Subtitles and transcriptions
Subtitles and transcriptions are available for selected materials for purpose of helping users understand the contents of the educational sessions.
Uncertain words have been indicated with ?? before and after the part. Parts that could not be understood at all have been indicated as [Audio Not Clear].

Every effort has been made to faithfully reproduce the audio of the sessions as recorded. However, no responsibility is accepted for mistakes or omissions. ESO does not endorse any opinions expressed in the presentations.

ABC of Statistics

Prof Dafni: Hello, everybody. I'm here to talk to you about ABC of statistics and I'm thrilled to discuss the basic of statistics in this forum. And have Dr Vadim Lesan asking his questions so that we can make it more interactive. I have to say that we all use statistics and I'm sure that in research, you have encountered both the need to estimate the parameter and use inference in testing hypothesis. Descriptive statistics are the first important step that give us the knowledge of a distribution based on observed values. While with inference, we can actually ask questions about certain hypothesis. For example, to explore association of variables of interest, like what is the effect of smoking in the incidence of lung cancer? One would need an epidemiological study to answer this question. While when we are interested on the effect of a treatment on outcome, we would prefer to have a clinical trial and intervention in a clinical trial. What are the descriptive statistics, first of all? Well, they're used to describe or summarise data in ways that are meaningful and useful. And they are the heart of all quantitative analysis. So, how do we describe data? As I said earlier, we're talking about the distribution of values. So, what matters to us is to have a measure of the centre of the distribution. These are called measures of central tendency and a measure of the variability. How far, what is the range, how far from the centre the actual observed values are? So, there are two types. Measures of central tendency and measures of variability or dispersion. When we're talking about the distribution, again, if we were to have the most appropriate measure of central tendency for a specific data set and we have a symmetric unimodal distribution like the one shown here, we could use to describe the centre of the distribution, either the mean, the average, the median or the mode. And in all cases, in the ideal case of the symmetric unimodal distribution, we will have end-up with the same value or in the real-world, very similar values. If we were to have a symmetric bimodal distribution, you can see that using the mean or the median which would end up in the middle here, won't really represent the centre of gravity, meaning where we have most of our observations. Instead, it is better to report the two modes of these two distributions. If we have asymmetry, either left or right asymmetry, then the median is preferable because again, the median in this case will be closer to where the higher height is of the distribution, which would mean, again, it'll represent the point where we have more of the values. But what can happen if we have here, as you can see, two identical as far as their centre distributions? They're symmetric, they're unimodal, and they do have the same mean median and mode, but as you can see, they're very different. Why? Because one of them, the lower height one is spread out. So, there is more variability, their values here all along this continuum that go further away from the middle, so, have bigger variability, while the high distribution is really having all the values very close to the middle. So, to describe the distribution apart from the centre, we would like to describe the variability, the dispersion. Also, another way of describing what's happening with the distribution of values is to use plots. And here, you can see if we have a discrete variable, a variable that takes very specific few kinds of categories as values, you can use a biplot. What is in of interest here is the percentage of patients with adenocarcinoma, squamous or other histology in non-small cell lung cancer. And you can see directly from the plot, over 50% are at the adenocarcinoma, for example. A very simple way of looking at the distribution for a discrete variable. In the middle here you can see a histogram, a very useful again way of describing the distribution for continuous variables. Here for example, we see the Q mutation burden values and you can see the range from zero to close to 80. Here it says that the median is close to 8, the mean is around 16 for the 78 observations that we have available. And finally, here, on this slide, you can see a plot that