DOI: 10.4081/jsas.2015.6410

STATISTICAL ERRORS IN MEDICINE

Marco B.L. Rocchi

Department of Biomolecular Sciences, Service of Biostatistics, "Carlo Bo" University, Urbino, Italy

Abstract. In this paper we deal with the statistical errors in medicine, analyzing the way they should hide in several phases of experimental process; particularly, we underlined relevant errors in which the researchers can stumble in every steps of the experimental research.

Key words: Statistical errors; medicine.

INTRODUCTION

"Statistics is the mathematics for experimental sciences"; this sentence, often quoted, underlines the relevance of the statistical thought, although it probably undervalues the deductive basis of math, compared to the inductive basis of statistics.

Anywhere, statistical error in medicine should hide in several phases of experimental process; in this paper we will consider the errors in which the researchers can stumble in experimental research; the examples will be drawn primarily from Phase 3 clinical trials [1].

Probably, most relevant statistical errors should be avoided if a statistician is involved in the first of these two steps; regarding this point, sir Ronald A. Fisher declared: "To consult a statistician after a project is finished, is often merely to ask him to conduct a post-mortem examination". Thus, statisticians not only make computations, but they chiefly plan the study design and actively participate in the development of the study protocol [2].

In fact, the most important statistical errors occur in planning stage of an experiment and, *sensu lato* in writing the protocol. In this document, fundamental for both the reliability and the validity of an experimental study, mistakes can hide in many different places; we will analyzed all them, generally following the points of an experimental protocol, as they were described and classified by Pocock [3].

Rationale and general objective of the study

It is important that the rationale (*i.e.* the formulation of the reasons which led to the experiment), as well as the general objective of the study were adequately described and specified.

Although this point seems not involving any statistical issues, nevertheless it represents the basis to correctly define the following protocol point (*i.e.* the definition of the specific objectives).

<u>Relevant errors</u>: to provide a not up-to-date bibliography; to not specify the rationale.

Specific objectives of the study

Once the rationale and the overall goal of the research has been clarified, it is necessary to identify the hypothesis to assess, a process involving the strict definition of both a main objective (which will be the one which will be used to calculate the sample size) and few secondary objectives (related to the main objective, limited in number and defined *a priori*).

The definition of the main objective is necessary in order to establish clearly and in advance which are the basis on which it will be determined the therapeutic efficacy of the treatment, so to avoid methodologically negative phenomena such as post-hoc analysis and data dredging.

We define post-hoc analyses those that are carried out after viewing the data, with a high risk of distortion of results (in particular a high risk of committing false positive errors).

Data dredging (sometimes also defined as fishing expedition) is the phenomenon in which the statistical analysis to be carried out are not fixed *a priori* (*i.e.* in the protocol), in search of significant results (also in this case, with a high risk of committing false positive errors) [4].

Correspondence to:
Marco B.L. Rocchi
Department of Biomolecular Sciences, Service of Biostatistics, Urbino University
Piazza Rinascimento 6, 61029 Urbino (PU), Italy
Tel.: +39.722.303304
E-mail: marco.rocchi@uniurb.it

These remarks, obviously, don't imply that post-hoc statistical analyses could not be executed, but that they only should have an exploratory interpretation; they can be a good starting point for new research, to be planned with new protocols.

Relevant errors: to specify more than one main objective; to indicate too many secondary objectives; to commit "data dredging"; to perform "post-hoc" analyses.

Criteria for patient selection

Establishing the criteria for selection of patients is not simply indicate the characteristics that patients should have to be enrolled in the clinical trial, but - most importantly - means to establish the population of future patients who potentially will benefit from treatment.

From a statistical point of view, this coincides with the idea of inference from a representative sample of a population to the same population from which the sample was extracted. The sample is a population in miniature and, in order to allow to draw conclusions on the population from which it is extracted, it shall be representative of that population.

This consideration may seem trivial and obvious, but the experience of many experiments warns us from this hasty judgment.

Regarding the selection criteria of the patients, we can choose between two opposing strategies (possibly opting for a reasonable compromise between the two) each of which has, of course, advantages and disadvantages; see for example [5].

