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.

Introduction to statistical analysis with R software for cancer scientists

Prof Fenga: It's an honour and pleasure to be here on the behalf of the European Institute of Oncology to share information about the software R and how this software can be helpful to cancer scientists and the cancer researchers. So, this slide here is for you to be aware that you can anytime raise a question or ask information, and please do it anytime. I'll be more than happy to answer your questions. So, a quick overview of my talk today. So, at first, I would like to introduce a bit of R and then, I'm going to try to make you aware that R is a kind of friendly, useful package. It's not an enemy. It doesn't require a lot of mathematics, a lot of statistics. It's just something that can be used in a user-friendly mode. And to do so, we need to change some paradigm here, in which way? So, basically, I just made an example from plots to LATTICE~~S~~. So, we need to change the way we think. And basically, I'm going to bring up to your attention this transitioning from plots to LATTICE~~S~~ in order to make you understand how it's possible and is pretty feasible to change perspectives in order to be more effective in cancer resources. So, then I think this is the most important part of my talk. And so, it is a bit of a practical session. So, I'm going to introduce you the Neural Network approach with cancer data and the power of simulation delivered by R. So, those guys are just two examples for you to become aware on how is effective and powerful, how is so powerful and effective and to gain some confidence by just working together on it. This is the big deal. So, instead of doing old book, old style academic lecture, I'm going to make you feel comfortable, hopefully, with this fantastic tool, which is R. So, the first provocative question: as an MD and a researcher, do I need to know R? The answer is yes. Maybe 20 years ago, 30 years ago, maybe not, but now, yes. Things move faster and faster, and we need to access our data in real-time and R is a great tool to organise your data. And you can work on your data real-time just basically running some pre-built routine. So, can I learn R? Of course, yes. But now, how? Well, R, first of all, is an open source, so, it's for free, and there is a lot of tutorials and ad-hoc resources online. So, the best way to learn R is just learning by doing, by copying and pasting the code you can find online. Basically, you just have to become a little bit familiar with this tool. And then, but just by cutting and pasting lines of code, you can easily, easily learn R and of course, improve your research. This part here, the interdisciplinarity, is of most importance. So, the winner scientist is the scientist that understands how it's important to work together. So, of course, the medical doctor is not supposed to know in details, for instance, a statistical technique or how to code efficiently. But it is not required. There are PhD students. There are master's students with the optimal command of this language that can be somehow hired on purpose, so, to carry out some statistical analysis. And once this work has been done it's going to stay forever. You have your dataset, you have your code, you just run the code anytime and you get the same result. In fact, this point here is, "should I know statistics?" Being a good practitioner is usually more than enough. Again, connecting with the mathematicians, physicists, statisticians is not just a good idea, is something that should be done. So, practically, working with R is not big deal. So, just to download the code R engine or RStudio engines, so both these engines here are just available at this website. So, nothing to write about. So, it's very easy to connect and now allow R anytime. So, installing R is really not a big deal, we just follow the directions and then,

