VIA MEDICA

REVIEW ARTICLE

Marcin Kaczor[1, 2], Rafał Wójcik[2], Joanna Połowinczak-Przybyłek[3], Piotr Potemski[3]

[1]Jagiellonian University Medical College, Krakow, Poland
[2]Aestimo, Krakow, Poland
[3]The Department of Chemotherapy, Copernicus Memorial Multidisciplinary Centre for Oncology and Traumatology, Lodz; Chemotherapy Clinic, Medical University of Lodz, Poland

Critical appraisal of clinical trials in oncology — part I

Address for correspondence:
Dr n. med. Marcin Kaczor
II Katedra Chorób Wewnętrznych
im. prof. Andrzeja Szczeklika
Uniwersytet Jagielloński
Collegium Medicum w Krakowie
e-mail: marcin.kaczor@uj.edu.pl

Oncology in Clinical Practice
2019, Vol. 15, No. 2, 89–103
DOI: 10.5603/OCP.2018.0057
Translation: dr n. med. Dariusz Stencel

ISSN 2450–1654

ABSTRACT

The main concept of evidence-based medicine is that all therapeutic decisions should be based on results from relevant, credible, and up-to-date clinical trials. Availability of a publication presenting a description of a clinical trial conducted with reliable methods and its high-quality results seems to be an ideal situation from the practitioner's point of view. However, reading only the abstract or just the author's conclusions may not always be sufficient to make the right clinical decision. For this purpose, several aspects of the clinical trial should be put under assessment, namely the methodology, its quality, internal and external credibility, clinical and statistical significance, as well as consistency of the results. The ability to perform the proper assessment of clinical trials may prove to be very helpful for practicing oncologists, especially in the case of new, emerging therapies, specific clinical situations, or when salvage treatment is necessary. It is also worth emphasising that the outcome assessment in oncology trials is specific, mainly due to the role of the survival analysis, which is relatively difficult to interpret. In this paper we tried to present in a clear and intelligible way the theoretical basis and subsequent steps in the critical appraisal of methods and results of clinical trials in oncology.

Key words: oncology, randomised clinical trial, critical appraisal, statistical analysis, survival analysis

Oncol Clin Pract 2019; 15, 2: 89–103

Introduction

Decisions regarding the choice of treatment are made based on correctly performed and reliable clinical trials. The results of clinical trials are used to develop the current guidelines for clinical practice in accordance with the principles of evidence-based medicine (EBM). To assess whether the conclusions from the study are appropriate, first of all it should be critically analysed for **internal credibility**. In order to do this, it should be assessed whether the study has been carried out correctly (an appropriate methodology ensuring reliable and undistorted inference and proper statistical analysis) and whether there is **internal consistency** of conclusions in a range of individual endpoints. **External consistency** assessment can be also helpful, determining whether a similar effect was observed in other clinical trials. Then an assessment of **external credibility** should be made, to find out whether the results of a clinical trial recognized as internally reliable can be extrapolated to the population subjected to treatment under real clinical practice, and whether similar clinical effects could be expected in these circumstances (patients' characteristics, additional medical procedures, appropriate comparator, compliance of study participants). Finally, clinical significance of the results should be assessed to answer the question of whether the magnitude of the observed effect indicates **significant clinical** benefit (taking into account the prognosis in a given patient population) and whether it really should lead to a change in clinical practice [1].

The individual elements of critical appraisal of clinical trials are discussed below. In addition, taking into account the specifics of clinical trials in oncology, the analysis of "time to event" endpoints is presented in more detail.

Clinical trial methodology

Correctly designed and conducted, blinded, randomised clinical trials (RCTs) provide evidence with the highest level of credibility [2]. These are experimental tests that assess at least two therapeutic interventions, and their use in patients is strictly controlled according to a previously developed study protocol. In oncology these studies usually have the character of trials with parallel groups. In some populations of oncological patients it is difficult to carry out a randomised trial, which may be due to the low prevalence of some cancers or small numbers of patients in specific clinical stages or treatment lines. It these cases it is necessary to conduct the study without a control group (single-arm). However, the methodological quality of such studies is initially lower than that of randomised trials. The same applies to cohort studies, which include a control group, but, due to the lack of randomisation, the non-random distribution of disturbing factors is a burden, and inference about the observed differences in the effectiveness of therapy is limited [1].

Randomisation

Randomisation (random allocation of patients to respective groups) is used to obtain as similar as possible or almost identical baseline clinical and demographic characteristics of patients, which, with an appropriately large population, ensures balanced distribution of all potential, as well as unknown, confounding factors. Therefore, the randomisation procedure cannot be carried out on the basis of simple assumptions, such as a medical history number or date of birth, because it allows the prediction of which group a patient will be allocated (this is called pseudorandomisation). Randomisation methods providing full randomness, i.e. unpredictability, include those in which the lists of random numbers are created with use of a computer or special tables (such a method is called simple randomisation). In the case of a small target number of patients (sample size) in the study the probability of unbalanced number and distribution of patients' characteristics in individual groups is higher; in this case more complex randomisation methods can be used. Examples of these include: block randomisation (patients are assigned to individual interventions in blocks, or groups with a specific sequence of subsequent patients allocation), stratified randomisation (independent stratification in any previously defined layer, such as country origin, gender, or type of previously used treatment, especially when differences in the effectiveness of assessed intervention between these subgroups are expected), or adaptive randomisation (in which the probability of allocation to a given group changes during the study, allowing the control of distribution of individual features in particular groups) [3]. In some studies unequal distribution **to the studied groups** is used, e.g. in a 2:1 ratio, which may increase the amount of information about a new therapy, especially regarding safety as well as recruitment capacity (patients are more willing to participate in the study due to a greater chance of receiving experimental therapy), but it adversely affects the statistical power and requires a higher sample size compared to allocation with a 1:1 ratio [3].

Allocation concealment and blinding

With random assignment of patients to study arms the process of allocation concealment is very important, to prevent access to information about the group to which the patient was assigned — which is possible with use of central randomisation, performed regardless of the individuals participating in the study. Additionally, allocation concealment allows elimination of influence of the researcher on patient assignment to particular groups, thereby reducing the risk of selection bias. The second step that ensures greater credibility is the introduction of **blinding**; therefore, the patient (single-blinded) or the patient and investigator (double-blinded) or patient, investigator, and team analysing the results (triple-blinded) are not aware of which intervention is received by each patient. It provides higher credibility of the study due to elimination of some confounders — a terminally ill patient, who knows that he/she was assigned to a placebo group instead of an active intervention group, may present much worse results than a patient who is unaware of the study assignment [4, 5]. In the case of medicines blinding is ensured through their preparation in the same form (e.g. in visually identical vials), and for different routes of administration or collation of different treatment methods an additional important role is played by proper masking (dummy) of intervention, e.g. simultaneous administration of two interventions that differ by administration routes, but in each study arm a different intervention is replaced with a placebo. In some cases, e.g. different medical procedures it is difficult to ensure blinding or it is associated with high burden to the patients. It should be remembered that the lack of blinding significantly affects primarily the evaluation of the subjective endpoints, independently assessed by patients (PRO, patient-reported outcome; e.g. scoring of symptom severity, quality of life) or safety analysis, but does not disturb unambiguously objective endpoints, such as death (and hence survival outcomes) [3]. During endpoint evaluation with use of pathological or imaging examination, or standardised criteria (e.g. RECIST — Response Evaluation Criteria in Solid Tumours) the risk of systematic error is ambiguous. On the other side, in oncology studies, despite the blinding of imaging tests assessing a progression (response to treatment), they are centrally confirmed by an independent and

blinded committee. However, there is a risk, especially in placebo-controlled trial with crossover after disease progression, that a lack of blinding will result in performing imaging examinations faster than planned, even upon mild symptoms. There are also clinical trials in which blinding of all researchers or patients is not required, but instead analysts evaluate the endpoints — these are referred to as PROBE (prospective, randomised, open, blinded-endpoint evaluation).

