Transcript – WP3 Metadata Catalogue Helen Collins

Slide 1

Hello, I am Dr Helen Collins and I am going to talk to you about the RECAP Preterm Metadata Catalogue on behalf of the RECAP Work Package 3 team.

Slide 2

In this video you will learn about:

What study level metadata is

Why the metadata catalogue is important

How we developed the RECAP Preterm metadata catalogue

And where is cohort metadata is found on the RECAP Preterm data platform

Slide 3

For the RECAP Preterm project we collected a range types of study documentation that describe the cohorts and their methods. This includes protocols, questionnaires used in the study, assessment records, ethics documents, as well as publications from the cohorts and other relevant documents. We also developed a metadata request form to collect information about each cohort, such as the study investigators, the population, number of participants and when participants were followed up. You can see an example of the form on the right.

Slide 4

In order to collect the cohort metadata and documentation, we developed a metadata request form which was sent to all cohort Principal investigators. We also asked them to send us all protocols, questionnaires and other useful documents. All this documentation has to be curated and stored so it can be used by other researchers across the RECAP preterm project – we have hundreds of documents supplied by the cohorts that we have to manage.

We also used these documents to create a metadata catalogue that describes each of the cohorts in a standardised way so that researchers can easily find out more about the cohorts and make decisions about whether they should include the cohort in their analysis. We started by deciding on the format for the metadata catalogue and standardised terminology, created a template and extracted the metadata for each of the cohorts and this was used to create the metadata catalogue on the RECAP Preterm data platform.

As the project has continued and we have been using the metadata catalogue we have continued to make changes to the design and the type of information we include. As the studies and their investigators change over time, we also regularly need to update the information in the metadata catalogue – this process is always ongoing.

Slide 5

The metadata catalogue describes the cohorts and their data.

The catalogue is an open resource that any researcher can use to find out more about the cohorts, so it highlights opportunities for external researchers to carry out research re-using existing cohort data and facilitates collaboration between researchers.

It provides important context about the cohort studies and the data that were collected. Before starting an analysis, it is important that you understand how a study was carried out to check if it is appropriate to use for your research question – this is particularly important where you will be combining data from several sources such as in a pooled analysis across multiple cohorts.

When deciding whether to include a cohort in an analysis, you should consider a number of factors:

- the aims of the study – why was the data collected, what research questions were the cohort aiming to answer.
- The characteristics of the study – such as dates of recruitment and follow up and the size of the cohort.
- Population – who was included, who was excluded, where did they select their participants from and does this result in any biases. Is there a control group - how and when were they recruited.
- Methodology – how was the study performed, what age were the participants assessed at, what methods did they use and are these comparable to the other studies you plan to include in your analysis, what questionnaires or tools did they use.

If there is marked heterogeneity between the studies you are using for a pooled analysis, it can bias your results. So it is important to assess the potential for bias in your sample and try to harmonise the study populations and methods used for data collection, to reduce the risk of bias due to methodological differences.

Slide 6

The Metadata Catalogue for The RECAP Preterm is made up of several sections:

- Objectives
- Study personnel
- Study overview
- Populations
- and Data collections

Next we will see how these metadata are displayed in the metadata catalogue.

Slide 7

The Studies page on the RECAP platform forms the metadata catalogue. As you can see each of the cohorts is listed with some basic summary information such as the number of participants and the number of variables available. By clicking on the read more button you can access the metadata catalogue page for that cohort.

Slide 8

The metadata catalogue entries for each study follow a standardised format.

At the top of the metadata catalogue page there are links to the cohort datasets and variables. Below that the aims and objectives of the study and the cohort contacts are listed. Information is provided about the number of data collection events, dates of follow up and the number of very preterm or very low birth weight participants. Significant publications such as the first publication describing the cohort, or the protocol publication are also listed in the marker paper section.

A timeline for the cohort’s data collections is automatically generated based on the dates of the data collections listed. Below this is information about the population including the selection criteria for the very preterm or very low birthweight group and the control group

The data collection events are listed at the bottom of the page. By clicking on the link for the data collection event you can see a description of the assessments carried out and the number of participants assessed.

The information in the metadata catalogue is searchable using the platform’s search function allowing researchers to identify cohorts for inclusion in their research project.

Slide 9

Our experiences of collecting cohort documentation are that it takes longer than expected. It may be necessary to go back to cohorts to collect missing documentation, or you may need to collect additional information later that you hadn’t realised you would need. In some of the adult cohorts where patients were recruited almost 40 years ago, there can be missing documentation as the original copies of protocols or questionnaires have been lost over the years. Documents may also only be available in the local language. To enable these documents to be used for further research they must be professionally translated to ensure that the meaning is not changed.

Many of the questionnaires and tools used by cohorts are copyrighted meaning that they cannot be shared. Researchers within RECAP have to check whether materials are copyrighted before sharing with other researchers or making the information publicly available.

The process is ongoing. Cohorts have new follow ups or the investigators change meaning that updates to the documents and metadata catalogue are needed. We also make improvements to the design and structure of the catalogue meaning that the existing content needs to be changed.

The metadata catalogue has been a vital tool in development of our initial demonstration projects. For example, understanding the sample sizes and the specific populations within each cohort has been important in the demonstration project assessing sex differences in perinatal survival. All the information is available on the data platform, meaning that researchers don’t have to contact the investigators for each cohort individually for every analysis project.

Slide 10

References and further information are listed here

Slide 11

Thank you for listening.