again is very useful because it compares the distribution of the LDL value in patients that took tamoxifen versus patients that took exemestane. And here, you can read very well where the median is around a little less than 150, not much difference between the two distributions as far as their centre, but difference as far as their variability. You see that for tamoxifen, the range of values is much bigger. Finally, for another plot that you have seen a lot is the Kaplan-Meier plot, the survival plots that are all in all manuscripts. And here, what we're using for describing the survival distribution is the median survival. And what does that mean? If we were to look at the blue line, which represents patients on one treatment and the red line that represent patients on another treatment, we can see that the patients taking the blue treatment are actually having a median survival close to 10 months while the others on the red treatment, 25 months. So, the way to actually see this from the plot is just getting a 50% line across the horizontal axis. Again, another example where you see each of these points are a measurement and, in all cases, we do have the same average, 15.5, but the standard deviation which shows the variability of these three distributions is pretty different. Well, where do we use that? First of all, we use it when we describe, for example, the baseline patient characteristics in the manuscripts. For the age, the randomization here, we see in years, 68 is the median and the range from 34 to 85, overall, for the osimertinib bevacizumab group, while for the osimertinib group is 66, we range from 41 to 83. I wanted you to see that all the rest are just discrete variables. Sex, for example. Male, female. All we need to know is the percent. Ethnicity, the percentage. So, it's a different kind of approach when we describe discrete variables versus when we describe continuous variables. Also, a pause here because what is important to remember is that what we want to estimate is the truth, but what we'll have in our hands is the result of one clinical trial. So, for example here with 50.000 simulations, we have simulated 50.000 trials and we wrote here under from .4 to 1.1 we observed hazard ratio. This 50.000 observed hazard ratios that come from a true hazard ratio of .7. So, you can see that the median, the centre of the actual distribution will be the .7, the true value. But we would have observed many more values around .7, but even some values very far away. Now, these simulated trials, in fact, were using a power of 80%. It meant that we designed them so that in 80% of cases, if the truth was hazard ratio of .7, then we would be able to reject the null hypothesis that the hazard ratio was equal to 1, 80% of the time. And you can see here, how close you get with the simulations, it's 79.6%. And what does that mean? That any observed value that was actually lower than .78 gave a statistically significant result. We rejected the null hypothesis of no difference between two treatments and the truth was that we actually had a 30% reduction in the survival, hazard ratio of .7. We can research associations with many different ways, types of studies, starting from case reports, case series, database analysis, then moving on to the observational studies, what we call epidemiological studies. But the more important, the higher quality studies are the clinical trials. Why? Here, the treatment assignment is by design and then points and the analysis are planned in advance. So, everything we gather is prospective and by design. The first question to ask as far as quality of the results that we see and how much we believe are the actual conclusions in the manuscript is whether there was an intervention, in which case we're talking about clinical trials or no, in which case we're talking about observational studies. And not only that, but beyond that, whether there was randomization in the interventional study, in the clinical trial. And if yes, we have a randomization then we're talking about the higher level of evidence. As you can see, the blinded randomised clinical trials, the randomised clinical trials or the meta-analyses that come from them, which is a summary of the descriptive results that we have from each of these trials, are the higher as far as the quality of evidence followed by the non-randomized clinical trials. Below them are the observational studies with the prospective ones being higher and so on and so forth till we reach the case reports. Now, I would be very happy to answer questions at this point. There is one question in the chat. Marina Puchinskaya was asking how to choose what data to present for ranges. The upper and lower values or the one should choose maybe the interquartile range.

