VIA MEDICA

REVIEW ARTICLE

Kazimierz Drosik

Randomized clinical studies — science, belief or advertisement only

Address for correspondence:
Dr n. med. Kazimierz Drosik
ul. 28 Lawendowa, 45–470 Opole
Phone: +48 506 197 837
e-mail: drosik@go2.pl

Oncology in Clinical Practice
2020, Vol. 16, No. 3, 133–139
DOI: 10.5603/OCP.2020.0022
Translation: dr n. med. Dariusz Stencel

ISSN 2450–1654

ABSTRACT

The results of randomized clinical studies are important in evaluation of drugs' or medical procedures' efficacy. Statistical analyses are meaningful in publications and conference presentations. This paper discusses the role and value of selected statistical methods. It is clear that in clinical studies patients and time to event (recognized as end study point) are essential. According to that the question arises — do all these statistical analyses are important or do they play only a role in drug promotion.

Key words: clinical studies, statistical analyses

Oncol Clin Pract 2020; 16, 3: 133–139

Introduction

Randomized clinical trials (RCTs) play the most important role in evaluating new drugs and new treatment methods, as well as in establishing the standards of care. The results of these studies are widely accepted; however, their interpretation may raise some doubts. The main problem is connected with the statistical methods used, which are often poorly understood by the average reader or reluctantly scrutinised due to the unlimited reliance on the authors' interpretation.

The value of new anti-cancer drugs and other treatment methods is assessed in the development process, including different phases of studies having varied aims. The study could be aimed at determining the maximum tolerated dose (MTD), determining drug activity, or evaluating its efficacy in terms of the effect on patient survival.

Cancers constitute serious clinical and social problems because they frequently shorten life expectancy. Therefore, determining the effect of a drug or treatment method on extending overall survival is of key importance. Clinical trials with survival as the study endpoint are therefore the most important, and they generally summarise the results of the earliest phase studies. An alternative to overall survival as an endpoint is time to relapse or time to disease progression. Although there is ongoing discussion about the superiority of one endpoint over another, this is not relevant for the purposes of this publication. It is important that in both cases the time is measured from the patient's entry into the study (in a randomised study it should be the date of randomisation) until the event, which may be relapse, progression, or death. The research methodology is the same at least from a statistical point of view. The differences are only associated with the ability to determine the time of the event — the time of death is a single point in time independent of study assumptions. Recurrence or progression of the disease is also a one-off event in the timeline; however, they are most often diagnosed during pre-planned periods in which subsequent control examinations are performed. If, then, the results of the study are presented in the form of a graph, the curve presenting the time to progression or relapse will have a stepped shape, while the overall survival curve will be continuous. For the above reason, it will be easier to discuss the problem of interpretation of study results based on a model with relapse or disease progression as a study endpoint.

Randomised clinical trials

Stratification and randomisation

The greatest difficulty in planning and conducting a clinical trial with "time-to-event" as a study endpoint

is that it is not possible to determine in advance at which time point the event can or should occur. If it was known how long the untreated patient would survive until the event occurred, it would be easy to show how much longer the survival time would be after using the study drug or other treatment. Each patient participating in the study would be his/her own control. Unfortunately, this is not the case, and therefore the clinical trial is based on a comparison of the results in an "experimental" group of patients with the results in the control group. The basic condition is that the patient groups are so similar that the only difference between the studied and control arm is the drug (combination of drugs) or treatment method used.

