TRANSCRIPT: Harmonisation on the RECAP Preterm Data Platform, ECR module

Hello, my name is Gonçalo and in this video I will be talking to you about harmonisation on the RECAP Preterm data platform. I will give you an overview of how we used the platform to harmonise data for the EPT demonstration project, which as I'm sure you've heard on previous videos of this module, is a demonstration project inside the RECAP Preterm project that looks at sex differences in perinatal survival of extremely preterm infants.

But before we actually start talking about harmonisation, let's just have a look at what the RECAP platform actually is, if you haven't watched the previous videos about the RECAP data platform, which are WP4 videos of the RECAP Summer School, those have a lot more specific information and detailed information about the platform. So this is just a quick, brief overview. So let's start.

So this is a map with all the nodes that compose the RECAP Preterm data platform. So there are 13 nodes hosted across Europe and the platform is a federated infrastructure, meaning that all the nodes, those 13 nodes that you see here on the map, they all have data. So the data is distributed, as you can see here on this second point, nodes host data from their respective cohorts. So some RECAP partners that have conducted cohorts in the past or still are conducting cohorts, they have data from those cohorts, and they host that data on their own nodes. So each partner hosts their own node where they put their own data.

And also, among these 13 notes, there is a special kind of node, which is hosted here in Portugal, which is the central node. And the central node is structurally the same as the other nodes but it's used for a different purpose. It's basically used to collect data dictionaries and summary statistics from all of the other nodes. So the other nodes share that information, the data dictionaries and summary statistics, with the central node. And then the central node takes all of that information collected from the other nodes and assembles that information into what we call the RECAP Preterm catalogue. And as you saw in the previous video, that's where you can go to browse through the studies that exist on the RECAP Preterm platform, search for variables, also see summary statistics of some variables. So that's how we are able to build the RECAP Preterm catalogue.

Now there are challenges with harmonising data that is distributed across multiple, physically distant places. So the way harmonisation works on the platform, is each partner is responsible for harmonising their own data. So harmonisation on the platform works in a distributed fashion. Let's see how it works here. So this is harmonisation across multiple RECAP nodes, I'm just using two nodes, just to simplify. So here we have two nodes. And let's say each of those nodes have a dataset, a table there. That table would be data that was collected from that node’s cohort, let's say. So each of those nodes, each of these nodes have the original collected data. And as you saw in the previous video, Helen Collins talked about how you would go about selecting and choosing your variables, the variables that you're going to need for your analysis in order to answer your research question, and so in this slide, we're sort of picking up where the last video left off, which was the point at which you have already selected your variables and you have created a harmonisation dictionary where you describe your variables, the variables that you're interested in analysing. So, that would be the starting point of the harmonisation process on the platform: is coming up with the dictionary, the harmonisation dictionary, then what you would do is send that harmonisation dictionary to the cohorts. So in this case, to each of the nodes that would will be participating in your study.

And what they will do is use that dictionary to create what we call a view over their original tables. A view is still a table, but it's a special kind of table. It's a table that has a reference to some other table. So in this case, this view has a reference to that original table. And what a view you can do is pull data from that table, transform it in some way, according to whatever is defined on the harmonisation dictionary.

Now the intermediate step that I mentioned (transforming the original data into something that complies with that harmonisation dictionary), that's done using harmonisation scripts as you can see here. So in order to create a harmonised view of the original data, you would have to feed it the harmonisation dictionary and also the harmonisation scripts. Now, the thing to note here is that the harmonisation dictionary… there's a single harmonisation dictionary that is the same for all the participating nodes. They will all take that harmonisation dictionary and harmonise their own variables to the variables that are defined on that dictionary. And so that piece, the harmonisation dictionary, is the same across all nodes. Now this piece right here, the harmonisation scripts, those of will be different because it will depend on how the original variables are defined on this original table. So those will be whatever they need to be to transform that data… to pull that data and transform it in a way that it would comply with what was defined on the harmonisation dictionary.

And so this is the process that all nodes must undertake. And then once they do, that will produce a harmonised view. So this will be… it's technically called a view, but it's still a table with data, still a dataset but the data that is there… so the variables that will be here on this table, will be the ones that are defined on the harmonisation dictionary. And the data will be the data that was here, but was transformed using harmonisation scripts, and now it's in a format that complies with that harmonisation dictionary.

