniquet pressure and intra-individual variability on plasma fibrinogen, platelet P-selectin and monocyte tissue factor. Clin Lab Haematol 2000;22:369-72. https://doi.org/10.1046/j.1365-2257.2000.00337.x.*
21. *Lippi G, Salvagno GL, Montagnana M, Guidi GC. Short-term venous stasis influences routine coagulation testing. Blood Coagul Fibrinolysis 2005;16:453-8. https://doi.org/10.1097/01.mbc.0000178828.59866.03.*
22. *Lippi G, Salvagno GL, Montagnana M, Franchini M, Guidi GC. Venous stasis and routine hematologic testing. Clin Lab*

Haematol 2006;28:332-7. https://doi.org/10.1111/j.1365-2257.2006.00818.x.

23. *Cengiz M, Ulker P, Meiselman HJ, Baskurt OK. Influence of tourniquet application on venous blood sampling for serum chemistry, hematological parameters, leukocyte activation and erythrocyte mechanical properties. Clin Chem Lab Med 2009;47:769-76. https://doi.org/10.1515/CCLM.2009.157.*
24. *Lippi G, Salvagno GL, Montagnana M, Brocco G, Guidi GC. Influence of short-term venous stasis on clinical chemistry testing. Clin Chem Lab Med 2005;43:869-75. https://doi.org/10.1515/CCLM.2005.146.*
25. *Lippi G, Salvagno GL, Montagnana M, Brocco G, Cesare Guidi G. Influence of the needle bore size used for collecting venous blood samples on routine clinical chemistry testing. Clin Chem Lab Med 2006;44:1009-14. https://doi.org/10.1515/CCLM.2006.172.*
26. *Lippi G, Salvagno GL, Montagnana M, Poli G, Guidi GC. Influence of the needle bore size on platelet count and routine coagulation testing. Blood Coagul Fibrinolysis 2006;17:557-61. https://doi.org/10.1097/01.mbc.0000245300.10387.ca.*
27. *Lippi G, Salvagno GL, Guidi GC. No influence of a butterfly device on routine coagulation assays and D-dimer measurement. J Thromb Haemost 2005;3:389-91. https://doi.org/10.1111/j.1538-7836.2005.01163.x.*
28. *Sylte MS, Wentzel-Larsen T, Bolann BJ. Random variation and systematic error caused by various preanalytical variables, estimated by linear mixed-effects models. Clin Chim Acta 2013;415:196-201. https://doi.org/10.1016/j.cca.2012.10.045.*
29. *Lippi G, Salvagno GL, Brocco G, Guidi GC. Preanalytical variability in laboratory testing: influence of the blood drawing technique. Clin Chem Lab Med 2005;43:319-25. https://doi.org/10.1515/CCLM.2005.055.*
30. *Majid A, Heaney DC, Padmanabhan N, Spooner R. The order of draw of blood specimens into additive containing tubes not affect potassium and calcium measurements. J Clin Pathol 1996;49:1019-20. https://doi.org/10.1136/jcp.49.12.1019.*
31. *Salvagno G, Lima-Oliveira G, Brocco G, Danese E, Guidi GC, Lippi G. The order of draw: myth or science? Clin Chem Lab Med 2013;51:2281-5. https://doi.org/10.1515/cclm-2013-0412.*
32. *Indevuyst C, Schuermans W, Bailleul E, Meeus P. The order of draw: much ado about nothing? Int J Lab Hematol 2015;37:50-5. https://doi.org/10.1111/ijlh.12230.*
33. *Salvagno GL, Danese E, Lima-Oliveira G, Guidi GC, Lippi G. Avoidance to wipe alcohol before venipuncture is not a source of spurious hemolysis. Biochem Med (Zagreb) 2013;23:201-5. https://doi.org/10.11613/BM.2013.023.*
34. *Sarmah D, Sharma B, Sharma D, Mathew S. Alcohol used as disinfectant before venipuncture does not lead to sample haemolysis or sample dilution. J Clin Diagn Res 2016;10:BC16-8. https://doi.org/10.7860/jcdr/2016/15967.7245.*
35. *Lippi G, Avanzini P, Musa R, Sandei F, Aloe R, Cervellin G. Evaluation of sample hemolysis in blood collected by S-Monovette using vacuum or aspiration mode. Biochem Med (Zagreb) 2013;23:64-9. https://doi.org/10.11613/BM.2013.008.*