Evaluation of study quality

The simplest **assessment of study credibility** is possible with the use of the Jadad five-point scale [6]. It assesses whether the study was described as randomised, whether double blinding was used, and provides information on how many patients discontinued the study and for what reasons. Additional points can be granted or deducted depending on whether randomisation and blinding were or were not performed correctly. However, this scale allows only for very general assessment of study quality and does not take into account other factors that could result in systematic error (bias). A more comprehensive method is use of the Cochrane Collaboration recommendations [7], according to which the following aspects are assessed:

— selection bias — whether the correct method of randomisation and allocation concealment was used;
— blinding of patients and medical staff (performance bias);
— blinding of assessment of results (detection bias) — whether investigators were blinded or whether the authors of the publication justified that the lack of such blinding does not affect the assessment of a given endpoint; in the case of assessment of endpoints with different susceptibility to bias resulting from the lack of blinding, it is necessary to carry out the evaluation for each of them separately;
— incompleteness of results and loss of patients from the study (attrition bias) — the low risk of this type of error is when the data lost does not interfere with the assessment of endpoints, investigators have applied the right method of imputation of missing data (e.g. LOCF [last observation carried forward], in which for patients lost from observation, the individual values of assessed endpoints recorded during the last control visit are imputed for each subsequent time point until the end of the study), and the percentage of patients excluded is not different between the groups; in practice, it is assumed that if more than 10% of patients have been lost from the study, the risk of systematic error resulting from data incompleteness is high, unless the frequency of individual causes of exclusion is similar and the percentage of patients lost to follow-up is small;
— selective presentation of results (reporting bias) — whether the study protocol is available, and the publication presents the results for all predefined endpoints;
— other factors (other bias) — whether no other potential sources of reduced reliability of presented results were found (such as incorrect study design or the allegation of dishonesty).

It is worth noting that currently multicentre trials are preferred with appropriate representation of different geographical regions [8], although they are associated with the risks of lowering the standardisation of the interventions used as well as the results [1].

Defining the studied population

The target study population should be described in details and defined based on inclusion criteria. They are analysed to conduct an external credibility assessment, e.g. determining the characteristics of patients for whom the conclusions of the study may be generalised. Too narrow and detailed inclusion criteria may limit the possibilities of recruiting patients to the study and the possibility of generalisation of conclusions, but too general inclusion criteria can cause dispersion of the assessed effect in subgroups with different characteristics, making it difficult to randomly distribute confounders and preventing subgroup analysis [3].

Defining the comparator

Another key element is the choice of a proper **comparator (control group)**, which determines the possibility of further extrapolation of results on the target population and the study's external reliability. The optimal and desirable comparator is the current clinical practice, consistent with widely accepted recommendations and guidelines [9]. However, placebo is often used in the control group. This is justified when new therapy is an add-on treatment to the current standard (then placebo is used only for blinding, and it's the current practice that is in fact the comparator) or when there is no other therapeutic option available in real-life conditions except symptomatic treatment, e.g. when the evaluated intervention is the very last treatment line. A comparison with placebo is usually aimed at demonstrating the superiority of the new treatment. The choice of an active intervention as a comparator always brings additional challenges, also in the context of sample size, but use of placebo would be simply unethical. In the case of comparison with active treatment, testing of the non-inferiority hypothesis may be considered [3]. The rationale for the selection of active intervention as a comparator should also be assessed in the context of changing clinical recommendations, especially in the case of clinical

trials planned a few years before. In oncology, in view of diversified chemotherapy regimens and treatment methods, the investigator's choice of therapy is often accepted as a comparator. In such a situation, it should be assessed which interventions were used in the comparator group and how they were distributed, especially when symptomatic treatment is an option, and whether they reflect clinical practice and possible extrapolation of conclusions (external credibility).

Defining the endpoints

Endpoints (outcomes) should be accurately defined in the study protocol, with the specification of primary (for which sample size and statistical power are estimated) and additional/secondary endpoints. Evaluation of clinically relevant endpoints, such as overall survival (OS) and quality of life (general and aimed at evaluation of symptoms associated with a given type of cancer) is also desirable. The possibilities to assess the impact of interventions on overall survival will depend on the type of cancer and its clinical stage. This analysis will undoubtedly be difficult in the case of assessment of early stages of therapy with curative intent (e.g. neo- and adjuvant treatment), when the expected further survival could last for decades, and additional effects of subsequent treatment lines, implemented after later recurrences or progression, will have an impact on the observed differences in survival. In such cases, surrogate endpoints may include disease-free survival (DFS), event-free survival (EFS), relapse-free survival (RFS), e.g. the time since the date of inclusion to the study to the date of occurrence the first documented clinical event or death (whichever occurs earlier), for therapies used in the early stages of cancer, or progression-free survival (PFS), e.g. the time from the date of randomisation until the date of progression or death, in the advanced stages. Clinical events included in PFS/DFS definition are observed earlier than death, therefore the observation period necessary to show a statistically significant difference between the interventions is usually shorter than for OS. Hence, the PFS assessment is preferred for example when high clinical needs exist (no other effective treatment available), because the registration of the drug in a given indication can be obtained much faster (even by several years) than if it would be necessary to wait for OS outcome. In addition, the observed differences in PFS are not affected by successive treatment lines and possible cross-over because further treatment is not usually introduced before disease progression. Progression-free survival is assessed in the majority of studies with the treatment of advanced cancer stages; nevertheless, it is considered as a surrogate endpoint. There are many publications assessing the correlation between PFS and OS regarding PFS usefulness as OS predictor, although so far the conclusions presented by many authors are ambiguous [10]. There are other commonly used endpoints such as objective response rate (ORR) based on imaging tests for solid tumours or haematological remission for haematological malignancies together with the time of duration of response (DoR). As the alternative to PFS time to disease progression (TTP) is sometimes used. It differs from PFS in that, that it comprises only events of progression, while observation of patients who died before its occurrence are censored at the time of death. Related endpoints, although much less frequently used in the assessment of palliative care effectiveness, include time to treatment failure (TTF) and time to next treatment (TTNT).

Considering the diversity of evaluated endpoints, the internal coherence of presented results should be highlighted, i.e. demonstrating a significant impact of the studied intervention on ORR, PFS, and then on OS. However, the following should always be remembered: the differentiation of studied populations in terms of type and stage of cancer, prognosis and time of expected survival, and even the type of intervention used, e.g. demonstration of the impact of immunotherapy on OS, in the absence of effects on PFS due to pseudoprogression [11]. Regarding studies in oncology, particular attention should be paid to safety assessment, including undesirable or fatal adverse reactions, which in turn should include toxicity specific to the intervention. Finally, the benefit-risk ratio should be evaluated, taking into account the prognosis in a specific patient population [12, 13].

Information about planned statistical analysis

The scope and type of **statistical analysis** in a correctly performed clinical trial should be predefined as part of a previously accepted protocol, together with predefined matching factors and subgroup analysis.

The initial estimation of **sample size** (statistical power of the study) is one of the key elements of the statistical analysis. It allows assessment of whether the sample is large enough to confirm or exclude differences between interventions. The assessment of the sample size refers to the main (primary) endpoint (or endpoints). It requires the determination of the expected frequency of events in the control group, the magnitude of the intervention effect that the study is aimed to detect (an alternative hypothesis), the assumption of the ability to detect the real effect (statistical power of the study), and the selection of the statistical significance level. In oncological studies with a long observation period the expected discontinuation rate should also be taken into account. Because statistical power depends on the number of patients experiencing a given event during observation, in oncological studies it is often assumed

that patients will be observed until the occurrence of an expected number of events (e.g. deaths or deaths and disease progressions in evaluation of OS or PFS) [3].