So let's now have a quick look at how this part actually is done on the platform. So the harmonisation itself. And there's two main ways of harmonising variables on the platform. And it depends on the type of variable that you are harmonising to. So it depends on the target variable. And by “target variable”, I mean a variable that is defined here, on the harmonisation dictionary. And let's say you have a target variable that is a categorical variable. And you have one of your variables on your original dataset that can be harmonised to the target variable. And if your original variable is also categorical… so you're mapping one categorical variable to another categorical variable here. That's the easiest way of harmonising variables on a platform, because if you're harmonising from a categorical variable to another categorical variable, you can use the graphical web interface of the platform to do that. And it's very easy. As you can see here from this screenshot. This is a screenshot from a harmonisation of a “sex” variable. So you can see here the original categories of the original variable, which were 1, 2, 3, and 9. And the labels were “male”, “female “,”undetermined” and “missing”. And you can then just choose to which categories on the target variable, you are going to map those original categories. So you can see here, for example. This category here, coded with 1, which corresponded to “male” on the original variable, is now going to be mapped to 0 on the target variable. And 2 is going to be mapped to 1, 3 to 9 and so on. So this is how you can map categories. So it's very easy. And this was the case, actually, for most of the variables in the EPT demonstration project. Most of them were pretty straightforward because you could use this graphical web interface.

Now the second way, is if the target variable is not categorical, but is instead a continuous variable. Then you cannot use the graphical interface. You would have to manually code a script for that variable. And the scripts, the harmonisation scripts, on the platform are done using a language called MagmaJS. And I have two examples here to show you. I'm not going to go into much detail here. If you want to know in more detail how this works, you can watch the fourth video from the WP4 module of the RECAP Summer School, where I talk in much more detail about harmonisation and actually a demonstration of harmonising variables using both the graphical interface way and this manual script way. So these are two examples. This first example is harmonisation for “gestational age in days”. So here we're taking a variable from the original table, multiplying it by 7 and then adding the values from some other variable. Then on the second example, which is a harmonisation script for “intubation at delivery”, it’s slightly more complex, but it's still not very hard to do. Once you get used to the syntax of the language, it's pretty much like any other language that you've used.

So these are the two main ways of harmonising variables on the platform. So remember, these two ways that we saw of harmonising variables, are this step right here. The harmonisation scripts that go into the view, that are used to pull data to the view… remember, the view pulls data from the original table, uses those harmonisation scripts to transform that data and produce the final harmonised view. So this is what… in the EPT project, we asked all the participating partners to do this on their own nodes. So we gave them the harmonisation dictionary. They took that harmonisation dictionary, put it on their nodes and created a view over their original data and harmonised their variables to comply with the harmonisation dictionary that we had given them. And once all the partners did that and produced their respective harmonised views, we were able to do something like this.

So as you can see here, all the participating nodes, after having done the harmonisation and produced their harmonised views, would send that the information about the harmonisation that they had done to the central node. And the central node, receiving that information, is then able to produce an overview of the status of harmonisation across all the participating cohorts. So as you can see here, the first column are the 22 variables on the initial harmonisation dictionary of the EPT demo project. Then we have one column for each of the cohorts that have done the harmonisation so far. And then for each cohort, you can see the status of harmonisation of each variable. So if they were able to harmonise the particular variable, you will see a checkmark and if they were not, for some reason, you will see across, a red cross. You could click on any of these checkmarks or crosses and you will see the… let's say you click on this one, for example. What you would see, is the way that this particular cohort harmonised this variable, so you will see the harmonisation script for example, you will also see summary statistics of this variable in this cohort.

So you can see individual summary statistics and individual scripts for each cohort and for each variable. And that's if you click any of these checkmarks or crosses. But if you click on any of these variables… let's say we click on this one, for example, “alive at discharge”, you will see something like this. And these are summary statistics. These are just frequencies. But these are global frequencies. So, frequencies across all the cohorts. You can see here, there's a drop down box. So right now it's “All”, so these are global statistics, but you can then click on it and select one of those cohorts right there.

So having this whole harmonisation process documented here on the platform, I think it's pretty obvious how useful it is to be able to have something like this, where you can see the status of harmonisation across all the participating cords. Because if you can, you can sort of check in with the cohorts. Because if you see… let's say you see a cross for a particular variable in one of the cohorts. You could then contact them and say: “So I saw you weren't able to harmonise that variable. Do you need any help? Do you need more information?”. So that was really helpful to have the process documented here. You can see the scripts for each variable of each cohort.

So the process of harmonisation on the platform, you could say it's self-documenting, because as you do it, the information will show up here on the central node and you can just keep track of the progress of the harmonisation across all the nodes. So that's very useful.

And so, these are the some references. So the first one is the link to the central RECAP catalogue. And the second one is a wiki that we've developed for the partners, the participating partners on the EPT project, just to help them maneuver their way on the platform, in terms of harmonising their data. And then this is the link to the OBiBa website. OBiBa be developed the software that we are using for the RECAP platform. And then also a link to the DataSHIELD website, which is the analysis software that you are going to see on the next video.

So if you have any questions, you can use the RECAP forum to post your questions there or you can just email me directly on this email address. Thank you for watching.