basically, is a self-installing code, programme. So, you just follow the directions and you can have your R on your computer in a matter of minutes. Now, so, I'm not going to go through all these steps here, but at the end, you just have to click on the last version. I think we have now the fourth version, or something. And this is the file. So, you just download this file, and you just double click on this file and R would install itself easily. So, once you got your software installed, well, that's what you got here. So, this is the first, this is the first thing you're going to see. And you just can easily type your own code at this point here or better, using a simple text editor. R comes in two main ways. You know, the first one is this one here, more basic. And this a little a more complex way. So, you can easily manage four different windows, where on the first window here, just the console. So, this part here is basically this one here. And we have three additional windows where you can store your code, all the dataset available, all the variable available on your working space and then, a multifunctional window with the files open, the plots, the packages, working packages and so on and so forth. At first, I wouldn't recommend this approach here, so, RStudio, but I would just encourage you to use the standard version of R. R is very simple. So, when you are prompted to write your own code, you can write something like a 3+5, or do some math. And it's very intuitive. The way R present itself is very intuitive. So, we have the logarithm, the factorial, the Pi. So, nothing to write on about. Now, this is the example I was talking about before. So, we need to change perspective. What does it mean change perspective? And I just wanted to submit to attention the transition from plots to LATTICE. So, basically, the real challenge here is to abandon the Excel-like word to something more complex. So, Excel is considered very well an important package from many points of views. So, Excel is a great, great programme which is not designed for statistics. It is not. You can do some statistics; you can do something. It's very easy, but it's very, very limited. So, the first thing a cancer scientist might want to consider is to abandon, right away, in a flash, cold-turkey, whatever, the relaxing Excel-like environment. When I say Excel, I extend this critique to software packages like SPSS. Unfortunately, for those of you loving this software package, again, SPSS is very easy to understand, it's very easy to work on but it is a statistical package. But it is so, so rooted in an Excel-like environment, and this doesn't help. So, first of all, let's abandon this chartered territory in favour of something much more complex, but much, much more rewarding, and let's dive a little bit on this: what's LATTICE? It's a plot, but it's not a plot. So, when we ask Excel to run a plot, that's fine. You have your vectors, your data, you just, you know, use your mouse and select the variables you want to select, and then Excel would provide you with a nice plot. Well, I picked this example because cancer cells, some live in a three-dimensional space and in a sort of grid. So, in order to understand and to properly process information on cancer cells, we needed to abandon the classic plot and start thinking to our plots in terms of something that you put on grid. So, basically, we have a random process, so we can somehow conceive cancer cells, like obeying a random process such that these tumours are placed in a theoretical way, in a grid, which populates itself according to some scheme, some growing path. I hope, and I really hope I was able to make you feel the difference. It's not a plot, but it's a is grid on which I impose a random process, and this random process evolves according to some rules on this grid, on this LATTICE. So, this is just an example of the transition cancer scientists might really want to consider doing in order to improve and perform better and better. It takes time. It takes time. Maybe, you might consider taking on some course or, but like I said, there are so many resources online that you can easily build up your knowledge yourself. Any questions, or? Okay. So, with that said, I would like to work with you right now, to work with you and see how R practically helps us. Well, for instance, you have a problem, you just Google it, and then, you add the magic word R. So, the search engine, typically Google, would address you to the right topic. And by adding the word R, it would easily bring you something related with R itself. Let's make an example here. So, this is just an example of how the R-community is very well aware of the importance of R in cancer research. So, just by typing on Google: cancer plus R and you are likely to get this web address here. So, I wanted to draw your attention to how effective is to work on R. So, in this very simple webpage, you have the main purpose, you have stated the main purpose of this library. So, R is able to perform a graphical user interface for accessing and modelling the cancer genomics data. So, what does it mean? R comes with a lot of data, so, you don't have to figure it out, how can I run this programme? No, no, no, no, you got the data. Usually, those software package libraries, those libraries are somehow