In the statistical analysis the researchers may adopt different analytical approaches — the hypothesis of superiority is tested most often, especially in the early stages of clinical trials and in placebo-controlled trials. The second approach is based on the assessment of whether the intervention is not inferior in terms of clinical efficacy to currently used methods (non-inferiority), especially with a better safety profile. In this case, there is a need to establish a clinically acceptable variability of effectiveness in terms of primary endpoint, and if the appropriate confidence interval (CI) for the difference between interventions does not exceed the set level, the intervention is considered to be no worse than the control. The use of the non-inferiority approach reduces the required sample size [14, 15]. There are also studies assessing the equivalence of interventions, where acceptable variability is assumed in both directions, but they are rarely used to assess clinical endpoints — laboratory parameters or pharmacokinetics are used instead. **The population included in the analysis** is also important — it may vary depending on the assessed endpoint. The ITT (intention-to-treat) population is included in the analysis of the results of all randomised patients, regardless of whether they received an assigned intervention and regardless of how long they remained in the observation (this usually applies to the assessment of OS or PFS). Sometimes a modified ITT (mITT) population is defined, i.e. randomised patients who have received at least one dose of the study drug — a safety analysis is usually performed in this population. Population PP (per-protocol) refers to patients who additionally did not discontinue the treatment, did not violate the protocol, and for whom a complete set of information is available — it is often used to compare the effectiveness of interventions in non-inferiority trials [14]. ITT analysis is more conservative because it tends to underestimate the beneficial clinical effect, whereas PP analysis allows a comparison of therapeutic options in conditions of a complete observation. If the results obtained in ITT and PP analysis clearly differ, this may indicate reduced reliability of the study. The evaluation of objective response rate is often carried out in the population of patients for whom additional imaging results are available, i.e. there is a possibility to assess the progression.

Evaluation of results

The analysis of results of a clinical study begins with a detailed assessment of the description of the population included and tables with **baseline characteristics**, which should include basic demographic data, disease severity, previous treatment, and other factors that may affect the effectiveness of the assessed therapy — their scope and type depend on the type of cancer and should also be adapted to disease severity. The analysis of these parameters can be used to assess the correctness of randomisation and eliminate the influence of confounders (analysis of such a table can refer only to those known factors, but at the same time it could be assumed that random assignment to groups with appropriate sample sizes also ensures equal distribution of other, unknown prognostic factors). It is necessary to distinguish subgroups defined for randomisation with stratification and subgroups, within which predefined analysis or possible unplanned post-hoc analysis will be performed. This information is also helpful in determining the external validity of study results. It allows also the assessment of whether the analysed population is close to the one in which the evaluated intervention is to be applied [8].

Study outcomes in the form of categorical (nominal) variables are usually presented as numbers and percentages, while continuous variables are usually presented by means of a measure of central tendency and dispersion — usually mean and standard deviation (SD) values, and in the case of variables that present the normal distribution, median and range, possibly interquartile range (IQR), are used (see below). In addition, some continuous variables can be transformed into ordinal variables (e.g. the percentage of patients above a given age). Referring to the previously mentioned assessment of the accuracy of randomisation, it should be checked that there are no significant differences in baseline characteristics between the groups — the authors should provide P-values in the table or declare no significant differences in the publication text [8].

The study should also include detailed information (usually on the appropriate diagram) on **patient flow** from the screening period (i.e. from consent to participation in the study to inclusion) until a possible additional follow-up period. As was already mentioned, this is an important element of assessment of study reliability — the size of the loss of patients from observation should be assessed, as well as how it can affect reliable analysis of results, and the occurrence of differences between groups.

The next step involves quantifying the differences between interventions, exposing the uncertainty of these estimates by means of confidence intervals, and evaluating the strength of evidence, i.e. confirming by means of P-value (statistical significance test) that the observed difference is true and not by chance [8, 16–18].

Because, for obvious reasons, it is not possible to test all patients in the considered clinical conditions, a sample should be selected (a group included into the clinical trial), and, based on observed effects, conclusions should be drawn with some approximation about the

real overall effectiveness of treatment. In the language of statistics, it is said that, based on the selected measure of the effect determined in a random sample from this population, a parameter in the general population could be estimated, which in this approach comprises all possible results of the experiment.

At the beginning a **research hypothesis** should be precisely formulated and then tested with statistical methods for its acceptance or rejection. By default, it is assumed that evaluated interventions influence the health effect in a similar way (so-called null hypothesis), and the observed differences are a result of random variation derived from the limitations of the experiment (e.g. the group of patients being too small). An alternative hypothesis is that the observed differences are true and not just a random observation. The role of statistics is to indicate which of these hypotheses is more likely.

But how will we know that the groups do not differ from each other? It could be intuitively said that the lack of differences between these groups will be surely confirmed by the same percentage of patients with a response in the intervention group as in the control group. However, if there are any differences, the probability is estimated — designated as a **P-value** — to obtain a difference in treatment at least as high as that observed (in both directions, i.e. in favour or not of the intervention being analysed) in a situation in which a null hypothesis were real. If the probability of the lack of differences between groups is below the **statistical significance** threshold of 5% adopted in the biomedical sciences ($P < 0.05$), it is assumed that these differences exist and are not the result of chance, so the null hypothesis is rejected with conclusions of significant differences between the groups. In other words, this means that the probability of obtaining at least such a difference as demonstrated is less than 0.05. Thus, the lower the P-value for a given estimate, the stronger the evidence against the hypothesis of the lack of differences and the greater the conviction about the effectiveness of the intervention. Obviously, the statistical significance of the result indicates only that the observed reliance is more likely than would result from a simple random case, but it does not mean that the observed effect is real. Critical appraisal should also take into account internal reliability and the influence of confounders related to study methodology and conduction (including randomisation, blinding, and loss of patients). In addition, it is important to distinguish the difference between statistical and clinical significance and to further assess the magnitude of the observed effect in the context of prognosis in a specific population.

The uncertainty of estimates can be assessed by analysing the **95% confidence interval** (CI), assuming there is a 2.5% probability that the real effect is below and a 2.5% probability that it is above this range (such a confidence interval results from the assumption of P-value < 0.05). The accuracy of the estimation increases with the sample size: the larger the study, the more accurate the estimate and the narrower the confidence interval for the assessed parameter. With many repeated tests for effect measurement CI can be calculated in each of these tests, and 95% of them should contain a true value. The designated CI may also be used to assess the statistical significance of the result if the entire interval indicates a coherent effect, i.e. it does not contain a value indicating no differences (0 in the case of difference in continuous variables or probabilities and 1 when considering the hazard or probability ratio in both groups). The intervals that exclude these values indicate significant differences between compared groups.

Usually, a statistical evaluation will assess the three types of variables in clinical studies: dichotomous (binary) (e.g. response to treatment/no response), continuous (e.g. average body weight), and time-to-event variables used in survival analysis (e.g. OS).

Evaluation of dichotomous variables

One of the most common types of variables is the number of patients in whom the assessed event occurred or not. Usually it is expressed in the form of numbers and percentages. It should always be remembered that this is the number of cases with the first event being assessed, which is not a problem if they are unique (e.g. death) or rare. However, if they can be repeated many times, such as febrile neutropaenia, assessment of the overall number of events can be more informative, preferably calculated per observation period (incidence rate per patient-year). The number of events can therefore be analysed as a continuous or dichotomous variable.

Let us assume that patients in the study were randomly assigned to two groups of 100 patients, one receiving the active drug and the other placebo. After a year of observation, a clinical response was observed in 80 patients in the group with active treatment (80%) and only 40 (40%) patients in the control group. The authors of this illustrative publication presented four parameters, reflecting the differences between both groups in the frequency of response rate (it should be remembered that the event probability is 0.8 in the intervention group and 0.4 in the control group):

— RB = 2.00 (95% CI: 1.54–2.59); $P < 0.0001$;
— OR = 6.00 (95% CI: 3.19–11.29); $P < 0.0001$;
— RD = 0.40 (95% CI: 0.28–0.52); $P < 0.0001$;
— NNT = 3 (95% CI: 2–4).