It would be best to carry out the study on identical twins, but even in this case it would be doubtful whether all events would be the same and occur at the same time. The questions even arise when the study is conducted with unrelated patients. In order to enrol maximally comparable patients into study arms, the principle of stratification and random assignment of patients to particular study groups (randomisation) was introduced. The purpose of stratification is to evenly separate patients in terms of prognostic factors with a known impact on the occurrence of the endpoint event. Obviously, there are an increasing number of factors that should be included in the stratification, which results from the in-depth knowledge about the biology of a given disease. The goal of randomisation is equal distribution of unknown prognostic factors. It is assumed that due to the random distribution of patients, these factors with an unknown effect on the event will be distributed equally in both arms. By definition, these factors are unknown, which makes it impossible to determine them at a given time and to obtain real comparability of patient characteristics in both study arms. One can only believe that it is so. Instead of proof, there is only the belief that we have proof. It is important at this point to pay attention to the correct qualification of patients for the study. Qualifying patients who do not fully meet the inclusion criteria (not entirely eligible) to enable participation "at all costs" can clearly affect the outcome. There should also be no individualisation of decisions to include patients into the study — the rule is that in the participating site, every patient who meets the inclusion criteria should be qualified if he/she agrees. However, any patient who meets the criteria but is not included in the study, for whatever reason, can affect the final result and the quality of the study. Obviously, it is unacceptable to conduct two or more trials with identical selection criteria. The assignment of patients to different concurrently conducted studies based on the doctor's decision completely distorts the sense of randomisation. This important error is unfortunately difficult to detect; one can only appeal to the ethics of investigators.

Course of the study

The patient enrolled into the study, assigned based on stratification and randomisation to the examined or control arm, receives appropriate treatment — it is a new drug (or combination) or a new method of treatment, and in the control group, for example, this is a standard of care. In the case of a study with relapse or disease progression as an endpoint, follow-up examinations are carried out at regular and predetermined intervals. If the assessed event is found, the patient discontinues the study but continues the toxicity (safety) follow-up period. In the present study this aspect is completely omitted because the methods of toxicity assessment are simple and do not require any special knowledge. Interpretation of effectiveness analyses is a real problem.

In the study assessing the effectiveness there are only two elements: the patient and the time to event. All patients had to meet the inclusion criteria and are therefore similar in each arm. However, it should be remembered that within the same arm the patient population is diversified even in terms of stratifying factors.

As a result of subsequent follow-ups, patients with endpoint events are excluded from further evaluation. It is best to examine this on a simple chart in which the number of patients is on the ordinate axis and the time (time intervals in which follow-ups take place) on the abscissa axis. This would be the simplest and most realistic way to present the study results. There is no need to recover an exemplary clinical trial at this time because it is possible to create many models of such a trial and to make charts based on the above principle. If the charts of some real study would be taken as the basis, a model of this study could be also developed. Although there are no absolute numbers in the presentation of the actual study result, only probability curves, at this moment only the shape of the curve is of special interest. This problem will be explained in a later part of this publication. If instead of the probability the actual number of patients were inserted, then it could be revealed that in both arms the number of patients still living event-free decreases; however, in a positive study the number of patients without an event decreases faster in the control arm. In many studies, however, these differences are not large. Many models could be created and then compared to actually published studies. An example would be a study with an equal number of patients in both arms at baseline. If, for example, during the first two assessments the number of events is equal in both arms, both curves on the graph would overlap. Now it could be supposed that in the third assessment the number of events is higher in the control arm. The curves on the chart will spread apart (so-called curve separation) by a size that is the difference in the number of patients with a given event in the examined and control arms. If in subsequent assessments the number of patients without an event decreases in

both arms by the same amount, then the curves will run in parallel and will be further "separated". This gives the impression that there are still differences between the arms, although in reality there is no difference, because the number of events in both arms is the same. There are only fewer patients without an event in the control arm as much as there were more events in this arm at the third assessment. This is presented in Table 1 and Figure 1. In this model, the difference between events in the A and B arms at the third assessment is 10 patients. When an event in the control arm occurs in the middle patient, the curve for that arm will cut a line from the middle of the ordinates. The curve for the studied arm will cross this line with a delay. It could be concluded that the median time-to-event increased for experimental arm.

Let us perform the next experiment and increase the number of patients in the control arm who had an event at the third assessment. The curves will even more separated, and at the same time the difference between the medians will increase, as shown in Table 2 and Figure 2. It follows that increasing or decreasing the difference between medians depends primarily on the difference in the number of events in both arms. If the time point of assessment in which differences were found were changed (not the third one, but any subsequent one), it will transpire that this does not affect the difference between the medians. Still, this difference will depend only on the differences in the number of events, which is presented in Table 3 and Figure 3. Obviously, it is important that these differences occur in the first half of the total number of patients participating in the study. These differences could be identified during each subsequent assessment, and then the sum of them would affect the median.