In the first case, we can use very restrictive criteria: this strategy has the advantage of a more accurate comparison between the two treatments, because some variables, potentially confounding, will be kept under control. Moreover, following this strategy, the results will be moderately affected by the variability of the population. On the opposite hand, the disadvantages are represented by the increase in costs (time, money and human resources) for the recruitment of patients, but above all by the limited generalizability of results (because a sample selected with very stringent criteria corresponds to a limited population on which to infer).

Relevant errors: to indicate either too large or too restrictive criteria for the selection of patients; to infer towards a population different the population from which the sample was extracted.

Methods of assessing the patient's response

The principle which is the basis of this point is that all the procedures of evaluation of the patient should be determined and standardized in the protocol, for both the baseline evaluation and the outcomes.

The outcomes are represented by one or a few variables, which should be consistent with the objectives of the trial [6, 7].

In the choice of endpoints the use of surrogate endpoints (*i.e.* a criterion of evaluation of the patient who does not represent the real benefit of the patient, but - in fact - a surrogate) should be avoided, because a theoretical benefit may sometimes not reflect a real clinical benefit; for example, the decrease of cholesterol level does not necessarily correspond to a lower risk of heart attack.

Another important point of these aspects is the need to use assessment tools (*e.g.* clinimetrical tests) which have been validated according to the criteria of reliability, validity, responsiveness and definition of the least clinically relevant difference.

Relevant errors: to indicate outcomes not consistent with the objectives; to indicate surrogate endpoints; to use not validated clinimetrical tools.

Experimental design

We must therefore emphasize that the Phase III studies, in ideal conditions, should be:

- *Controlled*: that is, it must provide for the presence of a control group which can, depending on the experiments , assume placebo or a standard drug for the disease that is intended to cure;
- *Randomized*: that is, the allocation to one of the treatment arms must be based on criteria of randomness;
- *Blinded*: that is, as far as possible, both patients and experimenters must be unaware of the treatment that each patient actually take (*i.e*, if they assume experimental drug, the standard drug, or placebo).

Despite these ideal conditions are very clear, in practice different aspects of the experimental design will have nuanced connotations.

Regarding the experimental design, a Phase III study could be planned according either to a between patients or a within patients design [8, 9].

In "between patients design" the experiment is usually done by comparing two groups of patients, one subjected to the experimental treatment (experimental arm), and the other assuming a standard drug or a placebo (control arm).

It is however also possible to provide more than two experimental arms, for example in the case that there are two separate control groups (each of which assumes a different standard drug for the disease to be treated), or in the case that you want to experience two new therapeutic approaches comparing them with a control group.

In "within patients design" a single group of patients, assuming first the one and then the other treatment, is considered [10, 11]. Randomization, in this case, determines the order of intake of the two treatments.

The main advantage of these experiments is related to the fact that each patient is, in some way, control of himself; this implies that there will be fewer confounding variables to limit the interpretations of the results, because all the variables pertaining to the characteristics of the patient will remain substantially unchanged in the two stages. Ultimately, this will result in a more accurate comparison between the treatments.

On the other hand, being dual participation request from the same patient in the study, this technique can only be achieved in the case of stable disease. In order to prove the stability of the pathology, it is usually foreseen before the start of the experiment a run-in of the patients, during which all assume only placebo.

Moreover, between the first and the second treatment, the patients have a wash-out period, during which they assume no treatment to avoid phenomena of carry-over, (*i.e.* residual effects of the first treatment during the period of second treatment).

We will neglect, in this paper, other less frequent designs, such as factorial e n-of-1 plans.

Relevant errors: to choose not suitable experimental design.

Randomization

Randomization is one of the fundamental aspects for the success of an experiment, and it represents one of the largest commitments for statisticians involved [12, 13].

Randomization is the random assignment of a patient enrolled in a trial to one of the experimental arms. This is one of the focal points of clinical trials because this procedure allows to avoid any form of selection in the allocation of patients, which, if happened, could have biasing effects, with serious repercussions the reliability of the outcome of the trial. Randomization ensures a certain degree of homogeneity in the two groups of patients.

In this paper we only consider the most relevant and used techniques of randomization.

Simple (or complete) randomization: it ideally corresponds to a coin toss, the result of which assigned the patients to arm A or arm B. Of course, from the operational point of view, the process of launching the coin is replaced by the random number generation (procedure now also achievable with the most common spreadsheets): for example, if the value is between 0 and 0.5, the patient is assigned to arm A, while between 0.5 and 1 is allocated to arm B .