approved by the scientific community. So, in this case, we have a very, very recent publication, and then, you can just go online and find it and read it. So, you can read the original paper, but not only, you can have some additional, some big net, so, some additional information in a very user-friendly manner. So, please try it, try it and you will see how easy it is to work on R. It's just a matter... for instance, to install this package, start R, and it's just the version, so basically, you need the version of 4.2 or superior and enter this. You don't have to type it, just cut and paste on your text editor and R will do the job for you. The documentation is always, in R case, the community is very strict so, you are likely to end up with a very nice and tested codes. So, usually, there are no surprises, and usually, things work as they should. So, this is just an example. Now, I wanted to work with you more closely on real data. So, this is a little bit of practical session. We had two sessions. The first one is just how to practically use R for cancer research. And then, the second one is just a little presentation on how to simulate tumour growth. So, I picked those two topics because I do believe it's very important for you to become a little bit more familiar or familiar with this tool here and in general with R, and then, how to simulate some tumours. So, a little bit of, okay, let's skip it and let's share, okay, you should be able to see by now the R. So basically, this is how R works. So, here have a simple text editor and here we got the console. So first of all, we need... So, R works according to libraries, so, a set of instructions needed by R in order to perform the required elaboration. So, by doing that, install the package, and then, the name of the package, R will automatically install the required library, like a Neuralnet, Deepnet and MLBench. So, and in order to run these, I just type Control R on my keyboard. So, I load my libraries, and here are the data. So, the library MLBench comes with breast cancer data. So, let's visualise those data. So, we have for each patient, we have a... I apologise, I don't know anything about this. Yeah, the cell shape, I can imagine, but the nuclei here or the chromatin, I don't know anything about that, but I suppose that of course they are cancer-related. So, I have a set of patients, for each patient I got a series of variable, cell data, so mitosis, so. And then, I have the outcome: benign, benign, malignant, benign, benign, so. This is my data set. So, well, this is a bit of a cleaning, so, I just want to get all the complete cases. I don't want blanks. So, that's what I'm going to do. I can explore a little bit of my dataset by doing head plus the dataset, names plus the dataset. The last column is related with, like I said, with outcome. And then, by doing this, I just record my dataset according to something that R likes. So, it's no big deal. You just use your data. The y variable is just the 11th vector. So, it's just benign or malignant. So, we change the benign, the malignant using 01. So, we just do something that R likes. And then, we just have to run our neuro-network. We use five neurons and basically, that's it. Now, we want to know how effective has been our analysis by running something called the confusion matrix. So, basically, the code wasn't able to classify the code zero, which was benign in five-case and failed to classify correctly 23 malignant cases. And then, can have an assessment of the precision, so 95%. So, what's going on here is very simple. You might be a little bit confused by this, but I promise you, is a matter of 15 minutes to understand this line of code here. I just selected which YY, which is the prediction coming from my neural-network, and I just force the code to put number 1 to those values which overcome the mean value. So, it's really, it's really simple, it's really simple. It might be confused, but it is really nothing to worry about. And now, I wanted to introduce you to the neural package, which does the same exact job, but is a different code. So, again, instead of using this syntax here, this one here, it's maybe a little bit more intuitive. So, I run this common here, Neuralnet. I add all my variables here, and I use five neurons. Again, it takes few seconds. Again, you got the nice results. So, it doesn't fail any zero. And it fails just 11 malignant-case. This is the plot of the neural-network, which is not much informative for us. And now, the last example is related to the simulation. And again, I installed the package SITH which I already did. I run the library. I set a seed, is just a number which I can invent, I can put 12. So, I just needed to, if I want to replicate my exercise, I need to generate an artificial seed. Then, by using this comment here, simulate tumour, I can simulate a tumour. Now, I wanted to make you see how easy is to work on R. So, suppose that I don't much about this comment here, and I want to go deep. I want to understand a little bit more. Or maybe I forgot it. Maybe, I read the publication, the article, but at this moment, I don't know how to work with this. Well, double click, copy and paste. Where? Help. So, I go to the R functions, and again, I paste simulate tumour here. I press the okay button, and that's it. R goes online. And I have everything I need to know about simulate tumour. So,

max_pop is number of cellular tumours. Div rate is cell-division rate. Mut rate is mutation rate. So, I can customise my simulation according to those simple examples, and then, I can run my examples as well, because each time you ask R for help, R gives you very nice explanation. In this case, we have a Poisson distribution, again, we have LATTICE here, three-dimensional LATTICE. So can you see how things transition from simple to complex, but please know, in a friendly way, and you can change colours, the simulated time in days, and then, you got your nice example. Now, back to our code. Same thing here. If you want to know visualised tumour, well, I go here, help and function, I change, and then visualise tumour. And that's what I got. So, really, really easy, easy job. So, let's simulate our tumour here. It takes a while. Okay, and that's my tumour here, is simulated. By doing names out, out the simulation variable, you know, I can extract the colour scheme, the drivers, the time, so everything I need. And you can see that is 3D plot. So, for each genotype, I have the 3D coordinates. I'm running out of time here. And again, this is how I can simulate the tumour itself. And so, what I wanted to do is just to show you how easy it is to work on R. And before heading over to Dr Bertolaccini, our discussant today, I wanted to let you know that my e-mail is there, is l.fenga@exeter.ac.uk. And you might consider sending me some emails if you need any help. I would be glad, I'm glad to help. And I can give you some suggestions, so run some code with you. And so, please, let me know if you need any help, I will be glad to help. Thank you so much for your attention. Let's stop sharing here.

Dr Bertolaccini: Thank you, Livio, for your beautiful presentation about the best statistical software, the best and free statistical software. We have to remind every time that R, it's completely free. Okay, we have some points of discussion from the floor. The first is when an MD should learn R, during the medical student or after, during the residency?