36. *Lippi G, Ippolito L, Zobbi V, Sandei F, Favaloro EJ. Sample collection and platelet function testing: influence of vacuum or aspiration principle on PFA-100 test results. Blood Coagul Fibrinolysis 2013;24:666-9. https://doi.org/10.1097/MBC.0b013e32835fada7.*
37. *Yawn BP, Loge C, Dale J. Prothrombin time: one tube or two. Am J Clin Pathol 1996;105:794-7. https://doi.org/10.1093/ajcp/105.6.794.*
38. *Brigden ML, Graydon C, McLeod B, Lesperance M. Prothrombin time determination. The lack of need for a discard tube and 24-hour stability. Am J Clin Pathol 1997;108:422-6. https://doi.org/10.1093/ajcp/108.4.422.*
39. *Gottfried EL, Adachi MM. Prothrombin time and activated partial thromboplastin time can be performed on the first tube. Am J Clin Pathol 1997;107:681-3. https://doi.org/10.1093/ajcp/107.6.681.*
40. *Bamberg R, Cottle JN, Williams JC. Effect of drawing a discard tube on PT and APTT results in healthy adults. Clin Lab Sci 2003;16:16-9.*
41. *Raijmakers MT, Menting CH, Vader HL, van der Graaf F. Collection of blood specimens by venipuncture for plasma-based coagulation assays: necessity of a discard tube. Am J Clin Pathol 2010;133:331-5. https://doi.org/10.1309/AJCP9ATB0AXPFJCC.*
42. *Masih M, Kakkar N. Routine coagulation testing: do we need a discard tube? Indian J Hematol Blood Transfus 2014;30:347-50. https://doi.org/10.1007/s12288-013-0285-9.*
43. *Tekkeşin N, Esen OB, Kilinc C, Eviyaoglu O. Discard first tube for coagulation testing. Blood Coagul Fibrinolysis 2012;23:299-303. https://doi.org/10.1097/MBC.0b013e328351ebbf.*
44. *Lippi G, Salvagno GL, Montagnana M, Guidi GC. Influence of primary sample mixing on routine coagulation testing. Blood Coagul Fibrinolysis 2007;18:709-11. https://doi.org/10.1097/MBC.0b013e32828621a0.*
45. *Parenmark A, Landberg E. To mix or not to mix venous blood samples collected in vacuum tubes? Clin Chem Lab Med 2011;49:2061-3. https://doi.org/10.1515/CCLM.2011.705.*
46. *Lima-Oliveira G, Lippi G, Salvagno GL, Montagnana M, Gelati M, Volanski W, et al. Effects of vigorous mixing of blood vacuum tubes on laboratory test results. Clin Biochem 2013;46:250-4. https://doi.org/10.1016/j.clinbiochem.2012.10.033.*
47. *Riggins RB, Thomas KS, Ta HQ, Wen J, Davis RJ, Schuh NR, et al. Physical and functional interactions between Cas and c-Src induce tamoxifen resistance of breast cancer cells through pathways involving epidermal growth factor receptor and signal transducer and activator of transcription 5b. Cancer Res 2006;66:7007-15. https://doi.org/10.1158/0008-5472.CAN-05-3952.*
48. *Rifkin SI. Pseudohyperkalemia in patients with chronic lymphocytic leukemia. Int J Nephrol 2011;2011:759749. https://doi.org/10.4061/2011/759749.*
49. *Biau DJ, Kerneis S, Porcher R. Statistics in brief: the importance of sample size in the planning and interpretation of medical research. Clin Orthop Relat Res 2008;466:2282-8. https://doi.org/10.1007/s11999-008-0346-9.*
50. *McHugh ML. Power analysis in research. Biochem Med (Zagreb) 2008;18:263-74. https://doi.org/10.11613/BM.2008.024.*
51. *Bridge PD, Sawilowsky SS. Increasing physicians' awareness of the impact of statistics on research outcomes:*

comparative power of the t-test and and Wilcoxon Rank-Sum test in small samples applied research. J Clin Epidemiol 1999;52:229-35. https://doi.org/10.1016/S0895-4356(98)00168-1.