The main advantage of this method is that each patient has the same probability of being assigned to arm A or arm B. In other words, there is no other method that allocates patients with as much randomness.

As for the disadvantages, however, this method presents a risk - usually however modest - to obtain unbalanced allocations (meaning by this term a different number of the two groups). Of course, due to the law of large numbers, the risk of imbalance becomes small when the sample size becomes large.

Block permutation randomization: this method, one of the most used in scientific literature, it has been proposed to avoid the risks of unbalance that the simple randomization presents. The idea is to establish some blocks characterized by an even number of patients, for example 4. Of these 4 patients, 2 will have to be assigned to the group A and 2 to the group B. It is clear that there are different orders (technically we define them as permutations) with two elements assigned to group A and two elements assigned to group B. Therefore it will be necessary to consider all the possible permutations of 4 patients. In the case of blocks of 4, the possible permutations will be the following 6:

AABB, ABAB, BBAA, BABA, ABBA, BAAB

In this case, to each random number (or range of random numbers) it will not be associated to a single patient, but to a block of 4 patients.

It is clear that the number of patients for each block may be also larger than 4, but always equal in number to ensure the balance. In fact, at the end of each block, this technique ensures the balance between the groups A and B.

A modest limit is represented by the predictability of the allocation of the last patient of each block (for example, if the first three patients are ABB, the fourth can only be A); in extreme cases, the predictability may also cover half of the patients of each block (for example, is the case of the blocks AABB and BBAA).

Stratified randomization: stratification is a procedure to be applied in conjunction with some techniques of randomization (simple, randomized block, and so on) in order to obtain a homogeneity between groups A and B with respect o the prognostic variables, i.e. variables that potentially could influence the therapeutic effect of the drug or at least the natural course of the disease [14].

Once you have identified prognostic factors with respect to which stratify, the procedure consists of building separate randomization lists for each level (or combination of levels) of prognostic variables.

For example, let's consider the following situation: we know the prognostic role of variables gender (male *vs* female) and age (<50 *vs* > 50). So, there are 4 combinations of levels (strata) of the two variables, namely:

males <50, males > 50, females <50, females> 50.

The stratification will consist in assigning to groups A and B patients belonging to the 4 strata described above, using 4 different randomization lists.

Another case that often uses stratification is represented by multi-center trials, that is where the enrollment takes place in different recruitment centers (sometimes located in different countries). In this case, to avoid distortions related to procedures (evaluations, clinical examinations, etc.) that could be different from center to center and from country to country, we decide to stratify with respect to recruitment centers (that is, each center will have its own list of randomization).

In general, it is considered good practice to operate the stratification only in cases where there is no uncertainty about possible confounding factors. It is also not advisable when the size of the research is very large (in this case, in fact, the law of large numbers alone guarantees a good homogeneity between the groups).

Relevant errors: to not randomize; to choose not suitable techniques of randomization with respect to the experimental situation (*i.e.* simple randomization for small sample size, choice of too many confounding variables for the stratification).

Blindness

Blindness or masking is the way by which patients and clinicians remain unaware of which treatment each patient is assigned. A study conducted in the absence of blindness is defined as "open study".

Blindness should include:

- Patients, to avoid that the improvement or absence of improvement (or simply a public stake) are due simply to psychological effects;
- Clinicians providing treatment, to prevent them from transmitting, with words or nonverbal behavior, greater or lesser enthusiasm to the patient; also to prevent tend to care more closely of the patients assigned to the experimental therapy;
- Clinicians involved in patients evaluation, to ensure maximum objectivity of judgment;
- Statisticians, to avoid any form of manipulation.

Studies can be conducted in single blind when blindness concerns only the patients, and in double blind when blindness concerns both patients and clinicians providing treatment; finally, studies can be conducted in triple blind study, when blindness concerns even the clinicians that evaluate the patients, if they are different from the clinicians providing treatment.

Although triple blindness is the methodologically the ideal model, there are situations in which it is reasonable to use double or single blindness. The following considerations appear valid, according to the scheme formulated by Pocock [3]:

- Ethical aspects: sometimes blindness is not ethically correct (*e.g.* when blindness of patients requires not necessary invasive methods, as a placebo to be administered repeatedly by injection);
- Concrete feasibility: for example, there are cases in which the comparison is made with non-drug therapies (surgery, psychoanalysis, physiotherapy, and so on), with respect to which the conditions of blindness are not feasible;

Anyway, in cases where it is impossible to obtain the condition of blindness for patients and/or for the clinicians providing treatment, we should use a form of partial blindness, concerning at least the clinicians that assess the final evaluation of patients.