The question is how many times the probability (risk) of a given event is higher in the intervention group compared to the control group; the answer is then a relative

parameter that in the case of a negative event is called relative risk (RR), while in case of a positive event it is called relative benefit (RB). If the frequencies of events (probabilities of their occurrence) are the same in both groups, their ratio will be 1 — this is therefore a neutral value, indicating lack of differences between interventions. Values greater than 1 indicate an increase in the probability in the intervention group, while less than 1 — the probability being lower. In the presented example the RB value is 2, so the probability of response is two times higher after using the study drug than with the placebo. It is also shown that the confidence interval constructed for this value ranges from 1.54 to 2.59 and does not contain the value 1, so it could be concluded that the differences are statistically significant, which is also confirmed by the quoted P-value (< 0.0001).

Instead of probability, the so-called chance of occurrence of a given event in each group can be calculated. Odds are defined as the ratio of the number of patients in whom the event was observed to the number of patients without such an event and hence determines how many times an event could be more frequently observed than could not. In the presented example, 80 patients responded in the intervention group and 20 did not, so the odds is 80/20 = 4 (it could be said that the chance of a response in this group is like 4 to 1). On the other side, in the control group, the chance of obtaining a response was much lower and amounted to 40/60 = 0.67. Then, by analogy with the relative benefit, the ratio of these odds values in the analysed groups could be calculated. Such a parameter is called the odds ratio (OR), and in the presented case it was shown that the chance of getting a response was six times higher in the intervention group than in the control group, which was statistically significant, as shown by the P-value and confidence interval not containing a neutral value of 1. Although the RR/RB values are more intuitive in interpretation, the publications often present calculation results in the form of ORs, which are a natural result of the statistical methods commonly used in the assessment of dichotomic endpoints (logistic regression, often also taking into account adjusting factors). It is worth noting that in the case of very high frequencies of any event in both compared arms, the OR calculation may have an advantage over RR, which in such situations will be close to 1 and will not accurately illustrate the actual difference between the groups.

In addition to the relative parameters presented above, absolute parameters can also be estimated, which are considered more informative, because they additionally show the real frequency of events — whether they are extremely rare or occur in a significant percentage of the population. Risk difference (RD) or absolute risk reduction (ARR) — in the presented case it could be called absolute benefit increase (ABI) — is a simple difference between probabilities in particular groups. In the presented example this difference is 0.40, and it could be concluded that the probability increases by 40 percentage points in the intervention group in relation to the control group. In practice, this means that for every 100 patients receiving intervention a response to treatment will be recorded in an additional 40 patients compared to the control treatment. In the presented example a confidence interval constructed for the calculated risk difference and P-value is also provided. Hence, the range (0.28–0.52) does not contain the value 0, so it could be assumed that the observed differences between groups are statistically significant.

This relationship can be also reversed, and the question could be asked, for how many treated patients one more event will occur. Such a parameter is called the number needed to treat (NNT) in respect of beneficial effect of treatment, and number needed to harm (NNH) when the event is unfavourable. The ratio shows that this number will be 1/0.40 (NNT = 1/RD), i.e. 2.5; because it refers to a number of patients, the result should be rounded up to the total number, so it shows that providing three patients with an intervention instead of control for a given time (year), one additional response could be expected [16–18]. Interpreting the results for dichotomous variables, in particular the values of absolute parameters, the observation period for a given endpoint should be taken into account. For example, NNT obtained in studies of different duration may not be directly comparable because the NNT value may vary with the observation period.

Evaluation of continuous variables

In clinical trials, the parameters determining the severity of disease symptoms, laboratory tests, or quality of life on a certain scale are often assessed. In each of these situations, the obtained results have a continuous nature, i.e. they take any value expressed in a real number from a given range. For example, patients may be asked to indicate their well-being on a scale from 0 to 100, from the worst to the best. We also have continuous results when measuring body mass, height, blood pressure, average white blood cell count, haemoglobin concentration, etc.

Such results for the studied groups of patients are usually summarised by presenting the central measure — mean or median value — as a reminder, the mean (arithmetic) is the sum of the results obtained for each patient in the group divided by the number of patients in this group, while the **median** is the middle value that divides a group of patients into half (i.e. half of the patients have a score below the median value and the other half, above). The set of results of a given effect expressed in a continuous variable is also characterised

by a certain variability that can be represented graphically as a spread of observed values around the mean. The parameter indicating the magnitude of this variation is the **standard deviation** (SD) — lower values indicate a small spread of results around the mean, while higher values indicate a large variability and large differences between the results and the mean value.

As in the case of event frequency analysis, also in the analysis of continuous data, the overall effect is assessed based on a sample from the general population. Because such sampling results in different mean values distributed around the mean value in this general population, an additional measure of this distribution can be introduced, precisely defining the distribution of the mean values from samples around the mean value in the general population; this parameter is called the **standard error** (SE) and is equal to the standard deviation in the sample divided by the square root of the sample size. Standard error decreases with increased sample size, and the lower its value, the better the approximation of the true value in the general population by the given sample.

If the study results are presented in the form of medians, the range in which the observed results are found is also usually given. As already mentioned, the "median" is a median value, i.e. which divides a series of data into two equal parts. However, we can determine several such "aliquots" of the data set, depending on the adopted criteria — in general they are called quantiles, and the median is a special case of such a quantile. The set can be also divided into four parts — then the "aliquots" are called quartiles (it is worth noting that the median is also the second quartile of the set), and in the case of dividing the set into 100 equal parts — percentiles (the median is the 50^{th} percentile of the set). Sometimes the authors present median values along with the so-called **interquartile range** (IQR), i.e. the distance between the first and third quartiles.

The reasoning when assessing the statistical significance of differences between groups for continuous variables is analogous to that carried out when describing the difference in the frequency of events in two groups — in general, average values of a given parameter should be determined in the analysed groups, and then their difference and the confidence intervals or P-value should be calculated. In case of variables with a normal distribution (also after the appropriate data transformation), Student's t- or ANOVA test is usually used to assess the differences, and in other cases, one of the non-parametric tests is used, e.g. U Mann-Whitney. Because in patients with higher values of a given parameter major changes can be expected during the test, an analysis of covariance (ANCOVA) is also used, which compares the mean values adjusted with the baseline values. As in the case of the risk difference described earlier, statistical significance can be assessed based on confidence intervals, where "0" is a value indicating no differences between groups.

The parameter most often presented in the study is the **mean difference** (MD) between the analysed groups. Assessment of the results should be carried out carefully because the authors of the study can present them in several ways. For example, when assessing the quality of life, the average score can be determined at the end of the observation period in both groups, and then the difference in mean values between them can be calculated (it is important always to make sure that no significant differences in the measurements were initially observed). It is also possible to assess the mean score change during treatment with respect to the baseline and to calculate the mean differences for such changes — this approach is used more often because it allows assessment of the effect of treatment with matching in regard to an already existing effect.

In clinical trials the **least square mean** (LSM) is also commonly used; this is simply an average adjusted for additional factors. For example, in a given group of people the average age can be calculated by summing the years of life of each person and dividing this amount by the number of persons in the sample; a simple mean value is then obtained. However, if there are many older women in this group, it could lead to overestimation of the average — in this case, the average age can be calculated first among women, then among men, and only averaging the age value for both groups the average value in the whole cohort could be obtained, with matching for gender [16–18].

Evaluation of "time-to-event" variables — survival analysis

Analysis of time-to-event data (e.g. death from any cause in OS analysis, death or tumour progression in PFS analysis) is associated with several problems. In general, the survival analysis is performed because during a sufficiently long period of observation the clinical events (progression/death) will occur in all or almost all patients. In this case, estimation of a simple parameter like RR will be useless (probabilities of events will be close to 100% in both groups). Additionally, in the long-term observation, apart from the cases marked as "with an event" or "without an event", there will also be patients who will be lost to follow-up during the study, whose state will remain unknown, or at statistical analysis they are still in observation but their future status is hard to predict. Finally, taking into account the different recruitment periods and the dates of inclusion of patients in the study, the observation will include patients with different periods of study [18–22].