52. Dell RB, Holleran S, Ramakrishnan R. Sample size determination. ILAR J 2002;43:207-13. https://doi.org/10.1093/ilar.43.4.207.
53. Koretz RL. Is statistical significance always significant? Nutr Clin Pract 2005;20:303-7. https://doi.org/10.1177/0115426505020003303.
54. Ialongo C. Understanding the effect size and its measures. Biochem Med (Zagreb) 2016;26:150-63. https://doi.org/10.11613/BM.2016.015.
55. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175-91. https://doi.org/10.3758/BF03193146.
56. Dupont WD, Plummer WD, Jr. Power and sample size calculations. A review and computer program. Control Clin Trials 1990;11:116-28. https://doi.org/10.1016/0197-2456(90)90005-M.
57. Meysamie A, Taee F, Mohammadi-Vajari M, Yoosefi-Khanghah S, Emamzadeh-Fard S, Abbassi M. Sample size calculation on web, can we rely on the results? Journal of Medical Statistics and Informatics 2014.
58. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab 2012;10:486-9. https://doi.org/10.5812/ijem.3505.
59. de Winter JCF. Using the Student's t-test with extremely small sample sizes. Pract Assess Res Eval 2013;18:1-12.
60. Tanizaki H. Robustness and power of parametric, nonparametric, robustified and adaptive tests—The multi-sample location problem. J Appl Stat 1997;24:603-32. https://doi.org/10.1080/02664769723576.
61. Zimmerman DW, Zumbo BD. The effect of outliers on the relative power of parametric and nonparametric statistical tests. Percept Mot Skills 1990;71:339-49. https://doi.org/10.2466/pms.1990.71.1.339.
62. Nybo M, Grinsted P, Jorgensen PE. Blood sampling: is fasting properly defined? Clin Chem 2005;51:1563-4. https://doi.org/10.1373/clinchem.2005.051789.
63. Bender R, Lange S. Adjusting for multiple testing--when and how? J Clin Epidemiol 2001;54:343-9. https://doi.org/10.1016/S0895-4356(00)00314-0.
64. Feise RJ. Do multiple outcome measures require p-value adjustment? BMC Med Res Methodol 2002;2:1-4. https://doi.org/10.1186/1471-2288-2-8.
65. Armitage P, Berry G, Matthews JNS, eds. Statistical methods in medical research. 4th ed. Malden, MA: Blackwell Science;2001. p. 817.
66. Oberg AL, Mahoney DW. Linear mixed effects models. Methods Mol Biol 2007;404:213-34. https://doi.org/10.1007/978-1-59745-530-5_11.
67. Hayes F. A primer on multilevel modeling. Hum Comm Res 2006;32:385-410. https://doi.org/10.1111/j.1468-2958.2006.00281.x.
68. Ibrahim JG, Zhu H, Garcia RI, Guo R. Fixed and random effects selection in mixed effects models. Biometrics 2011;67:495-503. https://doi.org/10.1111/j.1541-0420.2010.01463.x.
69. Sylte MS, Wentzel-Larsen T, Bolann BJ. Estimation of the minimal preanalytical uncertainty for 15 clinical chemistry serum analytes. Clin Chem 2010;56:1329-35. https://doi.org/10.1373/clinchem.2010.146050.
70. Ma Y, Mazumdar M, Memtsoudis SG. Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. Reg Anesth Pain Med 2012;37:99-105. https://doi.org/10.1097/AAP.0b013e31823ebc74.
71. Little RJ, Raghunathan T. On summary measures analysis of the linear mixed effects model for repeated measures when data are not missing completely at random. Stat Med 1999;18:2465-78. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2465::AID-SIM269>3.0.CO;2-2.