So, you can see here that the median is more a measure of the number of events, but not the time at which these events occur.

In the created model there are still two separated curves that run parallel to each other. It is assumed that the curves separated because patients in the experimental arm received more effective treatment. What happens, however, when effective treatment is

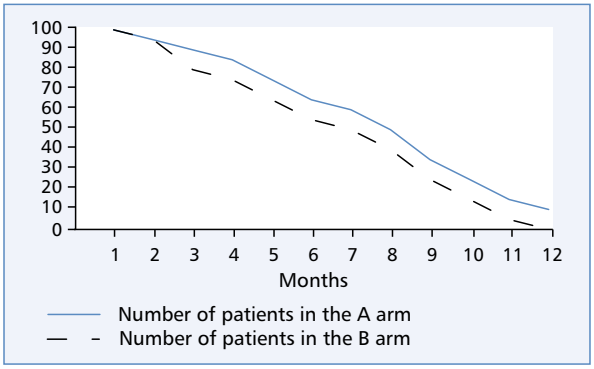

Figure 1. The data are presented in Table 1

Table 1. Columns A1 and B1 present the number of patients starting a given observation period, columns A2 and B2 the number of patients with an event in a given period, and columns A3 and B3 the number of patients without an event at the end of the evaluated period

Months	Number of patients					
	The A arm			The B arm		
	1	2	3	1	2	3
1	100	0	100	100	0	100
2	100	5	95	100	5	95
3	95	5	90	95	15	80
4	90	5	85	80	5	75
5	85	10	75	75	10	65
6	75	10	65	65	10	55
7	65	5	60	55	5	50
8	60	10	50	50	10	40
9	50	15	35	40	15	25
10	35	10	25	25	10	15
11	25	10	15	15	10	5
12	15	5	10	5	5	0

only applied to patients in the control group, who have caused a difference in the number of events in the third assessment (first example) or in the subsequent assessment (second example)? If this treatment is more effective then there should be no events in these patients, and the curve in the control arm will not move down (both curves will still overlap). The above situation confirms the statement that a small number of patients may decide on the final study result. If the final result of the study were presented in absolute numbers, as in the proposed models, there would be clarity as to the actual effectiveness of the new drug, combination of drugs, or another new treatment method. The result would be visible on these simple charts without using any statistics.

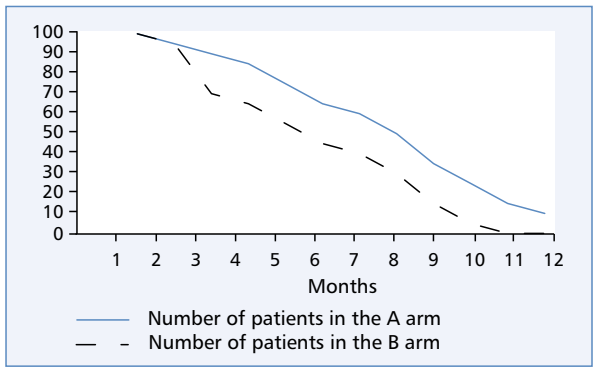

Figure 2. The data are presented in Table 2

Finally, it is worth checking on the models the number of possibilities for running of event curves for the same median difference. In each case it could be found that it always depends on the number of events (not always the time point at which these events occurred).

In publications presenting the result of a clinical trial in the form of graphs of overall survival or time to another endpoint the probability curves are shown instead of absolute numbers.

Kaplan-Meier estimator

The result of the study presented as absolute numbers can easily be understood. However, the main disadvantage of such a solution is that such an analysis would be possible only after completion of the study by all patients (after occurrence of an end point event in all patients). Such a study would last for a very long time, which would particularly apply to adjuvant treatment. For this reason, Kaplan-Meier estimator, i.e. calculation for incomplete observations, is used to analyse the study [1, 2].