Relevant errors: to not perform double blind studies when they are feasible; to not guarantee blindness of clinicians that assess the final evaluation of patients, in case of single blind and open study.

Placebo and active control

Depending on the case, the control group of a trial may assume either a standard therapy (usually, the therapy normally recognized as the best one for the disease) or a placebo.

Thus, trials can be either active-controlled or placebo-controlled.

The placebo is a pharmaceutical formulation that is completely identical to that of the experimental drug for what concerns each organoleptic character and appearance (flavor, color, smell, packaging), except that in the active principle, completely absent in the placebo.

The purpose of the placebo is to evaluate the effect of the experimental drug actually attributable to the drug itself, and therefore net of the placebo effect. We define "placebo effect" the improvement that the patient simply shows because of the belief of assuming therapy [15].

Placebo effect occurs together to the pharmacologic effect during any therapy; in other words, a certain amount of improvement related to the placebo effect occurs during any therapy, and is therefore net of this dimension that the experimental drug should be evaluated.

From an ethical point of view, it is clear that we should use a placebo control only if a drug considered valid for the disease is not available.

Finally, there is a case where placebo is not used to take under control the placebo effect, but it has the purpose of ensuring the blindness to treatment. Suppose that an experimental drug is compared with a standard drug, and that the two drugs have different formulations and routes of administration: in this case, we can use the double-dummy technique, so that each patient takes the medication given to him along with a placebo formulated for the route of administration of the other active drug. In other words, a patient assigned to the experimental drug also will take a placebo quite similar (even for the route of administration) to the standard drug; *vice versa*, a patient assigned to standard drug also will take a placebo quite similar (even for the route of administration) to the experimental drug.

Relevant errors: to plan a placebo-controlled trial when a standard therapy exists.

Sample size

The sample size of patients to be enrolled in a clinical trial is the result of an *a priori* analysis. The basic principle is that the size should result from a compromise between two opposing views:

- the first is that research with too few patients have a high risk of producing false negative results, *i.e.* not

to highlight differences between the comparing treatments, for simple effect of “lack of data”;
- the second is that too large sample size should provide positive results from the statistical point of view (*i.e*. statistically significant results), but not relevant from a clinical point of view.

Nowadays, methodologists suggest to limit the use of the two terms to particular sections of the clinical reports; they recommend that the term “statistical significance” should be reserved to the “Results” section and the term “clinical relevance” to the “Discussion and conclusions” section [16].

Practically, the definition of the sample size of a search is calculated first by means of statistical methods [17, 18, 19] and further basing on feasibility considerations.

Statistical methods

In order to correctly use of the statistical formulas for the computing of sample size, we have to answers to five key questions:
- What is the main criterion for measuring the outcome? From a statistical standpoint, this question implies that the classification of the response variable (*i.e.* qualitative: nominal or ordinal; quantitative: discrete or continuous) is known.
- Which statistical test should be used to analyze the data? The choice of the test obviously depend on the type of the considered response variable.
- What results are expected from the control group? The answer to this question can be derived for example from the literature, by any pilot studies of the trial, from the personal experience of researchers, and so on.
- What is the minimum clinically relevant difference? That is, what may be the minimum difference in response between the control and the experimental group to be considered relevant from a clinical standpoint?
- What is the degree of statistical safety that must be achieved? Here, the term “degree of safety” refers to the acceptable risk of making errors of the type: false positive (or Type I errors) and false negatives (or Type II errors). These risks are indicated by α (or significance level of the statistical test) and β (though usually in the protocols you prefer indicate $1-\beta$, that the power of the test), respectively. Usually, in clinical trials, these risks of errors are fixed in $\alpha = 0.05$ (5%) and $\beta = 0.10$ or 0.20 (10% or 20%, corresponding to a power of 90% or 80%) [20].

Once the answers to these questions are given, we use suitable formulas that provide the number of patients required for the experimental trial.