Prof Dafni: Okay, this is a very good question. If we have a symmetrical distribution, usually, mean and range allocate to report. But whenever you have outliers, both the mean as a measure of dispersion is not you know, the preferable way of presenting the centre. As we said in cases of asymmetry, median is really representing better the centre of the distribution. The same way the range is very much affected by outliers, you know, a very high point and a very low point. Although the centre of the distribution might be, you know, in much less range. In this case it is good to use the interquartile range.

Prof Lesan: Yeah, thank you. And the second question also from the same participant Marina Puchinskaya, she is asking if we should use the Shapiro-Wilk test to assess the type of distribution to choose the type of descriptive statistics.

Prof Dafni: Yes, this is a good way of getting, you know, to the conclusion of what you should use.

Prof Lesan: I also recommend maybe the Kolmogorov test. It's also One example.

Prof Dafni: Yes, yes. Absolutely.

Prof Lesan: But looking at the data is very important.

Prof Dafni: Yes, I see your answer here. You are absolutely right, Vadim. The first thing we do before running any test or any association is look at the data.

Prof Lesan: Thank you. There are no other questions, so I think we can.

Prof Dafni: Okay, let's move on then and if you think of any questions on what we have discussed so far, you can give it next also. Click on the Q&A button. No, how do I go? Ah, okay. So, in the clinical trials, as you very well know, after all the preclinical work has been done and there is a reason to move-on to believe that we're having a treatment that is worth exploring through clinical trials. We go to a phase I study, which is mainly a safety and a PK, pharmacogenetic study and with as few patients as possible because this is the first time we go to patients, then we move on if we're successful to a phase II study with a dosage usually that has been already chosen at the phase I. And here, we have actually started having bigger and bigger studies that we start actually looking at the first measure of effectiveness. But of course, the deciding step is the phase III study where we do measure effectiveness in big enough number of patients and comparing them, the new treatment to the standard therapy or if the standard therapy is placebo, to a placebo. This is where we know whether the new treatment is better than the standard therapy and should be given to the patients after successful completion of the study and the approval of the authorities. Phase IV, you might know also, that are to monitor the long-term side effects. Side effects that are very rare and you couldn't get in the few thousand patients that maybe a phase III trial had already shown. So, what is important here to remember is that whenever we're talking about clinical trial, we should really look at the declaration of Helsinki and follow the ethical principles because we're talking about experiment, a controlled experiment, an intervention with human subject. So, you can always visit this site. Now, the basic concepts that are really involved in an experiment, but specifically in the clinical trials we're discussing here, have to do with the variability and bias that we need to really address when estimating or making inferences. And I would like to talk a little bit about the difference between non-inferiority and superiority trials because a common mistake is to take a result of a superiority trial and wrongly assume if it is a negative trial that we're talking about equivalence or non-inferiority and you'll see why this is not correct. And if we have some time, which I hope we do, we'll talk about prognostic and predictive factors and what is the difference between them and the multiplicity problem. So, going back to the fact again that clinical trials are experiments, we want to answer a scientific question by isolating the intervention and the outcome from extraneous influences. And the goals are, first, to minimise the random error. This is to increase the precision, the decrease, the variability of what to estimate because results are always inaccurate due to sampling, right? I mean we're not taking the whole population; we're using a sample. Then, to eliminate the systematic error, the bias, which is any effect that would render the observed results not representative of the treatment effect. And finally ensure the generalizability of the study results by having the correct inclusion and exclusion criteria in our trial. So, very quickly, all of this can be addressed only with one-way correct study design. This is the only way to achieve these goals. Of course, you know, conducting correctly the trial and analysing the trial and reporting the trial is as important but the first step that you cannot correct for is the study design. I would like you to think of this little red rectangle in each of these circles as the truth. So, let's suppose that this is a true hazard ratio. Like we saw before, the hazard of .7 being the truth. Let's suppose that in a trial when compared to treatments, the correct, the true hazard ratio is .7. And this is represented by this little red rectangle. Then what's going to happen is if we have a big enough trial to have big precision in our estimate, which means small variability and if we have correctly addressed the issue of confounders without having any bias, then we can have one of these results. Every little dot is an observed hazard ratio from a trial. So, what is happening? If we were to do repeatedly these trials, we would expect a hazard ratio, an observed hazard ratio, which would be centring on the red triangle as the red rectangle, which is the truth. And it'll be close, the values will be close to the true value. Here we see no bias again. So, the centre, the observed values of the different clinical trials are correctly cantering around the truth. But we have larger variability because we haven't really used enough sample size. Here, we