Consecutive patients are included in the clinical study, and the time to event is assessed. The difference between the arms is of great interest, so it is prospectively determined. Based on the current data on the number and duration of events in patients treated by the method that will be used in the control arm, the number of patients (sample size) needed to prove the thesis that

Table 2. Columns A1 and B1 present the number of patients starting a given observation period, columns A2 and B2 the number of patients with an event in a given period, and columns A3 and B3 the number of patients without an event at the end of the evaluated period

Months	Number of patients					
	The A arm			The B arm		
	1	2	3	1	2	3
1	100	0	100	100	0	100
2	100	5	95	100	5	95
3	95	5	90	95	25	70
4	90	5	85	70	5	65
5	85	10	75	65	10	55
6	75	10	65	55	10	45
7	65	5	60	45	5	40
8	60	10	50	40	10	30
9	50	15	35	30	15	15
10	35	10	25	15	10	5
11	25	10	15	5	5	0
12	15	5	10	0	0	0

the difference in events will reach the assumed level and will be statistically significant is determined. Please note that it is assumed that the events will occur in both arms, but in a pre-planned period there will be fewer events in the examined arm. The study will not be conducted until all participants experience the end point event (death, relapse, or disease progression) but until the assumed number of events is achieved. Obviously, patients participating in the study, in whom end point event will not occur, will continue the treatment; however, it will no longer be a subject for fundamental analysis.

At the time when the assumed number of events is recorded, i.e. the study as such is completed, patients included in the study will have different follow-up periods, i.e. patients enrolled at the beginning have the longest ones, while patients included shortly before study completion have the shortest ones. Due to the different time of participation in the study, an assessment based on absolute numbers would not be possible. Therefore, estimation using the Kaplan-Meier method is used, i.e. the probability of surviving a specific event-free time is determined and absolute numbers are replaced by the probability. Let us assume that 100 patients were included in each arm. Each month of participation in the study is followed by an evaluation. If all patients survived the first and second month without an event, then after two months there will still be 100 patients in each arm. The probability of event-free survival will be calculated by dividing the number of patients who survived a given period without an event by the number of patients who started that period. In this case it will be 100 : 100 = 1 for the first and second month. However, if during the third assessment (after three months) end point event was detected in five patients in the examined arm, i.e. 95 patients survived without the event, the probability of surviving the given event-free period would be 95 : 100 = 0.95. At the same time, 15 events were found in the control arm, i.e. the probability of event-free survival without an event in the control arm will be 85 : 100 = 0.85. The probability of event-free survival is calculated for each period separately. Thus, 95 patients in the examined arm and 85 patients in the control arm will enter the next assessment period. For example, if during the next assessment an event is found in four patients in

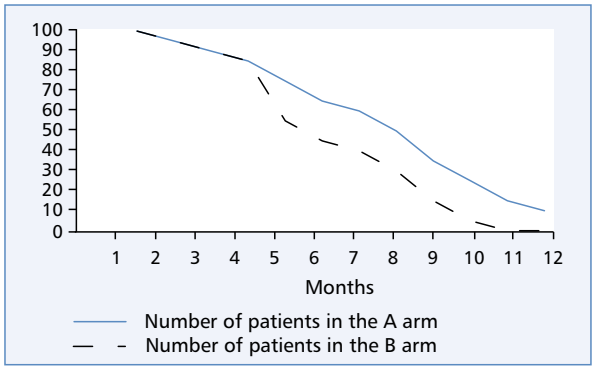

Figure 3. The data are presented in Table 3

Table 3. Columns A1 and B1 present the number of patients starting a given observation period, columns A2 and B2 the number of patients with an event in a given period, and columns A3 and B3 the number of patients without an event at the end of the evaluated period