The solution to these limitations is the survival analysis, considering not only the occurrence of an event, but also the time to its occurrence, and enabling the cutting-off (censoring) of patients lost to follow-up or with unknown further fate at the time of data analysis. It should be noted that the term "survival analysis" is not reserved exclusively for the assessment of overall survival (i.e. time to death) but applies also to all time-to-event endpoints (e.g. time to response, progression-free survival).

The simplest form of survival analysis is plotting of the **Kaplan-Meier curves**, based on which the probabilities of survival to a given time point and the median survival could be assessed. Then, by means of an appropriate statistical test (usually a log-rank test), a comparative assessment of time differences to the occurrence of an event between groups takes place. More advanced analysis, allowing us to take into account the matching factors (independent explanatory) affecting the survival time, is carried out using regression models, most often the Cox proportional hazard model (it should be remembered that Kaplan-Meier curves, as well as median time-to-event are still not adjusted in this situation). In general, the analysis of the magnitude and direction of differences in survival is based on the assessment of **hazard ratio** (HR), **median survival time** (until an event), and **survival probability at a specific time point** (e.g. 12-month survival). The HR value summarises the relative differences in survival between groups over the entire observation period. The assessment of differences in survival, consisting only of a simple comparison of median survival time in groups, is not sufficient to depict differences in the horizon of the entire study and can be quite misleading, especially in the case of lack of hazard proportionality (this issue is discussed later in this article).

The Kaplan-Meier curve is drawn based on the results of the study for individual patients, e.g. whether and when death occurred. However, it should be remembered that some patients "fall out" from the study for other reasons, and information about how long they live is lost. So, at the time of statistical analysis, it includes patients who live, patients who have died, and those who left the study some time ago and it is not known whether they are still alive, which is called censored observation.

The risk (hazard) determines the probability of an event occurring at a given time, assuming that the event has not occurred so far. The ratio of the hazard values estimated for the intervention and control group at a given time is called the hazard ratio (HR). Conceptually, in a simplified interpretation the HR is close to the RR, but it should be remembered that the HR includes data from the entire period of observation for survival in the study and censored cases, while the RR is carried out at a predefined time point (e.g. deaths after 12 months of treatment). Interpreting the HR value presented in the study, it is assumed that this ratio is approximately constant at any time point during the observation (assuming proportionality of hazards), i.e. if, for example, the HR value is 0.61, it is assumed that for patients in the intervention group the risk of death is approximately 39% lower than in the control group at each time point during the follow-up. This relationship can also be presented as the average prolongation of survival time by 64% (1/0.61 = 1.64) in the intervention group compared to control [18–22]. It should be noted that HR values cannot easily be translated into absolute differences in survival time — for example, in a population with low mortality, a 30% reduction of the risk of death (i.e. HR = 0.70) may be accompanied by an increase in the average survival time of 12 months, while the same relative reduction in death risk (HR = 0.70) in a high-mortality population may be associated with a much lower absolute effect (e.g. three months).

The simplest way to confirm the assumption about the **proportionality of hazards** is visual analysis of the course of Kaplan-Meier curves, assessing whether the difference between them is approximately constant and persists over time. Small deviations (decrease or increase in time differences in the course of curves) are acceptable (Fig. 1). With sufficiently long observation, especially in the final lines of treatment of advanced cancers, when events occur in all patients (with very mature data, when only a few patients remain in observation), after the initial period of maintaining differences in the course of curves their convergence can be observed (Fig. 2A). The opposite situation may occur in a population with low risk of death and expected long-term survival, when the curves can reach a flat course (plateau) due to very few deaths (Fig. 2D). Sometimes, at the very beginning of the observation, the curves intersect, which may happen because in the intervention group in the initial period the risk of complications may increase (especially if there is a significant difference between the observed procedures — e.g. surgery with chemotherapy vs. conservative treatment and chemotherapy), and the expected clinical benefit is only observed in the further period when the curves separate (Fig. 2B). In general, if the variability of the course of curves during the observation relates to the magnitude of the effect, but not its direction, it could be considered that the deviation from the assumption of hazard proportionality is insignificant, and the presented interpretation of HR could remain. However, if there is a significant change in the direction of action (Fig. 2C), the calculated HR value cannot be interpreted because it changes significantly over time. One solution is an attempt to perform subgroup analysis in order to detect the cause of differences in effectiveness over time, with some objections related to such an analysis (see below) [18–22].

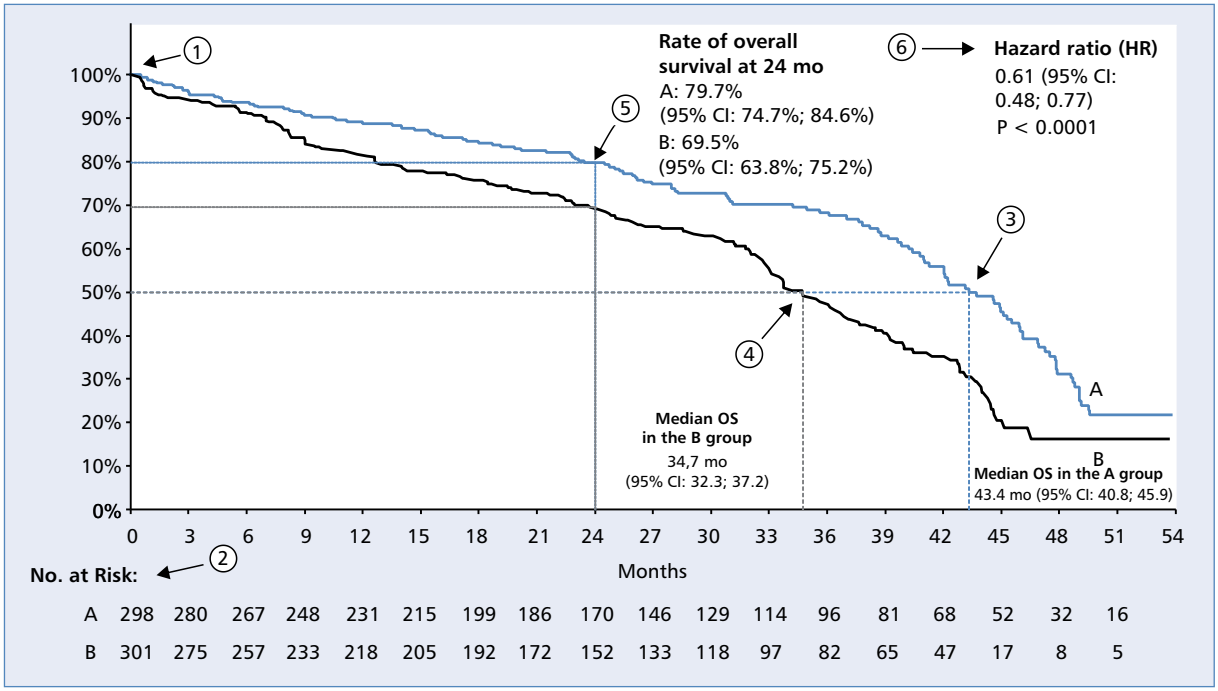


Figure 1. The Kaplan-Meier chart is located on the plane bordered by the Y-axis describing the probability of overall survival and the X-axis, on which the time from the beginning of treatment (observation) in the study is presented. While constructing the Kaplan-Meier curve, the subsequent time intervals take into account the number of patients with the possibility of measurement at the beginning of the interval (at-risk), the number of patients with the event, and the number of patients lost to observation. The graph shows the cumulative probability at a given time. At the start of observation (①) all patients are alive (OS = 100%). In fact, this is not the same moment in time (date of randomisation) for each of the participants, because they enter the study in different centres at different times. Then a decrease in the likelihood of overall survival in both arms over time is observed; however, it is always higher in group A (despite some variability in the size of differences between groups, the direction of the effect is consistent and it could be concluded that the deviation from the assumption of proportionality is slight — see text). Below the graph the number of patients in the observation at regular intervals should be given (at-risk) at the beginning of the given time interval (②). If at the end of the graph they are small (< 10% of the baseline value), the inference from the curves is limited in this section. Looking at the Kaplan-Meier graph, values from both curves can be compared horizontally, looking for a difference in time when the cumulative probability of survival reaches 50% — (③) and (④), and they provide median survival values for groups A and B, respectively (median OS in the A group, median OS in the B group), and the median difference is 8.7 months. Differences in survival can also be assessed vertically, comparing the survival values at a given time point. In our example, the two-year survival rate (rate of overall survival at 24 months) is 79.7% in the intervention group and 69.5% in the control group (⑤). This is the cumulative probability for this period; often its value is given with the confidence interval (which allows us to assess the accuracy of the estimation), and in the text a statistical evaluation of the result will usually find (P-value for differences in cumulative survival at this time point). Concluding the differences in survival throughout the observation period gives the relative hazard ratio (⑥); it can be seen that at a given time point the risk of an event (death) is lower in the group with intervention, and the result is statistically significant (looking at both the confidence interval and the given P-value)

