TRANSCRIPT: Interview about demo project, ECR module

JULIA DOETSCH

Hello everyone, my name is Julia Doetsch and I'm part of the ECR group. And today we're going to have a discussion on the demonstration project with members from the ECR group. So in RECAP Preterm, we have the possibility to do collaborative research with multiple preterm cohorts and register data. And we wanted to demonstrate this in a project on sex differences in neonatal mortality.

So hi, Andrei, could you tell us how you came up with the initial idea for the demonstration project? And what is it about?

ANDREI MORGAN

Hi, Julia. Yes. So, our first demonstration project using the platform was to look at the differences between the sexes and the outcome at discharge from hospital and whether extremely preterm infants survived to discharge or not. And we therefore were also able to look at some interim outcomes, such as just whether they were born alive or not born alive. And we decided to look at this as our first question, because differences between the sexes is not only an important and an interesting question, from a medical and the clinical point of view, and epidemiological point of view. But also, it's using some very simple variable and it's essentially binary variables like sex or survival, as opposed to some of the more complicated or variables that require more complicated harmonisation. And for a start, we wanted to have a project that would enable us to involve as many of the cohorts and as many of the investigators that are part of the RECAP Preterm consortium and have data on the RECAP Preterm platform as possible.

And so that was really our objective. So it's a demonstration project, not only for the scientific question, but also a demonstration project that we can do this kind of research using the platform that we've been constructing.

JULIA DOETSCH

Okay, and what was your initial hypothesis when you developed this demonstration project?

ANDREI MORGAN

We had a number of hypotheses, as I mentioned already, this is an interesting and important question from a kind of clinical everyday point of view — and also from an epidemiological point of view. The reason it's important clinically, is that as clinicians, we often tell parents of newborns or of mothers to be, for example, who have threatened preterm delivery, that survival might be better amongst female infants. But the evidence of that, particularly in an extreme preterm population, has been extremely weak and overall there is evidence to show that survival was better amongst female foetuses and female infants, but this is predominantly based on data collected from children at an older gestational age and not the extreme preterm population. And so there are very few data amongst the extreme preterm population. And we know that there are other differences between the sexes as well, in terms of morbidities, and so on. And in fact, there are sex differences throughout the lifespan.

We had three hypotheses that were related to this particular project. The first was that the sex ratio at birth of extreme pre terms is biased towards males — i.e. that more live born babies are male — and the differences are seen according to the reason for delivery. But, we then thought that the survival to hospital discharge of the live one babies is higher in females and also that live born extremely preterm male babies would have higher rates of neonatal morbidity than the females. And these are all thoughts that are consistent with what people observe in practice. So it's known that there is a higher ratio, higher number of male babies are born than female babies, but the females do seem to have better survival and better outcomes than males subsequent to that. So that was what we wanted to explore.

JULIA DOETSCH

How did you go from your research idea to start the collaborative research demonstration project?

ANDREI MORGAN

The most important thing, first of all, was to try and develop a protocol and to come up with the questions and to think a little bit about the background evidence that already exists. And in doing this, we first started to sketch out the idea and we then shared it amongst the broad range of collaborators that are involved. So we sent the protocol around to all the different collaborators, so that they could have the opportunity to inputs and to become part of the project and the fact is, the important fact is that this is a collaborative project involving lots of people, and we want all of those people to feel ownership and to feel like they're part of the project and not to just be a kind of small group of people who are dictating what everybody else has to do, as in the rest of RECAP Preterm. This is something that can only be done if lots of people participate. And so therefore, it's important to have everybody's input from a very early stage and subsequent to that, we've arranged lots and lots of meetings, we're meeting regularly, so that people can continue to have input and to participate in developing the project as it goes along.

JULIA DOETSCH

Thank you so much, Andrei. Helen, hi. Question to you.
So how did you select and harmonise variables for this demonstration project?

HELEN COLLINS

So, the initial set of harmonised variables were selected based on the research question we wanted to answer. And they were, as Andrei mentioned, selected with input from the other members of the project that were involved. And in this case, they were variables relating to the research question, which was sex differences in survival in these extremely preterm infants. And these included the exposure which is sex, and the outcome, which was survival to discharge and a number of covariates, and confounders which were likely to be important based on our hypothesis. And once we had this initial set of candidate variables, it was important to assess the availability of the variables and the quality of the available cohort data.

For this project, as well, we had to look at the study populations, because actually, the range of gestational ages available for each cohort and the point of recruitment made a difference as to which of the studies, the various cohorts, would be eligible for. And they also made a difference to… some of the recruitment time points actually became variables, harmonised variables, for the different studies. So we weren't just looking at existing variables, we were looking at the recruitment time points and creating variables based on those two.

JULIA DOETSCH

So what should you keep in mind when selecting and harmonising variables for a collaborative project

HELEN COLLINS

The most important thing is probably the balance between the quality of the variables and the number of cohorts that you're able to include. And you may be able to, in some cases, alter your variable definition slightly to enable you to include more studies, and so the variable definition you choose is really important. And you have to also assess the categorical variable categories, because while it may appear that you can harmonise these variables, it may be that the categories don't allow you to and you need to make sure that you have sufficient detail to be able to answer your research question. So in some cases, if you change the definition too much, it may mean that you can't actually do the analysis properly and get meaningful results from your analysis. So that balance between the amount of detail in a variable and the note number of cohorts is really important.