have a case of bias with small variability. This is one of the worst-case scenario because it means you used a lot of patients but you had the bias study. You didn't correct for the bias. And of course, this is bias and large variability, small study and bias not addressed. So, let's see, how do we address bias, to be sure that we address bias? When we have non-randomized studies, single arm, no control arm, we have a population of patients that are eligible to participate in the trial and, of course, we can have a single arm trial for treatment B, a single arm trial for treatment A. If we end up here having different outcome in each of these two trials, we want to know if it is because of the different characteristics of the patients that entered trial A, trial for treatment B, and trial for treatment A or it is because of the treatments. While if we do randomization, starting again with a population of patients that are eligible to participate in the trial, we centrally decide whether the patient will enter the control arm and the experimental arm. And then, of course, as patients come in randomly, we balance out their characteristics. So, the randomization tends to balance the distribution of important characteristics that might influence the outcome. And to make sure it balances characteristics, not only that we have thought of and we can show on our baseline table that are not different let's say in our manuscript, but also, characteristics that might influence the outcome and you have no idea what they are. So, this is very important. It's the only design, the randomised one, that would allow us to reach a valid conclusion relative to the existence of a treatment benefit. Why? Because if we see a difference in the outcome, we can claim that this difference in the outcome is derived from the only difference that we have allowed between groups. And this is the difference in the intervention. I hope this is clear. So, to address bias of the estimate, the design needs to include at least randomization. I'm saying at least because there are other additional ways of satisfying that we don't have bias and this includes certification or blinding. But what is the item that should be there is randomization. To address no variability of the estimate, as I alluded already, the design needs to include the adequate sample size. And I want, again, to go back to the ethical question because the sample size determination is foremost an ethical issue. Why? Because if we actually have more patients than need it, then we violate the individual human rights of the specific patients. Why is that? Because in fact, let's think that we have a clinical trial that you can actually reach our conclusion that the new treatment is better than the old treatment, the standard therapy with 500 patients. And instead, we have this trial with 1000 patients. What does that mean? That after the first 500 patients that could have had the result, right? And know that the new treatment is better, we randomised 500 patients more, which means half of them, 250 patients, will take an inferior treatment while we could have known and we would have given it to them. So, it's really important not to think that if you have a huge trial, we're ethical, right? Because in going further up than what we need is, again, not the ethical thing to do. Going with less patients, obviously what this does is that if there exists difference and the new treatment is better, we don't have the ability, the power to see the discrimination and the difference between the new treatment and the old, which means that we're really violating the human rights of the whole society because they end up not being able to receive, the patient, a better treatment. So, it's ethical to use the appropriate sample size. Now, this is a little difficult topic, but I want you to understand that the sample size comes from some very important items. First of all, the clinically significant difference. One needs to know, what is the clinically significant difference? The difference that would make a difference clinically. One shouldn't try to show a very minute difference in outcome, for example, I don't know, two weeks of survival, you know, one year and one year and two weeks because then, it might be different but it's not clinically significant, right? And we waste a lot of effort and patients in trials to do this as to claim that this is a significant difference even though it is not clinically meaningful. And then, is the power, the probability that we actually have from the beginning to show that this clinically significant difference exists, if it exists, usually, we have 80%, 85%, 90% in our trials probability of showing this difference if in fact, exists, okay? And the error of not showing this is the β, right? So, if it is 80% power to show that there is a difference when it exists, we have a risk, a probability of 20% not to be able to show it, right? And then, there is a significance level, which is the probability of wrongly rejecting the null hypothesis and saying that a difference exists. This is very commonly used as a 5%, as you probably know. So, in the phase III parallel trial design randomised trials, in all experiments in fact but for the one we're talking here, we have a null hypothesis and an alternative hypothesis and it's interesting to note that we call alternative hypothesis what we want to prove. So, it is the statement that we would like to prove if it is true. And what remains is a null hypothesis, which is a statement we would like to reject. Because if we reject it, then we'll accept the alternative. This is generally the statement that we do not want to be true. And then we have the superiority trials, which means that the null hypothesis we need to reject is that there is no effect, no difference. The alternative hypothesis is that there is a difference, right? And if we accept that, we have a significant positive trial. On the other hand, if what we're interested in has to do with no effect, no difference, that means that this should go