Transcript - WP3 – Anatomy of a data dictionary & the development of a data schema - Deborah Bamber

Slide 1

Hello. I am Deborah Bamber and I am going to introduce you to the RECAP Preterm data dictionaries and the data schema on behalf of the Work Package 3 Team.

Slide 2 - Learning Objectives

By the end of this session, you will:

- Understand what the RECAP Preterm data dictionary is and why it was developed
- Understand why and how the RECAP Preterm data schema was developed
- Understand how data mapping is carried out
- Appreciate the complexities of curating data from multiple cohorts and variable mapping as I share our experiences
- I will provide links to the data dictionary templates, the data schema and mapping instructions so you are prepared to create your own data dictionary if you have very preterm birth cohort data that you would like to include on the RECAP Preterm platform.

Slide 3 - The RECAP Preterm data dictionary

So what is the RECAP Preterm data dictionary?

The RECAP Preterm data dictionary is an Excel file that we created to capture variable-level metadata from all the cohort studies involved in the RECAP Preterm project. Cohort investigators completed the data dictionary with information relating to each individual variable collected as part of their study. They completed a separate data dictionary for each stage of data collection, for example at the pregnancy, antenatal and birth stage and for each stage of follow-up such as at 2 years, 6 years, 12 years and so on.

The data dictionary was designed to be compatible with the RECAP Preterm platform so it collects all the necessary variable-level information that is needed to populate the platform.

As shown here, the data dictionary includes the variable name, label, a detailed description of the variable, the unit of measurement, for example in years or kilograms, and value type which would be integer, decimal, or date. This information is all recorded on a sheet called 'Variable' and afterwards we added columns for tracking data dictionaries and variables for harmonisation, so columns for data dictionary version, an index number for each variable and a table name which is a short name for the data dictionary.

There is a separate 'Category' sheet where cohort investigators recorded category information for each variable where applicable, for example male/female, and information relating to missing values, for example 1 for missing.

So, the data dictionary allows all the variable-level information to be uploaded to the platform in a standardised, machine readable, way.

Slide 4 - The RECAP Preterm data schema

What is the RECAP Preterm data schema?

The platform allows the integration of a classification system, or data schema, to organise individual variables, to allow topics of interest to be searched and to facilitate harmonisation and analyses. Existing data schema's did not cover the range and detail of the topic areas relevant to very preterm birth cohort studies or were too complex in structure to organise cohort data and implement as a search function on the platform, so we created our own.

The RECAP Preterm data schema is now a fundamental component of the platform not only structuring and standardising data in preparation for harmonisation but it can also be used to inform future data collections.

So, this is the process we conducted to develop the data schema.

We started with a hierarchical structure comprising 4-levels: Module, Theme, Domain and a Variable level. We held a workshop with local clinicians and researchers interested in outcomes of very preterm birth to identify key topic areas. We then reviewed the literature to develop Modules and identify Themes and Domains before carrying out a consultation with RECAP Preterm partners to obtain levels of agreement as to the overall structure of the data schema and the content of each Module, Theme and Domain.

Twenty-three European clinicians and researchers experienced in carrying out research to understand and improve outcomes for children and adults born very preterm took part in the consultation where components of the data schema underwent two iterative rounds of consultation and levels of agreement were assessed pre- and post- round of consultation.

The schema was uploaded to the platform to provide a searchable structure for the cohort data available and all individual variables were added to the schema as a variable mapping activity, which I will come to shortly.

Slide 5 - The RECAP Preterm data schema

This is an example of the data schema in practice, with a Module of 'Antenatal & Birth', an accompanying Theme of 'Birth', and the Domain of 'Infant Characteristics at Birth'. The individual variable 'Infant Head Circumference' populates the Variable level as would harmonised variables generated from individual variables.

We created Domain definitions, and Variable definitions used by individual cohorts were included to provide detail and facilitate harmonisation.

Slide 6 – The RECAP Preterm data schema – pre- and post- consultation

This table shows the 10 Modules on the left which were created prior to the consultation and those on the right were a result of the consultation with expert partners with 14 Modules, 93 Themes, 345 Domains.

Slide 7 – The RECAP Preterm data schema

In line with feedback from the consultation, we also included cross-cutting taxonomies to capture additional variable-level information so those relating to Target (with whom the data relates),

Source (who collected or provided the data), Mode (how the data were collected for example by questionnaire, interview, register or routine data) and Instrument.

Due to the multigenerational nature of very preterm birth cohort studies, we included generation information to Target and Source to distinguish between the very preterm or control infant and their parents for instance.

For the instrument taxonomy, we referred to the cohort study protocols and questionnaires to identify instruments or measures used and that now exists as a project-specific taxonomy comprising 7 Modules each covering different topics.

A link to the schema is provided here.

Slide 8 – Variable Mapping

So how are the cohort data dictionaries and the data schema related?

As I mentioned previously, it is the individual variables that populate the Variable level of the data schema so the data dictionaries are used to map or tag variables to the data schema using a series of standardised dropdown lists. This enables each variable, and it's associated mappings to the data schema, to be uploaded to the platform. So variable mapping is incredibly important as it allows users to search for variables of interest on the platform.

So we map each variable in the data dictionary to a Module, Theme, Domain and Target, Source and Mode, and Instrument where applicable. Cohort study-level metadata and detailed variable definitions help with the variable mapping by providing context for each variable.

The data schema continues to develop as retrospective variables are mapped and harmonised variables are created.

Slide 9 – Mapping Instructions

We mapped all of the individual variables in each data dictionary on the cohort investigators behalf as part of the RECAP Preterm project but going forward, cohort investigators will need to do this for themselves to enable inclusion of their cohort on the platform.

We have therefore written instructions about data dictionaries and mapping variables to the data schema to enable cohort investigators to create and complete their own data dictionaries and to map variables to the data schema.

A link to these instructions is provided here.

Slide 10 – The RECAP Preterm Data Schema

The data schema on the platform as I said has a search function to allow variables of interest to be easily searched by Module, Theme or Domain and it is freely available for use worldwide to facilitate research to understand the long term impact of very preterm birth and to inform the development of future preterm birth cohort studies.

Slide 11 – Our experiences with data dictionaries and variable mapping

Over the past 4 years, we have worked with over 100 cohort data dictionaries and mapped over 22,000 individual study variables to the data schema in order to populate the platform for others to use.

- In our experience all of this preparatory work takes time so ensure you allow for this in your research timetable! For instance, collecting variable-level metadata from cohort investigators in the form of a data dictionary is dependant on them knowing what variables they want to make available to platform users and having the time to prepare a data dictionary or have the staff to do this for them.

- Data dictionaries vary in length and can change over time. We found that the number of variables in a data dictionary varied between cohorts and between different stages of follow-up within a cohort, so some data dictionaries had only 15 variables that needed mapping to the data schema whilst others had over 3000 variables. We also found that sometimes data dictionaries changed over time as new variables were added and existing variables, that we had already mapped, were removed.

- Mapping each individual variable to the data schema takes an long time, particularly when you think we had over 22,000 to map! Now that the platform is up and running, mapping variables to the data schema will be easier than it was for us doing it in an Excel spreadsheet.

- We found that labels and variable definitions in data dictionaries are really important. They need to be clear and unambiguous so that we can make decisions as to where in the data schema they should be mapped for instance and for platform users to understand. To help with variable mapping, we also referred to study protocols and questionnaires used to collect the data but these were not always available and often needed to be translated from a local language to English.

Slide 12 – References and Further Information

References and further information about data dictionaries, the data schema or variable mapping can be found here.

Slide 13 – Thank you

Thank you for listening.