Months	Number of patients					
	The A arm			The B arm		
	1	2	3	1	2	3
1	100	0	100	100	0	100
2	100	5	95	100	5	95
3	95	5	90	95	5	90
4	90	5	85	90	5	85
5	85	10	75	85	30	55
6	75	10	65	55	10	45
7	65	5	60	45	5	40
8	60	10	50	40	10	30
9	50	15	35	30	15	15
10	35	10	25	15	10	5
11	25	10	15	5	5	0
12	15	5	10	0	0	0

the examined arm, the probability of event-free survival in this period will be 91 : 95 = 0.96. However, to survive these four consecutive months, the patient had to survive the first three months. Because of this, the probability of surviving the following months is multiplied. Therefore, the probability of survival for four months will be $1 \times 1 \times 0.95 \times 0.96 = 0.91$. On the other hand, if in the control arm in the fourth month, for example, six events are found, then the probability of surviving the fourth month will be 79:85 = 0.93, and the probability of surviving four months will be $1 \times 1 \times 0.85 \times 0.93 = 0.79$. In this way, by multiplying the probability of survival of consecutive periods, the probability of survival for the entire follow-up period could be obtained. However, please note that at the completion of some real study the number of patients assessed in particular periods will decrease not only because of diminishing of patients due to end point event occurrence, but also because the patients later included into the study have insufficient follow-up period. For example, patients who have only been participating in the study for six months cannot be taken into account in calculating the likelihood of survival of eight months and longer. This is the advantage of the Kaplan-Meier estimator over the analysis based on absolute numbers. On the graph, the curves presenting the results of the study in absolute numbers are replaced with survival probability curves. If the study was to be completed only after the event occurred in the last patient and presented in the form of curves based on absolute numbers, these curves should coincide with the probability curves obtained earlier. In the publications of randomised clinical trials the graphs of survival probability curves are frequently, albeit not always, presented together with a table showing the absolute numbers of patients who were the basis for calculating the probability of survival for a given period. According to this, it can be seen that although the curve shows that 20% of patients would probably survive, in some cases it was calculated based on survival of only one or two patients. The course of the curves could allow, however, to read out the same thing that would be on the graphs of absolute numbers, i.e. it is possible to calculate the differences in terms of events in individual periods between the arms. The difference between the medians can be seen at the moment when the probability of survival in each arm accounts for 0.5. It does not change the fact that this difference will still depend primarily on the difference in the number of events between the arms. For this reason, the common claim that the "new" treatment increases the survival time by the difference between the medians is not justified. Perhaps, some patients have actually increased their survival to event, but it certainly does not apply to all patients. To be able to state that the new treatment method prolongs the survival time of all patients by the median noted in the study, patients from both arms would not have an event for some time, then at the same time all patients from the control arm would have to have an event, and none from experimental arm, and after some time, corresponding to the median difference, all patients from the experimental arm would have an event at the same time. The reality of such a situation is difficult to imagine.

In the majority of even "positive" trials, most patients have the same survival time in both arms at each assessment period. The final difference found in the study is the sum of the differences in subsequent assessments. The supposition that all patients in the experimental arm benefited is not substantiated. These kinds of statements, which we often find in presentations, however, are merely advertising the medicine or method.

Statistical significance in clinical trials

In order to authenticate study results, statistical significance tests are used. Even if the difference is statistically significant, it does not mean that it is clinically significant. This is observed when the real difference is small.

If there were 10 patients in each of the two arms of the study, and as a result of using the new drug in the examined arm only one patient had an event against nine in the control arm, the difference would be visible with the naked eye and statistical tests would be unnecessary. However, if there were five events in one arm and six in the other, there would be doubts as to whether this difference was not accidental. In this case, statistical tests are necessary as well as increasing the number of patients needed to prove the difference. This is already taken into account at the study planning level. Assuming the size of the clinically significant difference, the number of patients needed to prove the difference in statistical significance tests is calculated. However, an important fact is noteworthy — two arms are compared on the assumption that the only element determining the existence of the difference is the drug or method of treatment used. Unfortunately, it is not possible to prove that, except for the drug or method, the patients in both arms are identical. Only faith remains that thanks to stratification and randomisation this is indeed the case. In this way, however, the value of scientific mathematical proof depends on what we believe. Any misconduct during the study, at any of its stages, may undermine the value of the result obtained. So, science or just faith?

To validate the results, further statistical tests are used. One of them is the so-called hazard ratio (HR). It is calculated in such a way that the risk of an event in one arm (the number of patients with an event divided by the sum of patients with an event and without an event in a given arm) is divided by the equally calculated risk in the other arm. This can be applied to the total number of patients or to individual cohorts created, for example, by age, disease stage, or other criterion [3].