in the alternative hypothesis. And then, we have the non-inferiority trials. Where the non-inferiority trial null hypothesis is that there is a different effect or there is a difference and we would like to reject that, okay? So, always, our final conclusion has to do with the null hypothesis. We either reject it, so you have a positive trial, we're very happy, or we failed to reject it and I would like to go back for a second to see again that we have a true situation, a true state of affairs and we have a test result and we shouldn't confuse the two because if in the test result, we end up having as a conclusion the same as truth, we'll never know, right? We'll never know that. But if it is the same as the truth, we are doing well. But what we keep in mind is that in our conclusion, we might always have an error, either a type-1 error, as it's called, or a type-2 error and what is a type-1 error? While the null hypothesis is true? We rejected it, okay? This is what we usually have as a 5% or maybe 2.5%. We decide this from the beginning and the whole community is accepting that this is an error we can take, 5%. If we're rejecting our hypothesis and in fact it was false, so you know, there was a difference. The alternative was the correct one, this is the power, what we give to ourselves to have the actual result that we would love to have. It is true that they're different and we showed that it is different. This is the power and usually it's 80%, and the other mistake we could do, as we said, is to actually not reject the null hypothesis, not be able to discriminate between the two treatments in the superiority trial while, in fact, there was a difference. This is the type-2 error, the false negative result. To see this the way we saw it earlier, I want you to look and see that the true hazard ratio equals 1 under the null hypothesis. If it's true that the hazard ratio equals 1, nothing going, then, that means, as we said earlier, that we're going to have observed value somewhere in this spectrum, around 1, okay? This is going to be the theoretical distribution of our hazard ratios from many, many, many similar trials. What we're saying is that if my observed hazard ratio will fall in this rare unexpected area of α halves, if my α, well here, my α, was .5, but if it was 5% and it was one-sided, then, if we had the hazard ratio observed that it was in this little red area, we could reject the null. It might have come from this distribution that was centred at hazard ratio equals 1 as true. It might have come from this distribution, but we allow ourselves to do this error, type-1 error. Reject the null wrongly, this little red piece here. Now, let's go to the other truth. If in fact, we are in the alternative hypothesis where there is a difference between the two treatments and what is the minimum difference that we find clinically significant? Well, here is 23% less risk of death. So, from hazard ratio 1, when there is no difference, we go to hazard ratio .77. 23% less risk, okay? And in this case, we had already decided where our cut-off to reject the null is. This is the little red piece here and below; we're going to reject. Here and above, we cannot reject. So, here and below, if in fact the truth is the alternative, that means 0.77 or something further out and we have a result in this area and observe result in this area, we're going to reject correctly. What is this? Our power. 80% of this under this distribution. The area under of this distribution where you see the green piece is our power, 80%. And what is left here, this little white thing is the β, the case that we're going to have some observed hazard ratio from the truth of .77 and it's going to fall here, it'll be still in this, from this distribution, but it will fall beyond the red, right? So, we won't be able to reject. So, this is our β. So now, if we have our α, we have our β or 1 minus β, which is the power. We have a clinically significant difference, then what is our role to design a clinical trial? Well, our role has to do with how thin or not thin are these two distributions? And these two distributions can change their shape open or close as far as their range based on the sample size. So, that's where the sample size comes and says, if you're interested in this kind of difference as clinically significant, if you want this kind of α and this kind of power, what is your sample size that will actually satisfy all of this? This is what I'll show you here. This is under the null and under the alternative. Supposedly difference of 2 means equal zero under the null. Again, the difference is 2. I just want to show you, please take a look at this. We start with 30 cases, 60 cases 120, 240, 400. You see how based on the sample size; we actually can increase our power to the point we want. So, all that the researchers can do is from the beginning, decide what their errors are going to be relative to the conclusions they're going to have at the end of the trial. If they reject the null, the mistake they could have made is α. If they do not reject the null, the mistake they might have made is β. We have this from the beginning. We set this from the beginning for the clinically significant difference, Δ. And then, we can go ahead and design our trial with the correct sample size. Again, in the superiority trial, rejecting the null hypothesis means that we reject equivalence and we accept significant difference. But failure to reject the null hypothesis should not be confused with proof of equivalence, okay? Not being able to reject the null hypothesis could be because we didn't have enough power. Could be that we were unlucky and we got into the type-2 error. It doesn't mean that we have equivalents. I wanted to ask this question, if possible. In a clinical trial to address variability of the estimate, the design needs to include, randomization, certification and blinding, adequate sample size or randomization, stratification and blinding? Could you please answer this question? Variability? Okay, should I show the results?