It should be remembered that HR is a relative value that allows the assessment of the statistical significance of observed differences in survival, but, as mentioned, when making a therapeutic decision it is also necessary to assess the clinical relevance, also in relation to prognosis in a given population. The absolute impact of the intervention compared to the control can be assessed by analysing the differences in the medians or by comparing the probability of survival at a given time (e.g. an annual or two-year survival).

During the RCT, especially with a long period of observation, **preliminary (interim) analyses** carried out by independent researchers in an unblinded but confidential manner are necessary. An independent, committee with no affiliation to the study may decide to prematurely terminate the study, e.g. due to safety concerns or the spectacular effect of the new intervention (in this situation the decision to stop should be assessed carefully, because the clear effect is often overestimated in the short period of observation and the differences

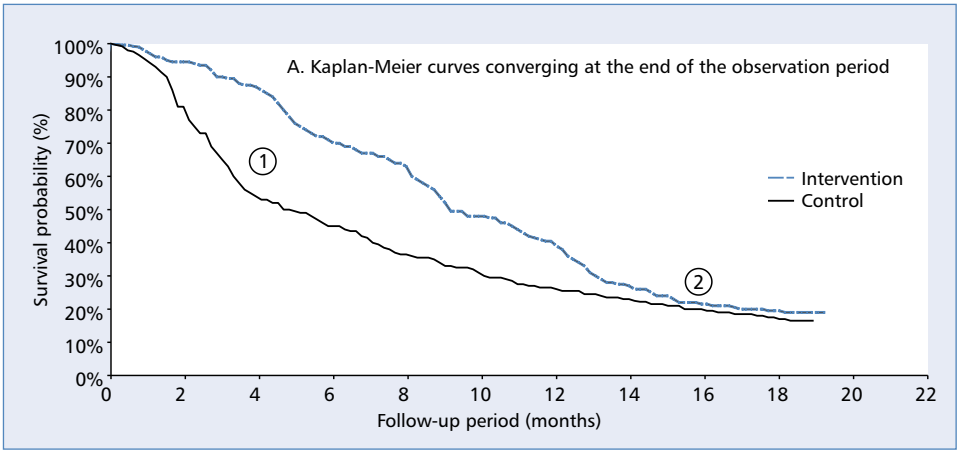


Figure 2 A. Kaplan-Meier curves converging at the end of the observation period. From the very beginning of the observation, there are differences in survival (①). The curves clearly "separate", but at the end of the observation period (②) the cumulative probability of death is practically the same in both arms of the study. Such a situation may occur, for example, at the terminal stages of cancer, where ultimately, regardless of the treatment used, an event (death) will occur in almost all patients. This is a permissible deviation from the assumption about the proportionality of hazards. Convergence of curves in the final observation period may also be caused by a high percentage of censored observations (i.e. a small number of at-risk patients), as a consequence of which the Kaplan-Meier estimation in the "tail" of the curve is impaired

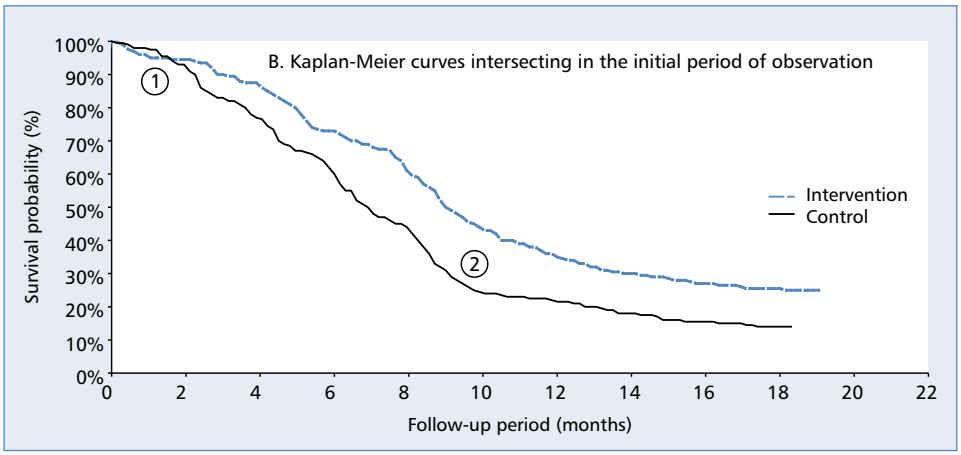


Figure 2 B. Early intersection of the Kaplan-Meier curves. In the initial period of observation, the curves intersect, and for a short period the survival is lower in the group with intervention compared to the control — e.g. when the intervention is associated with higher risk of complications (①), but then there are clear differences in survival in further observation (②). In this situation, it is also a permissible deviation from the assumption about the proportionality of hazards

between interventions become less visible in the longer period of observation) [14]. In statistical protocols of oncology studies further interim analyses are also predefined (when the statistical power calculation depends on the occurrence of a given number of events). Because of repeatedly testing the hypothesis, the risk of accidentally observed "statistically significant" results increases (the greater, the more pre-planned interim analyses, e.g. according to the O'Brien-Fleming criteria); it is worth noting that significant results will not refer to P-values of < 0.05, but assume a much lower threshold, e.g. < 0.001.

Finding significant OS differences between interventions according to the assumed statistical power may enable patients after disease progression to move from a control to an intervention group (**cross-over, treatment switching**), which acts in a conservative direction, overestimating the effectiveness of control intervention and reducing the estimated effect of the study drug. Obviously, this only matters for OS assessment, and PFS assessment itself is unaffected. In this case, it is possible to use appropriate methods of correction of the cross-over impact on OS, the simplest of which is

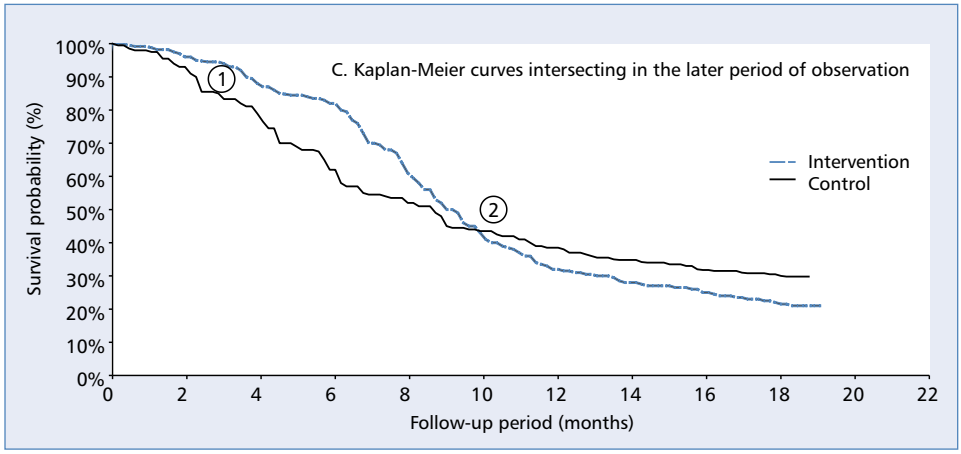


Figure 2 C. Kaplan-Meier curves intersecting in the later period of observation. The initial advantage of intervention ((1)) disappears unexpectedly during the observation ((2)), and the survival remains higher in the comparator arm until the end of the observation period. In this case, the proportionality of hazards criterion is not met, and no differences can be indicated between the groups. Probably there was an unknown confounder in the study that reversed the inference, and a detailed subgroups analysis is needed to identify it