Feasibility considerations

Once the statistical calculation is performed, it will be necessary to assess the feasibility of the trial; in particular, we have to establish whether the calculated number of patients is compatible with the expected rate of recruitment, if the economic funds are sufficient, and so on.

If such a realistic assessment is unfavorable, it is possible to operate - depending on the case - in different ways, such as:
- to increase the number of involved centers and investigators (in order to increase the rate of enrollment);
- to decrease the scientific safety degree (*i.e.* increase the risks a and b, but not exceeding the standard level of 5% and 20%, respectively).

Finally, if the previous strategies were not viable, it will be necessary to give up the search, because it would be unethical to start a research not leading to conclusive and reliable results [17].

Relevant errors: to not establish sample size; to not properly use the suitable formulas; to not consider the feasibility of the trial; to confound statistical significance and clinical relevance.

Monitoring

In case of long duration trial, we can perform some “interim” statistical analyses [21].

The purpose of these analyses can be schematically described in three following fundamental points

Monitoring the quality of research: it consists in evaluating the adherence of the researchers to the protocol. Different indicators of “loss of quality”:
- Changes in the rate of enrollment: an increase rate could point out an easing of the eligibility criteria; a decrease rate could indicate a lack of enthusiasm of the investigators;
- Changes in the distribution of patient characteristics: if the change takes place in the time and manner equals between the two arms, it may indicate an easing in the eligibility criteria; if it differentially concerns the two arms, it may indicate the breaking of blindness conditions;
- Changes in the level of response to treatment: if it concern both the arms (experimental and control) it may indicate a modification in the patient assessment; if it only concerns the experimental arm, it may be the signal of breaking of blindness conditions.

Monitoring side effects: usually, an independent monitoring committee evaluates the possible side effects (both expected and unexpected).

Monitoring the efficacy of the treatment: the purpose of these statistical analyses is to assess whether, before the experiments end, there is evidence about the greater or lesser effectiveness of one treatment.

In other words, the question can be asked by an ethical point of view: is it appropriate to continue to enroll patients in one arm of the trial if there are already significant evidence that the corresponding treatment is less effective than the other?

The problem, in this type of monitoring, is that each statistical analysis performed goes to raise the risk of false positives or errors of type I (remember that with the level of significance, normally fixed at 5%, the experimenter assumes the commitment to maintain this level within the risk of making a false positive, which is a type I error) [22].

The protocol should therefore fix the number, the manner and timing in which these interim analyses have to be performed during the trial. However, there are two main approaches to the problem [23, 24, 25]:

Group sequential design: it is based on the idea of performing repeated significance tests. In this case, to avoid exceeding the overall significance level fixed (usually 5%), and once that the required power (usually 80% or 90%) are known, a nominal significance level (lower than the global one) is fixed by means of special tables, in order to maintain the overall level within the preset limits.

Continuous sequential design: it is based on the idea to perform an analysis for each patients that ends the trial. It is obvious that the problems of increased risk of type I error result amplified. From a practical point, a purpose-built chart is used for this approach; see for example [26].

Relevant errors: to perform interim analyses not planned in the protocol; to use incorrect methods to control the global risk of type I error.

Deviation from protocol

A protocol must contain the way we deal with deviations from the protocol itself. Notwithstanding the fact that a protocol should be respected as much as possible, it should be provided for any violations that may occur and establish standard procedures to address them [27, 28].

It is clear that when deviations assume catastrophic proportions (for example, the drug proves to be unstable, some data were invented, and so on) there is no other solution but to close the trial. But when deviations involve individual patients, there are some possible approaches.

The main possible violations are the following:

- enrollment of not eligible patients, in which case the opinions are varied, but usually the idea of excluding them from the final statistical analysis prevails;
- incomplete adherence to therapy: this situation may be due to a lack of cooperation of the patient, or sometimes to a change of therapy determined by the attending physician; in this case the possible approaches are different and they will be described below;
- withdrawal of patients (dropouts): the patient may withdraw from the trial at any time, by choice or even in the judgment of the treating physician who believes having to transfer him to other therapy. Notwithstanding the need to continue the patient evaluation until the end of the study (if possible), also in this case the possible approaches are different and they will be described below.