Prof Lesan: Yeah, it'll be interesting to see.

Prof Dafni: Adequate sample size, very good. Yes, these 80% answered correctly. Thank you for that. And then, the same question, but to address bias of the estimate. Okay, let's close again. What is the result? Okay. I have to say that this was a little tricky because in fact, as we said, we need to include randomization. Having stratification and blinding on top of randomization it really not matters as much. Okay. Very quickly talk about equivalence of non-inferiority trial. Why? Because again, I don't want us to make the mistake of looking at the superiority trial that was negative and conclude equivalence or non-inferiority. If this is the question, we want to answer whether two treatments are equivalent because something else is good, maybe less toxicity, easier administration, shorter administration, etc., then, we might want to just show equivalence on efficacy. Or not an inferiority. So, in this case, we have again to decide what is it in our mind a predefined tolerance and non-inferiority margin that if it happens, let's say, if we have a five-year survival, 97%, maybe we can live with a 93% survival if all the rest are good, toxicity, etc. etc. So, in that case, we would have, what? A 4% predefined tolerance. In this case, what we do is we calculate our outcome, it might be a hazard ratio again. We create a 95% confidence interval and we see whether this confidence interval falls between minus Δ plus Δ. In this case, we reject the null that there is a difference and we accept the alternative that says that we have a non-inferior result. So again, even non-inferiority trial, what we do is we reject the null hypothesis, which would say now that there is a difference. We'll reject inferiority and we accept the non-inferiority in this case. Again, we pause for possible questions.

Prof Lesan: Yes, one very interesting question now in the chat also from Marina Puchinskaya. How do you define what hazard ratio you want to find? Just any we can think clinically meaningful, or how do you decide?

Prof Dafni: Again, an excellent question, Marina, thank you for this. Well, in general, you have to look at the particular kind of tumour or haematological or cancer that you have and what exists in the literature at that point. I have to say that the example we saw a little earlier with hazard ratio of 0.77 comes from the HERA trial where they gave trastuzumab in adjuvant way in breast cancer, early breast cancer in cases with HER2 positivity. And in that case, a hazard ratio of .77 was what was thought by the experts in the field, a 23% reduction in disease-free survival was really what mattered. Now, in general, we have started actually as a community trying again not to have, you know, very small kind of benefit giving a statistically significant result with a huge number of patients, etc. etc. So, it's more or less a consensus that the hazard ratio of .80 or less is within the realm of clinically meaningful difference. And if you're aware of the ESMO-Magnitude of Clinical Benefit Scale this corresponds more or less to having a lower limit of the confidence interval of .65. So, this is a possible value. And in fact, what you propose here, .73, .75, are very well within what we would consider clinically meaningful.

Prof Lesan: So, basically, you can use the studies that are already done in the field and if not, maybe you can base also your assumptions on the studies from phase I and II, if that' a medication?

Prof Dafni: Yes, but in fact, for the clinically meaningful, there are other issues also, the rarity of the disease, you know, how... if there is nothing to give to patients, so even a little benefit cumulatively can give you a result. But as you said, Vadim, it's really what the field says for the specific cancer at the specific stage of the specific timing point.

Prof Lesan: Yeah, thank you. So, I think we can continue now.