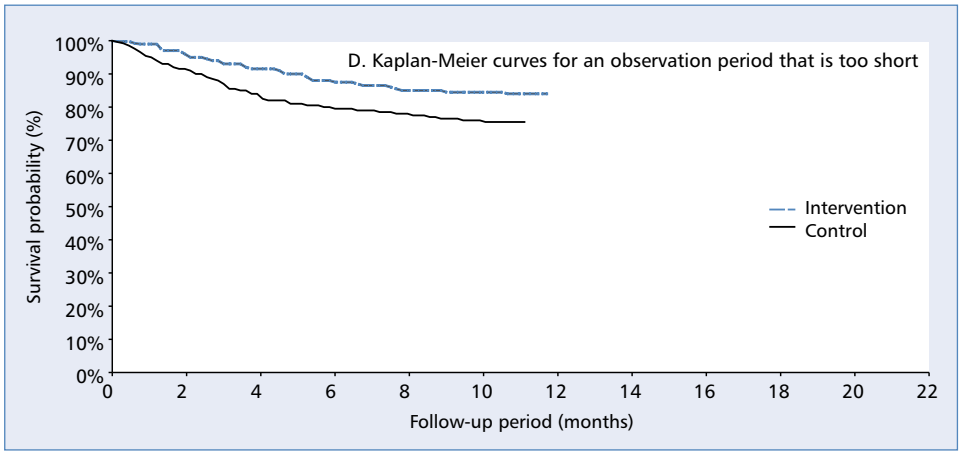


Figure 2 D. Kaplan-Meier curves for an observation period that was too short. Mortality persists at a relatively low level, and it is difficult to conclude about the fate of patients and differences between groups (assessment in the early stages of cancer, with expected long-term follow-up). Despite the initial differences with a longer observation period, when more observed events accumulate, the course of these curves may approach any of the previously described situations

the censoring of observations in patients changing the treatment [24, 25].

Adjusted results and subgroup analysis

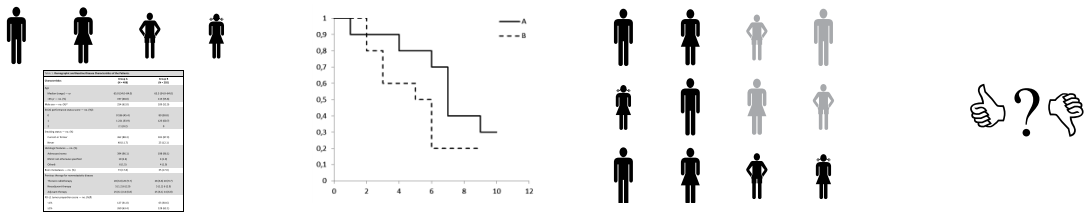
Analysis of results with adjustment to baseline characteristics within logistic regression (when OR values are presented as adjusted or multivariate) or Cox proportional hazard model (in the case of survival analysis) can be presented as primary or sensitivity analysis. It is worth noting that the adjustment should primarily include prognostic factors and — unless it usually affects the accuracy of the estimate — can modify the central measure of the estimate. However, taking into account the characteristics not related to the prognosis, although even including stratification factors (e.g. geographical location), does not significantly affect the results. Post-hoc selection of matching factors (not defined in the statistical plan) may raise the suspicion of purposeful data selection to achieve the desired effect; in this case it should always be expected to display adjusted and not adjusted results [23].

Subgroup analysis should also be predefined in the statistical plan. Considering the diversity of the general study population, it allows the assessment of whether the general results refer to all patients or whether there

are differences in the effectiveness of the intervention. If it is the case, it should be remembered that reliable evaluation of statistical differences leads to loss of statistical power. On the other hand, multiple defined subgroups and repeated testing increase the risk of completely random occurrence of statistically significant results. Above all, statistical significance or lack thereof in one of the subgroups is not sufficient to conclude real differences in the effectiveness of intervention. Usually a coherent effect is observed in individual subgroups, although with some variability of the central measure, and in smaller groups the confidence intervals become wider and, in some cases, may exceed the value of 1 (loss of statistical significance). In this situation attention should be paid to the significance of the **interaction test** (statistical analysis of whether the impact of interventions on an observed result depends on other factors) to assess whether there is a difference in the effectiveness of the intervention in a given subgroup (remembering, however, the possibility of obtaining false results in repeated testing). Subgroup analysis may also be helpful in seeking a narrowing target population with no significant result in the general cohort. However, it should be remembered that subgroup analysis is more exploratory and serves to create further hypotheses rather than make final conclusions [23].

Summary

Translating the results of clinical trials into clinical daily practice is the established method of EBM. For this purpose, however, many elements should be assessed, which together provide evidence of the reliability of the study and significance of its results (Fig. 3). It is not easy, especially since most of the publications are written in English, and the authors often assume in advance that the recipient is fluent in terms of statistics and detailed explanations are unnecessary. In addition to assessment of the methodology and the reliability of the clinical study, it must be ensured that the population being evaluated is representative, i.e. it has characteristics similar to those for which a therapeutic decision is to be made, and if there are discrepancies (e.g. different age of patients or presence of comorbidities), what is their meaning. Then, if the comparator in the study is not widely used or not available but has a similar mechanism of action to the current treatment standard, it should be assessed whether there is evidence of similar effectiveness, which would allow transferral of the inference from the study to clinical practice in this aspect. The sample size of the study is significant — if it was small and it was not due to the low prevalence of the disease, then it should be assessed whether the study had statistical power to indicate the differences.



Methodology	Primary results	Additional analysis	Others
• Baseline characteristics • Allocation, intervention, control • Comparator and local clinical practice • Research hypothesis • Randomization method, blind, open label, unmasked • Drop-out, discontinuation • Internal credibility	• Statistical significance, hazard ratio for progression-free survival, overall survival — P-value and confidence interval • Course of Kaplan-Meier curves • Median OS and PFS • Other end points — convergence of inference with primary endpoints — internal consistency (time to treatment discontinuation, objective response, complete response) • Quality of life, safety, adverse event	• Intention-to-treat, per protocol analysis • Subgroup analysis — convergence of inference with the ITT population • Statistical significance in subpopulations (interaction test) • Interim analysis, cross-over	• Subsequent treatment • The authors' conclusions — unambiguous or conservative • Consistency of results with other tests (external coherence) • The possibility of transferring conclusions to clinical practice (external credibility)

Figure 3. Aspects to which attention should be paid, making a critical evaluation of the methodology and results of a clinical trial in oncology