On this basis, it is concluded to what extent the new drug or method reduces the risk of the event. However, it is not about the risk of an individual patient, but the risk of an event in a group of patients from a given arm. Hence, this is the same as can be seen on the charts, but differently presented. Please note that the clinical value of HR will depend not only on the number of events, but also on the number of patients in the study arms or individual cohorts. For example, if there are three events in 10 patients in the A arm and six events in 10 patients in the B arm, the HR will be 0.3 : 0.6 = 0.5. If in another study in the A arm there are three events per 100 patients and in the B arm six events also per 100 patients, the HR will be 0.03 : 0.06 = 0.5. The same result will be obtained with, respectively, three and six events per 1000 patients, 10,000 or more in each of the arms. In all of the examples above the risk reduction was 50%; however, the same magnitude of this reduction will have a completely different clinical significance.

Another statistical analysis is the so-called "forest plot", which shows the results in individual cohorts of patients (by age, disease stage, and other parameters). If the analysis shows differences in the result between patients from different arms, it is most often concluded that all patients benefit from using a new drug or method. This is clearly an erroneous conclusion. The result of this analysis only indicates that patients who have benefited from treatment with the new drug or treatment method belong to all or almost all cohorts. In each cohort, therefore, there are both patients who have benefited and those who have not. This, unfortunately, makes it difficult to detect patients who can actually benefit from treatment with a new drug or method, making it possible to conclude that it is necessary to treat all patients meeting the study eligibility criteria. This is to be proved by statistical tests, which, however, regardless of their number, can show nothing more than the fact that the events are presented in both arms but that only in the experimental arm there are fewer ones (usually by just a small margin). Multiplication of statistical tests that show this in various ways resembles drug advertising more than scientific evidence.

What is missing in the presentation of the result?

The study compares the frequency of events between the arms. However, during each assessment, there are patients in each arm, who have or do not have an event at that time. Of note, the course of the curves indicates that in each arm there are patients in whom the assessed event is found already at the first examination, as well as patients who do not have events until the end of the study. We do not know the decisive differences in terms of patient characteristics because they all meet the same eligibility criteria. It could be assumed that this is due to enrolment in the study of patients with various known risk factors for the event (stratification), but this is not subjected to any analysis. It could be also expected that the end point event will firstly occur in patients with the highest risk of this event at the enrolment, e.g. due to disease stage. This group will be in the first half of patients with assessed events, so they will decide about the median. It is not known whether this is the case. However, if this were true, the outcomes of each study would result from results in patients at the highest risk of the event, and the others would be only a kind of "supplement" justifying the use of a new drug or method in all patients meeting the inclusion criteria, although for many of them it may not matter which arm they are in. It is also not clear why, when assessed at a given time point, events manifest themselves in both arms (although there are fewer in the experimental group). What common features do these patients have in both arms? If they have something in common, what causes the difference in the number of events between the arms? Such questions can be multiplied, but the answers to these questions are sought very rarely. There is no doubt, however, that this is the only way to find patients who should be treated with a new medicine or a new method of treatment, because only those will benefit from such treatment.

Summary

It seems that more weight is attached to convincing everyone that all patients who meet the eligibility criteria should be treated, although only a few will actually benefit. For most patients participating in the study, the time to event is similar in both arms.

In this way, the clinical trial becomes, first of all, a method of promoting the drug or method, for which the statistical analyses used are to make credible.

Conflict of interest

None declared.

References

1. Kaplan EL, Meier P. Nonparametric estimation from incomplete observation. Journal of the American Statistical Association. 1958; 53(282): 457–481 (dostępny: https://web.stanford.edu/~lutian/coursepdf/KM-paper.pdf).
2. Green S, Benedetti J, Crowley J. Clinical trials in oncology second edition. Chapman and Hall/CRC. Boca Raton, London, New York, Washington DC. 2003: 30–37.
3. Green S, Benedetti J, Crowley J. Clinical trials in oncology second edition. Chapman and Hall/CRC. Boca Raton, London, New York, Washington DC. 2003: 37–39.