Prof Dafni: All right. Oops. Okay. What I want to talk briefly about is, as you all know of course, you know, we have moved from chemotherapy to targeted therapies, immunotherapy, etc. which means that we end up with the different targets with smaller and smaller subsets of patients to use in a clinical trial. Which means that big randomised trials are not really the way to go all the time. And we do have two different ways of approaching the biomarker-based clinical trials. We either have all-comers that are stratified by the marker status or enrichment designs. And I'll go to the next slide to show you for the all-comers, what we do is we stratify by the biomarker, negative or positive, and then we randomise within the biomarker positive and within the biomarker negative levels. Now, this has a big advantage. What is the advantage that we know right away, whether the biomarker at the end of the trial, I mean, whether the biomarker is prognostic or predictive. Why? Because to be predictive, meaning that treatment A is better than B only when the biomarker is positive, for example. So, the targeted treatment is better than the standard therapy when the target is present, okay? To see whether this is not happening, when the target is not present, we need to have information for the biomarker negative as well. Predictive is a factor, a marker, a target only when its presence points to a specific treatment, okay? And if its absence, really shows that we don't have the same benefit of the new treatment. Now, if one is very sure, for example, Crizotinib for ALK positive, then one can assess the biomarker and keep in the study only biomarker positive patients, like here, the ALK positive patients and randomise only within this subgroup. Again, one has to be very sure that this is the case to go in that direction. This is the enrichment design, and I wanted you to see why it is important not to call a biomarker predictive only on single arm trials. And why randomising to the standard therapy is also very important. Here you see with the blue that the factor is present, with the red is not. And if we were just to run a single arm trial with experimental treatment, we would see, let's say if it's a good treatment, we would see that there is a difference, right? With the experimental. And we can see also that whether the factor was prognostic or predictive. Here, we can actually say looking at what happened with the standard treatment that we're talking about the prognostic factor here and a predictive factor here. Why? Because experimental treatment really had an effect only for the cases where the factor was present. So, the predictive role of the biomarker can be definitely ascertained only through randomly assigning patients to the standard treatment arm also. Finally, very briefly, the curse of multiplicity. I want us to be aware once more that the more tests we do, the bigger our type-1 error. So, for example, the probability of a false positive finding, that means rejecting the null and having a positive result, increases wrongly, increases as a number of subgroup analyses increases. And in fact, if we were to do 10 analyses, if there is a 40% chance that we'll find at least one significant result, while there is no significant result in truth, right? So, it's easy to find one subgroup in which the treatment appears to work because, you see, when we're talking about an analysis, think of five factors, mutation, age, histology, history of cancer, yes/no, previous treatment, okay? Two levels, its factors, that's 10 analyses. So, that makes it very important to actually be aware of this plot where with 10 tested, we have 40% chance of finding one positive result and we have a 10% chance to find two and so on and so forth. So, in summary, we discussed a little bit the sample size and its importance in trial design relative to the type-1 and 2 errors and the clinically significant difference. We emphasised that to address bias, the only guaranteed way is randomization. Mention the difference between superiority and non-inferiority trial conclusions. Pointing out that lack of rejection of the null hypothesis in the superiority trial does not equal to acceptance of lack of difference. Predictive versus prognostic factors. And again, what the role of randomization plays in ascertaining whether a factor is predictive or prognostic. And finally, the multiplicity problem, which to resolve it, we always need to report all analyses that have been performed so that the reader knows how many were done and maybe adjust the type-1 error for the number of tests that were performed. Thank you very much. This is everything I had prepared for you today and I would like to give the podium to Vadim, Dr Lesan, if you have any more questions or discussions.

Prof Lesan: Thank you very much Professor Dafni for this amazing session. I think you've touched very, very important points in the statistics. There are no questions, but for the participants, please feel free to put any questions you would like to get answers from us. I have one question, Professor Dafni, maybe, for you, how do you use the descriptive data to choose the test you're going to apply for the analytical part of the statistics?

Prof Dafni: Well, if we're talking within the clinical trial, we're not talking about the primary endpoint because this is decided from the beginning and all the design and the sample size is actually based on what tests we're going to be using. But if we're talking about secondary analysis or observational study analysis, we do look at the distribution of our data and then decide which is the more appropriate test. So, as you mentioned earlier, we can look at the normality and decide whether we should use a T-test for example, or a non-parametric test.