Thus, the problem is the following: patients who did not adhere fully to the protocol or who have retired, have or not to be included in the final data? There are two possible approaches to the problem:

Per protocol analysis (also named drug efficacy approach, or explanatory approach): in this his approach we take into account only patients who strongly adhere to the protocol. This approach is characteristic of phase II trials, *i.e.* studies aiming to an initial assessment of the therapeutic effects of the experimental drug evaluation that can be obtained only under a perfect adherence to the treatment protocol, as well as phase I trials, *i.e.* studies that aiming to determine the dosages based on the toxicity and the kinetics of the drug; in both cases, we need information that can be obtained only under conditions of perfect compliance.

Intention to treat analysis (also named pragmatic approach): in this approach we taken into account all the patients who participated in the study, whether or not they adhered to the protocol and whether or not they completed the study itself (of course provided we obtain the final evaluation). The only exception is represented by patients who have withdrawn before starting treatment. The idea is that if we exclude patients withdrawn or with poor adherence to the protocol, we would overestimate the effect of treatment (in fact, it is logical to think that withdrawals and no adhesions to the protocol are mainly caused by a dissatisfaction with the treatment). Furthermore, this approach tends to evaluate the effectiveness of the drug in an environment similar to standard clinical practice, where it is the norm that a patient tends to adjust the dose, change the timing of intake, discontinue therapy for short periods, and so on. For all these reasons, this is the approach characteristic of the phase III trials. However, we have to point out that, even at this stage, it is usual to combine the intention to treat analysis and the per protocol analysis, to get more information on the effectiveness of the drug.

Relevant errors: to not provide a way to deal with deviations from protocol; to choose incorrect approach with respect to the phase of the trial.

Plan for statistical analyses

In this section, we will list briefly some points that should be taken into account in the drafting and analysis of a protocol [29].

Descriptive indices (arithmetic mean, geometric, harmonic, median and mode, range, variance, standard deviation, coefficient of variation, interquartile range) should be chosen in an appropriate manner, taking into account the scale of measurement of the variables, the presence of censored data, the asymmetry of the distributions.

Plots should be chosen as appropriate in relation to the phenomenon being described, honest in choosing the axis scales, and if possible they should consider the individual observations rather than the values grouped into classes.

Measures of treatment efficacy should be always provided, and they should be specified in the protocol (*e.g.* ARR, or Absolute Risk Reduction; RR, or Relative Risk; RRR, or Relative Risk Reduction, NNT, or Number Needed to Treat; OR, or Odds Ratio).

Outliers (values strongly extreme into the distribution) should be excluded from the statistical analysis only if there are strong doubts about their credibility; specific statistical tests (such as the Dixon test) have to be used to support this decision.

Each estimate should be accompanied by a confidence interval, that takes into account uncertainty of the estimate. The technique of calculation of the confidence interval should take account of the distribution of the variable (for example, if the distribution is not normal Gaussian, one should make use of non-parametric techniques, such as the quantile method or bootstrap technique).

Statistical tests should only respond to questions submitted in advance (*i.e.* in the protocol) to avoid data dredging phenomena [30]. Other tests (post-hoc tests) should have a pure exploratory value.

Statistical tests should be selected properly based on the scale of measurement of the variables.

Statistical tests should always be performed in two-tailed version, unless there are reasonable reasons to choose the one-tailed form.

For each statistical test we have to verify the so-called parametric assumptions (such as normality and homoscedasticity for the execution of the Student's t test and of Anova). If these conditions are not satisfied, we can proceed to appropriate data transformations [31] and, in case of further failure in the basic assumptions, we have to perform the corresponding nonparametric test [32].

In case of multiple endpoints, we should proceed in order to avoid an excess risk of false positive error; we can choose either to apply Bonferroni's criterion (subdividing the overall significance level for the number of endpoints considered) or to apply multivariate analysis techniques (which consider simultaneously all endpoints) or, finally, to construct an overall score from single endpoints (thus obtaining a single response variable summarizing and takes account of all endpoints) [33].

<u>Relevant errors</u>: to use inappropriate descriptive indices; to use inappropriate or not-honest plots; to not provide measures of treatment efficacy; to exclude outliers without a valid justification; to not provide a confidence interval for the estimated values; to perform data dredging; to choose not suitable statistical tests; to perform one-tailed tests without a valid justification; to perform parametric statistical test without verification of the assumptions; to not use corrective methods in case of multiple endpoints.