72. Passing H, Bablok. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem 1983;21:709-20
73. Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part II. J Clin Chem Clin Biochem 1984;22:431-45. https://doi.org/10.1515/cclm.1984.22.6.431.
74. Bilic-Zulle L. Comparison of methods: Passing and Bablok regression. Biochem Med (Zagreb) 2011;21:49-52. https://doi.org/10.11613/BM.2011.010.
75. Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem 1979;25:432-8.
76. Linnet K. Performance of Deming regression analysis in case of misspecified analytical error ratio in method comparison studies. Clin Chem 1998;44:1024-31.
77. Magari RT. Bias estimation in method comparison studies. J Biopharm Stat 2004;14:881-92. https://doi.org/10.1081/BIP-200035450.
78. Magari RT. Evaluating agreement between two analytical methods in clinical chemistry. Clin Chem Lab Med 2000;38:1021-5. https://doi.org/10.1515/CCLM.2000.151.
79. Ranstam J. Why the P-value culture is bad and confidence intervals a better alternative. Osteoarthritis Cartilage 2012;20:805-8. https://doi.org/10.1016/j.joca.2012.04.001.
80. Linnet K. Limitations of the paired t-test for evaluation of method comparison data. Clin Chem 1999;45:314-5.
81. Westgard JO. Use and interpretation of common statistical tests in method comparison studies. Clin Chem 2008;54:612. https://doi.org/10.1373/clinchem.2007.094060.
82. Fraser CG. Reference change values. Clin Chem Lab Med 2011;50:807-12. https://doi.org/10.1515/cclm.2011.733.
83. Oddoze C, Lombard E, Portugal H. Stability study of 81 analytes in human whole blood, in serum and in plasma. Clin Biochem 2012;45:464-9. https://doi.org/10.1016/j.clinbiochem.2012.01.012.
84. Lund F, Petersen PH, Fraser CG, Soletormos G. Different percentages of false-positive results obtained using five methods for the calculation of reference change values based on simulated normal and ln-normal distri-

butions of data. Ann Clin Biochem 2016. https://doi.org/10.1177/0004563216643729.
85. Roraas T, Stove B, Petersen PH, Sandberg S. Biological variation: the effect of different distributions on estimated within-person variation and reference change values. Clin Chem 2016;62:725-36. https://doi.org/10.1373/clinchem.2015.252296.
86. Ricós C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez CV, et al. Current databases on biological variation: pros, cons and progress. Scand J Clin Lab Invest 1999;59:491-500. https://doi.org/10.1080/00365519950185229.
87. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10. https://doi.org/10.1016/S0140-6736(86)90837-8.
88. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60. https://doi.org/10.1191/096228099673819272.
89. Eksborg S. Evaluation of method-comparison data. Clin Chem 1981;27:1311-2.
90. Dunn G, Roberts C. Modelling method comparison data. Stat Methods Med Res 1999;8:161-79. https://doi.org/10.1191/096228099668524590.
91. Rauch G, Geistanger A, Timm J. A new outlier identification test for method comparison studies based on robust regression. J Biopharm Stat 2011;21:151-69. https://doi.org/10.1080/10543401003650275.
92. Shakespeare TP, Gebski V, Tang J, Lim K, Lu JJ, Zhang X, et al. Influence of the way results are presented on research interpretation and medical decision making: the PRIMER collaboration randomized studies. Med Decis Making 2008;28:127-37. https://doi.org/10.1177/0272989X07309640.
93. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. Aust Crit Care 2010;23:93-7. https://doi.org/10.1016/j.aucc.2010.03.001.
94. Simundic AM, Cornes M, Grankvist K, Lippi G, Nybo M, Kovalevskaya S, et al. Survey of national guidelines, education and training on phlebotomy in 28 European countries: an original report by the European Federation of Clinical Chemistry and Laboratory Medicine (EFLM) working group for the preanalytical phase (WG-PA). Clin Chem Lab Med 2013;51:1585-93. https://doi.org/10.1515/cclm-2013-0283.