And you may also have to look at the amount of missing data which is available within the harmonised variables you create. Because in some cases, while a variable could be harmonised, there'd be a lot of missing data and, therefore, that's a really important thing to assess too, as that could really impact in your analysis.

After the harmonisation, it's also really important to use descriptive statistics to be able to look at the harmonised data that you've produced from all of the different cohorts to assess whether it's actually comparable. And these differences could actually suggest real differences between the cohorts but, in some cases, they can also suggest that there's a problem with the harmonisation and that you need to look at that harmonised data more carefully before you include it in your analysis.

JULIA DOETSCH

Thank you so much, Helen. Gonçalo, question to you.
When you've selected your variables, how do you then operationalize them on the platform?

GONÇALO GONÇALVES

Well, after you've selected your variables, I think you could… well first thing you can do is just go on the central RECAP node and search for those variables. And you'll get a list of variables and the cohorts that have those variables. So you could then contact those cohorts asking them to participate. And those interested in participating, you would send them a harmonisation dictionary that contains those variables that you've selected and then the cohorts would take that dictionary and use it to harmonise their own data on their own notes. So each partner would harmonise their own data on their own. And then finally, each cohort that has harmonised data would grant you access to those data so you could perform your analysis and work on your research question.

JULIA DOETSCH

So what are the advantages of working with a platform?

GONÇALO GONÇALVES

So I think the best way to highlight the advantages of working with the RECAP platform, is by looking at the concept of FAIR data. And FAIR, if you don't know, FAIR is a set of guiding principles for scientific data management, and which became sort of the gold standard of over the last five or six years for data management. And FAIR is an acronym. It stands for findability, accessibility, interoperability and reusability. And I would argue that the RECAP platform does comply with those principles in a lot of ways.

So the first one is findability. And the data on the platform are finable. Most of the work on making the data findable was done by our colleagues at Work Package 3, gathering all the study-level metadata, and cataloguing and classifying all the variables from all the cohorts in RECAP. So all of that information helped to build the central RECAP catalogue. So that's findability: the first principle.

And then the second one is accessibility. And that means that users, after having found the data (by the first principle), must then be able to actually access those data. And that would be your data access management, and permissions management. And those are all things that you can certainly do on the platform: control who has access to what and what level of access they have. So that's accessibility.

And then the third one is interoperability. And that's the ability of the platform to integrate with other systems, other services and applications. And the platform does integrate very easily with other applications, such as DataSHIELD, for example, which works so seamlessly, that it almost looks as if it is part of the platform, but it's actually a separate thing. It just works with the platform. So that's what it means to be interoperable. Also, another example we could give is the RECAP MyLife mobile app, led by WP6, that's also an application that can export data from people's phones into the platform. We've made it work that way, so that goes for interoperability.

And then the last principle is reusability, which is sort of a natural consequence of the other three principles. It just consolidates the other three principles. I mean, if the data is findable and accessible and interoperable, that, of course, makes it easier to reuse those data.

So I think the compliance with all of those principles is certainly a major advantage of working with the RECAP platform.

JULIA DOETSCH

Thanks, Gonçalo.
And what's the added value of investigating your research question in several cohorts? Andrei?

ANDREI MORGAN

So there's quite a bit of added value, actually, the first important thing is that it gives us the power to look at this question in a larger number of babies, of subjects. And so therefore, there's increased statistical power. The second thing is that we're able to look at potentially how things change over time because the cohorts took part in… or happened at different periods over the past 30 or 40 years or so. So we can look at some impact from the temporal changes and then the other advantage that is kind of linked to that, is the geographical differences that have occurred as well, or that may potentially have occurred, and that may relate to, for example, differences in decision making across different countries. So those are the big advantages, I think from using it in the cohorts.

JULIA DOETSCH

Thanks, Andrei. Josephine, hi. So, question to you.
How are Nordic national registers different from cohorts? And what are the advantages and disadvantages of using these registers?

JOSEPHINE BILSTEEN

So I guess the main difference between cohorts and registers is that registers, they are collected for administrative purposes. So it's a government deciding, or the ministry deciding, that we should collect information about these individuals. And this means that we, in the Nordic national registers, have information about everyone born in the country over time. So one of the advantages is, as Andrei just pointed out, that you can investigate time trends because the Nordic registers were established as soon as… the Norwegian one was established back in the 60s. And other advantages is that you have larger numbers since you include everyone born in the country.

The disadvantages would be that you do not decide as a researcher what is collected. So this is a question of what is of interest for the national statistics and for the ministry. So you cannot ask more specific questions as you can when you set up your own cohort.

JULIA DOETSCH

So what are the strengths and limitations of combining cohort and register data for this demonstration project?

JOSEPHINE BILSTEEN

I think the registers can add something with looking at these temporal transfers without looking at one cohort and another one, then you follow a whole population, so this makes a nice interplay. So you can nest your cohort inside the registers, as we also do have some Nordic cohorts and I think we can mainly add to the temporal trends by the Nordic registers and then the power, if there's something specific you want to investigate, such as interactions.

JULIA DOETSCH

Thanks, Josephine. So now we're moving on to a session where we have individual presentations from each of the members present. So Helen will talk about harmonisation, Gonçalo, will talk about how to work with the platform, Andrei will talk about how to carry out the analysis and Josephine will give us some insights about the registers.