It is also necessary to pay attention to possible differences in baseline characteristics between groups — if they were significant, they may indicate a selection bias. An incorrect randomisation method can cause imbalanced distribution of confounders. In the absence of blinding, the assessment of differences in subjective endpoints is subject to the limitations. In the case of significant differences in the lost patient rate, it is important to know whether this could be related to the treatment applied. It is also worth checking the type of the research hypothesis tested. In oncology the primary endpoint will be the survival analysis — PFS and OS. If consistent, statistically significant, and clinically relevant differences are observed in favour of the intervention, there are strong premises about the superiority of the evaluated therapy. If, however, the significance of the results was observed only for PFS, it should be decided whether a further (final) assessment of the OS is planned, in which, with more matured data, the result could reach statistical significance. For some cancers, especially when assessing their early stages, it may be difficult to show differences in survival due to the expected follow-up, sometimes even decades. Such a long observation is an additional challenge, because during this time patients may be subjected, for example, to many different lines of further treatment, and the evaluation of ultimate survival differences is limited (in this situation, such endpoints as DFS or pathological response are of higher importance). It is important whether the possibility of changing the treatment after the progression (cross-over) was allowed, which could lead to OS overestimation in the control group. When HR for both OS and PFS didn't reach the significance level, the question should be asked whether this is not due to the lack of study statistical power, immaturity of published results (interim analysis), or high lost-patient rate. If none of these factors is relevant, there are probably no differences between the interventions. Assessing the consistency of results with other publications for a similar population may be very helpful. If PFS/OS medians differ from similar studies, the characteristics of the population should be carefully analysed. A valuable source of information is also subgroup analysis. Sometimes the result for the subpopulation becomes statistically significant, despite the lack of significance in the total population — this may be a premise of higher treatment effectiveness only in a specific subgroup, but on the other hand, one should bear in mind the exploratory nature of such analysis. If a similar trend in the results is observed depending on the presence or absence of a given criterion, but one of the subgroups lacks relevance, it is worth checking whether the size of this subgroup is not too small, as well as checking the result of the interaction test. If any therapeutic options are available in the analysed indication, the safety issues of the therapy being evaluated are extremely important. In oncology, treatment with higher efficacy is often associated with increased toxicity; this situation may be acceptable with clear clinical profit, such as prolonged survival.

In conclusion, the evaluation of a clinical trial consists of many elements discussed briefly in this paper. The authors hope, however, that they have addressed the most important aspects of the evaluation of clinical trial results and the terminology used, and that the article managed to show the complexity of the interpretation process. In the second part of the work, examples of clinical trials will be presented along with an assessment of their credibility and impact on clinical practice.

References

1. Govani SM, Higgins PD. How to read a clinical trial paper: a lesson in basic trial statistics. Gastroenterol Hepatol (N Y). 2012; 8(4): 241–248. PubMed PMID: 22723755; PubMed Central PMCID. PMCPMC3380258.
2. Bothwell LE, Greene JA, Podolsky SH, et al. Assessing the Gold Standard — Lessons from the History of RCTs. N Engl J Med. 2016; 374(22): 2175–2181, doi: 10.1056/NEJMms1604593, indexed in Pubmed: 27248626.
3. Pocock SJ, Clayton TC, Stone GW. Design of Major Randomized Trials: Part 3 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(24): 2757–2766, doi: 10.1016/j.jacc.2015.10.036, indexed in Pubmed: 26700838.
4. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008; 336(7644): 601–605, doi: 10.1136/bmj.39465.451748.AD, indexed in Pubmed: 18316340.
5. Poolman RW, Struijs PAA, Krips R, et al. Reporting of outcomes in orthopaedic randomized trials: does blinding of outcome assessors matter? J Bone Joint Surg Am. 2007; 89(3): 550–558, doi: 10.2106/JBJS.F.00683, indexed in Pubmed: 17332104.
6. Jadad A, Moore R, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Controlled Clinical Trials. 1996; 17(1): 1–12, doi: 10.1016/0197-2456(95)00134-4.
7. Higgins JPT GSe. Cochrane Handbook for Systematic Reviews of Interventions Version 5.2.0 [updated June 2017]. The Cochrane Collaboration, 2017. Available from: http://handbook.cochrane.org.
8. Pocock SJ, McMurray JJV, Collier TJ. Making Sense of Statistics in Clinical Trial Reports: Part 1 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(22): 2536–2549, doi: 10.1016/j.jacc.2015.10.014, indexed in Pubmed: 26653629.
9. Health Technology Assessment Guidelines. The Agency for Health Technology Assessment and Tariff System. 2016. Version 3.0:[Available from: http://www.aotm.gov.pl/www/wp-content/uploads/wytyczne_hta/2016/20161104_HTA_Guidelines_AOTMiT.pdf.
10. Li L, Pan Z. Progression-Free Survival and Time to Progression as Real Surrogate End Points for Overall Survival in Advanced Breast Cancer: A Meta-Analysis of 37 Trials. Clin Breast Cancer. 2018; 18(1): 63–70, doi: 10.1016/j.clbc.2017.07.015, indexed in Pubmed: 28818493.
11. Panasiuk AWR, Budasz-Świderska M, Kaczor M. Cancer immunotherapy in second-line treatment of non-small cell lung cancer — is there a need to change the approach to the assessment of clinical benefits? Journal of Helath Policy & Outcomes Research. 2017; 2(2): 65–77, doi: 10.7365/JHPOR.2018.1.9.
12. Wilson MK, Collyar D, Chingos DT, et al. Outcomes and endpoints in cancer trials: bridging the divide. Lancet Oncol. 2015; 16(1): e43–e52, doi: 10.1016/S1470-2045(14)70380-8, indexed in Pubmed: 25638556.
13. Wilson MK, Karakasis K, Oza AM. Outcomes and endpoints in trials of cancer treatment: the past, present, and future. Lancet Oncol. 2015; 16(1): e32–e42, doi: 10.1016/S1470-2045(14)70375-4, indexed in Pubmed: 25638553.
14. Pocock SJ, Clayton TC, Stone GW. Challenging Issues in Clinical Trial Design: Part 4 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(25): 2886–2898, doi: 10.1016/j.jacc.2015.10.051, indexed in Pubmed: 26718676.

15. Mauri L, D'Agostino RB. Challenges in the Design and Interpretation of Noninferiority Trials. N Engl J Med. 2017; 377(14): 1357–1367, doi: 10.1056/NEJMra1510063, indexed in Pubmed: 28976859.
16. West CP, Dupras DM. 5 ways statistics can fool you — tips for practicing clinicians. Vaccine. 2013; 31(12): 1550–1552, doi: 10.1016/j.vaccine.2012.11.086, indexed in Pubmed: 23246309.
17. Elwood JM. Interpreting clinical trial results: seven steps to understanding. Can Med Assoc J. 1980; 123(5): 343–345, indexed in Pubmed: 7260774.
18. Case LD, Kimmick G, Paskett ED, et al. Interpreting measures of treatment effect in cancer clinical trials. Oncologist. 2002; 7(3): 181–187, indexed in Pubmed: 12065789.
19. Pocock SJ, Travison TG, Wruck LM. How to interpret figures in reports of clinical trials. BMJ. 2008; 336(7654): 1166–1169, doi: 10.1136/bmj.39561.548924.94, indexed in Pubmed: 18497415.
20. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010; 1(4): 274–278, doi: 10.4103/0974-7788.76794, indexed in Pubmed: 21455458.
21. Fendler WCJ, Młynarski W. Methods of survival analysis applied in oncology — assumptions, methods and common pitfalls. Onkologia w Praktyce Klinicznej. 2011; 7(2): 89–101.
22. Barraclough H, Simms L, Govindan R. Biostatistics primer: what a clinician ought to know: hazard ratios. J Thorac Oncol. 2011; 6(6): 978–982, doi: 10.1097/JTO.0b013e31821b10ab, indexed in Pubmed: 21623277.
23. Pocock SJ, McMurray JJV, Collier TJ. Statistical Controversies in Reporting of Clinical Trials: Part 2 of a 4-Part Series on Statistics for Clinical Trials. J Am Coll Cardiol. 2015; 66(23): 2648–2662, doi: 10.1016/j.jacc.2015.10.023, indexed in Pubmed: 26670066.
24. Panasiuk AHM, Pawlik D, Prząda-Machno P, et al. Approach to uncertainty in health technology assessment in a Central and Eastern European country: appraisal of cancer drugs by a Polish HTA agency in presence of high crossover rates in clinical trials. Journal of Helath Policy & Outcomes Research. 2016; 10(2): 17–34, doi: 10.7365/JHPOR.2016.2.2.
25. Haslam A, Prasad V. When is crossover desirable in cancer drug trials and when is it problematic? Ann Oncol. 2018; 29(5): 1079–1081, doi: 10.1093/annonc/mdy116, indexed in Pubmed: 29648572.