95. Lippi G, Becan-McBride K, Behulova D, Bowen RA, Church S, Delanghe J, et al. Preanalytical quality improvement: in quality we trust. Clin Chem Lab Med 2013;51:229-41. https://doi.org/10.1515/cclm-2012-0597.
96. Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, et al. Reducing waste from incomplete or unusable reports of biomedical research. Lancet 2014;383:267-76. https://doi.org/10.1016/S0140-6736(13)62228-X.
97. Hanneman SK. Design, analysis, and interpretation of method-comparison studies. AACN Adv Crit Care 2008;19:223-34.
98. Lippi G, Salvagno GL, Montagnana M, Franchini M, Guidi GC. Phlebotomy issues and quality improvement in results of laboratory testing. Clin Lab 2006;52:217-30.
99. De Groote SL, Shultz M, Blecic DD. Information-seeking behavior and the use of online resources: a snapshot of current health sciences faculty. J Med Libr Assoc 2014;102:169-76. https://doi.org/10.3163/1536-5050.102.3.006.
100. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J 2008;22:338-42. https://doi.org/10.1096/fj.07-9492LSF.
101. Dewitte K, Fierens C, Stockl D, Thienpont LM. Application of the Bland-Altman plot for interpretation of method-comparison studies: a critical investigation of its practice. Clin Chem 2002;48:799-801.
102. McQueen MJ. The STARD initiative: a possible link to diagnostic accuracy and reduction in medical error. Ann Clin Biochem 2003;40:307-8. https://doi.org/10.1258/000456303766476940.
103. Hirst A, Altman DG. Are peer reviewers encouraged to use reporting guidelines? A survey of 116 health research journals. PLoS One 2012;7:e35621. https://doi.org/10.1371/journal.pone.0035621.
104. Dilauro M, McInnes MD, Korevaar DA, van der Pol CB, Petrcich W, Walther S, et al. Is There an Association between STARD Statement Adherence and Citation Rate? Radiology 2016;280:62-7. https://doi.org/10.1148/radiol.2016151384.
105. Smidt N, Overbeke J, de Vet H, Bossuyt P. Endorsement of the STARD Statement by biomedical journals: survey of instructions for authors. Clin Chem 2007;53:1983-5. https://doi.org/10.1373/clinchem.2007.090167.
106. Smidt N, Rutjes AW, van der Windt DA, Ostelo RW, Bossuyt PM, Reitsma JB, et al. The quality of diagnostic accuracy studies since the STARD statement: has it improved? Neurology 2006;67:792-7. https://doi.org/10.1212/01.wnl.0000238386.41398.30.
107. Korevaar DA, van Enst WA, Spijker R, Bossuyt PM, Hooft L. Reporting quality of diagnostic accuracy studies: a systematic review and meta-analysis of investigations on adherence to STARD. Evid Based Med 2014;19:47-54. https://doi.org/10.1136/eb-2013-101637.
108. Korevaar DA, Wang J, van Enst WA, Leeflang MM, Hooft L, Smidt N, et al. Reporting diagnostic accuracy studies: some improvements after 10 years of STARD. Radiology 2015;274:781-9. https://doi.org/10.1148/radiol.14141160.
109. Johansen M, Thomsen SF. Guidelines for reporting medical research: A critical appraisal. Int Sch Res Notices 2016;2016:1346026. https://doi.org/10.1155/2016/1346026.
110. Morton JP. Reviewing scientific manuscripts: how much statistical knowledge should a reviewer really know? Adv Physiol Educ 2009;33:7-9. https://doi.org/10.1152/advan.90207.2008.
111. Erb HN. Changing expectations: Do journals drive methodological changes? Should they? Prev Vet Med 2010;97:165-74. https://doi.org/10.1016/j.prevetmed.2010.09.011.
112. Linnet K. Necessary sample size for method comparison studies based on regression analysis. Clin Chem 1999;45:882-94.
113. Bablok W, Passing H. Application of statistical procedures in analytical instrument testing. J Automat Chem 1985;7:74-9. https://doi.org/10.1155/S1463924685000177.