Prof Fenga: Yes. Thank you, Luca. This is a very good question. Basically, R is ageless. So, when I encourage old guys like me to improve R in their 50s. And I just say, R is like play bowling, so you can become a world championship at 70-years old virtually. So, ideally, ideally during the, yes, at the undergrad level, each medical student should take at least one class, at least one class. And sadly, statistical classes are given sometimes without any computer assistance. So, just a bunch of notions which might not be super useful. I hope I answered the question.

Dr Bertolaccini: The second question is about the difference between R, the normal R and RStudio.

Prof Fenga: Yes.

Dr Bertolaccini: What do you suggest for me, for the audience, for start to move the first steps on R?

Prof Fenga: This is a very good question. Many people think that RStudio is the way to go. I personally suggest to use the simple text editor; is less confusing, and just copy and paste the instructions found on the net or some paper, whatever. Just practising on the simple text editor. And this kind of wording is less confusing because working with the four different windows might be a little bit too impactful. So, unless we are dealing with some 19-years-old guy with some already practise, or some ability on computers, I would really, really encourage you to use R in its simple version, just R plus a text editor. You cannot go wrong with that.

Dr Bertolaccini: A question from the floor is, "it's easy to do a multi-variable analysis?"

Prof Fenga: Yeah. No, it is not.

Dr Bertolaccini: On R.

Prof Fenga: Yeah, yes, yeah, yeah, thank you Luca, no, it is not. I don't know, Luca, if you can let me to just run a simple example on just by sharing the screen. So, suppose that I want to do some multivariate analysis. Okay. You should be able to see my screen now. So, I just go to R function help. And I know that I need to know that LM, is linear model, is the way R understands regression, and that's it. So, LM is to fit linear models, including multivariate ones. So, by using LM, I can do that. Well, this is the usage. I can just go to the final

thing and just run an example. So, what are you going to do? Are you going to do this cut and paste? Let's open a new text editor and that's it. So, this is my, whatever it is, is something, I don't know, some variables. This is my groups, this is the weight and then, this is the univariate analysis, univariate analysis without intercept. So, I can just do summary, lm.D9. That's what I got. And I just have to add a new variable here. So, let's call that AAA and C, and you write 4, 5. So, how many, let's see, I have 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, ... 4, 5. So, I just invented a variable, 2, 3, 4, 5, 6, 7, 8, 9, 10. I just add it here. So, that should be... I did some error here. Oh, yeah, because it's 2, 1, 2, 3, 4, 5, 6, 7, 8, and 10. 3, 4, 5, 6, 7. Oh, yes. Okay, yeah, I don't have time to correct. I just, yeah. But, you know, it's just a matter to add another variable and I can do my multivariate regression. So, it's very easy. It's just a matter to copy and paste. That's it. Let me...

Dr Bertolaccini: Thank you, Livio.

Prof Fenga: Sure.

Dr Bertolaccini: The last question is not a question; I want to know your personal point of view. I know very well your personal point of view, but I want you to share with us about the multidisciplinary approach to the research. We have to tear down the wall between statistician and clinical researcher.

Prof Fenga: Absolutely, Luca, yes.

Dr Bertolaccini: It's the time to tear down the wall?

Prof Fenga: Yes, yes. Well said, and well done, Luca. Yes, we need to tear down and start selecting the people we like to work with according to the attitude to work together. Scientists should mix-up. And this is the only way to overcome, unless we have five, six life spans, which we can't. We have a limited time from A to Zeta, and we can't conceive, you can't imagine to learn what we need to learn in order to provide good performance. So, we need to gather. So, we need to multivariate ourselves. We need to create a unique engine. And in the UK, that's what we are trying to do. Sometimes it's not easy when old school professors come into play. But in my humble and personal opinion, there is no way out, multidisciplinary and work with those professors, those researchers willing to open. This is my personal point of view. And thank you for asking, Luca.

Dr Bertolaccini: Thank you, Livio, for your beautiful presentation, with a lot of points to learn about R. I personally use R, and I could say that it's not easy to use, but it's easy to learn. So, don't be afraid of R and start. Try, start. Let's try. No worries about it. Take your time and try to use R. Thank you so much, Livio.

Prof Fenga: Thank you very much.

Dr Bertolaccini: And thank you to all attendants. And have a nice evening, night, and day.

Prof Fenga: Likewise. Thank you very much.