Prof Lesan: Yeah. Thank you.

Prof Dafni: You're welcome.

Prof Lesan: As you may know, so with great power comes great responsibility. So, unfortunately, in cardiology, if you look, we have like a lot of phase III studies with compound endpoints and they also enrol thousands of patients and you always find the P-value, which is significant, but with very small effect size and this is normal to see. Do you think we can still interpret these data if we decompound the endpoints? This compound?

Prof Dafni: Okay, using composite endpoints, yeah, it's very common in cardiology but just a little parenthesis, if you think about it, progression-free survival that we use is also a composite endpoint because it's both progression and death, right? So, very similarly but with a big list usually of items you have composite points in cardiology. Again, I have to say that due to the fact that you design a trial based on one major question and this is your primary endpoint. If you started having as a primary endpoint your composite, then anything else you do is of lesser value. Now, there are ways, and I mean this is maybe some insiders information that you might appreciate, there are ways of actually either splitting, as you know, the α into more than one test. So, maybe you can say composite endpoint, but also, I want cardiac death for example. And then, give 2.5% type-1 error to one and 2.5% to the other or another split, 1%, 4%. But also, there is what is called hierarchical testing, which means that if you're lucky enough and your primary endpoint is statistically significant, then you can say conditional on having my first test resulting in a significant result, rejecting the null. I can use the full type-1 error for a secondary question. So, you might start and say I have this composite endpoint, design, sample size, everything is based on that 5% α. I run the trial, I have a significant result. I have already said that if this happens, then I'm going to use my full 5% to test survival for example. So, there are ways of looking to more than one, but usually, this increases the sample size.

Prof Lesan: Yeah, of course, yeah. And then it comes also in question again, the ethical part of the studies because the last congress I visited, they were talking in one study that they enrolled many more patients in the study just to be able to do some post-hoc analysis. What do you think about that?

Prof Dafni: No, I wouldn't agree to that. If I were the statistician for this study, I would quarrel against it.

Prof Lesan: And one more question, we don't have any other questions in the chat. What would you prefer? Hazard ratio of 0.8 but with a P-value of 0.01 or hazard ratio 0.6 with a P-value of 0.06?

Prof Dafni: I don't think this is a legitimate comparison. You have to remember that what you see is one data point, right? With the variability that we discussed. So, you might get .6 while the truth is .8, you might get .8 while the truth is .6. Okay? So, and you have a very rigid kind of rule for your P-value. So, your P-value value is either significant or non-significant. Again, there is randomness in it, right? So, what matters is that we set a boundary for it.

Prof Lesan: Definitely. I agree with you definitely. And do you see any movement from the P-values to the effect size in the studies? Like getting more effect size mentions as P-values?

Prof Dafni: Yes, it is true that it is much better. And nowadays, I mean, it's more and more common to discuss based on the confidence intervals because the confidence intervals show also the variability of the estimate. While with the P-value, you're not sure. So, if you have, let's say, a lucky trial, you know with small patients, a small number of patients and getting a significant P-value, you can see that very easily if you look at the huge confidence intervals.

Prof Lesan: Yeah, I definitely agree. I pay attention all the time at the confidence intervals. So, I hope that our young participants now for this session, they also pay attention on the confidence interval and not just the P-value. Okay, I don't see any other questions in the chat, so, I think this was a great e-session and for somebody else interested in the statistics as me and I hope all our participants, they're also very interested and we see, Professor Dafni, that you really enjoy what you do and thank you very much for explaining to us all these basics concepts from statistics, but are very, very important for us, not just in doing clinical trials, but also in interpreting everything that we read. Thank you very much.

Prof Dafni: Thank you so much. And you were a great discussant and thanks for the questions. And the chance to talk to you through the European School of Oncology.

Prof Lesan: Thank you very much.