CONCLUSIONS

In this short review, we have tried to show that statistical error in medicine should hide in several phases of experimental process, and not only in the execution of statistical tests.

Thus, statisticians should be also involved in the experimental planning, in order to avoid many types of error.

REFERENCES

1. *Altman DG. Statistics and ethics in medical research. Misuse of statistics is unethical. Br Med J 1980;281:1182-4.*
2. *Peto R. Clinical trial methodology. Biomedicine 1978;28:24-36.*
3. *Pocock SJ. Clinical trials, a practical approach. New York: Wiley; 1983.*
4. *Norman G. Data dredging, salami-slicing, and other successful strategies to ensure rejection: twelve tips on how to not get your paper published. Adv Health Sci Educ Theory Pract 2014;19:1-5.*
5. *Foster D, Shaikh MF, Gleeson E, et al. palliative surgery for advanced cancer: identifying evidence-based criteria for patient selection: case report and review of literature. J Palliat Med 2015 [Epub ahead of print]*
6. *Slankamenac K, Nederlof N, Pessaux P, et al. The comprehensive complication index: a novel and more sensitive endpoint for assessing outcome and reducing sample size in randomized controlled trials. Ann Surg 2014;260:757-62.*
7. *Simon EG, Fouché CJ, Perrotin F. Introduction à la méthodologie des essais randomisés : le choix du critère de jugement. Gynecol Obstet Fertil 2011;39:595-6.*
8. *Brown BW. Statistical controversies in the design of clinical trials, some personal views. Contr Clin Trials 1980;1:13-27.*
9. *Cochran WG, Cox GM. Experimental designs. New York: Wiley; 1957.*
10. *Armitage P, Hills M. The two-period crossover trial. Statisticians 1982;31:119-31.*
11. *Hills M, Armitage P. The two-period crossover clinical trial. Br J Clin Pharmacol 1979;8:7-20.*
12. *Pocock SJ. Allocation of patients to treatment in clinical trials. Biometrics 1979;35:183-97.*
13. *White SJ, Freedman LS. Allocations of patients to treatment groups in a controlled clinical study. Br J Cancer 1978;37:849-57.*
14. *Zelen M. The randomization and stratification of patients to clinical trials. J Chron Dis 1974;27:365-75.*
15. *Gribbin M. Placebos: cheapest medicine in the world. New Sci 1981;89:64-5.*
16. *International Steering Committee. Uniform requirements for manuscripts submitted to biomedical journals. Can J Publ Health 1978;69:454-8.*
17. *Altman DG. Statistics and ethics in medical research. How large a sample? Br Med J 1980;281:1336-8.*
18. *Freedman LS. Tables of the number of patients required in clinical trials using the logrank test. Stat Med 1982;1:121-9.*
19. *Gore SM. Assessing methods, trial size. Br Med J 1981;282:1687-19.*
20. *Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The impor-*

tance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 negative trials. N Engl J Med 1978;299:690-4.

21. *Pocock SJ. Interim analysis for randomized clinical trials: the group sequential approach. Biometrics 1982;38:153-62.*
22. *Pocock SJ. The size of cancer clinical trials and stopping rules. Br J Cancer 1978;38:757-66.*

23 *Armitage P. Sequential medical trials. Oxford: Blackwell; 1975.*

24. *Demets DL, Ware JH. Group sequential methods in clinical trials with a one-sided hypothesis. Biometrika 1980;67:651-60.*
25. *Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika 1977;64:191-9.*
26. *Wu J, Xiong X. Group sequential design for randomized phase III trials under the Weibull model. J Biopharm Stat 2015;25:1190-205.*
27. *Wolf GT, Makuch RW. A classification system for protocol deviations in clinical trials. Cancer Clin Trials 1980;3:101-3.*
28. *Schwartz D, Lellouch J. Explanatory and pragmatic attitudes in therapeutic trials. J. Chron Dis 1967;20:637-48.*
29. *Armitage P. The analysis of data from clinical trials. Statistician 1980;28:171-83.*
30. *O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics 1979;35:549-56.*
31. *Gore SM. Assessing methods, transforming the data. Br Med J 1981;283:548-50.*
32. *Siegel S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill; 1956.*
33. *Lord SJ, Gebski VJ, Keech AC. Multiple analyses in clinical trials: sound science or data dredging? Med J Aust 2004;181:452-4.*