Transcript - WP3 – Introduction to Harmonisation – Charlotte Powell

Slide 1

Hello, my name is Charlotte Powell and I will be giving you a brief introduction to harmonisation on behalf of the WP3 team.

Slide 2 – Learning Objectives

By the end of this session, you should:

- understand why data harmonisation is important?
- understand the RECAP Preterm approach to harmonisation
- know where to find the RECAP Preterm harmonisation guidelines
- understand the importance of your research question
- Be able to identify some key steps in the harmonisation process including:
    - What do you need to consider when harmonising data?
    - Ideal variables versus reality
    - What is a harmonisation dictionary?

Slide 3 – The benefits of data harmonisation

Firstly, to think about the benefits of data harmonisation. There are many benefits to harmonising data, here are just a few:

- Most simply, harmonisation enables assembly of large amounts of data across different studies.
- Harmonisation increases generalisability of results – by bringing together data from different areas and regions, your data then reflects broader geographic areas and also cultural and health system adversity. So, increasing confidence that outcomes observed in your results are less likely to be due to localised variations
- Harmonisation increases sample sizes – this is of particular important when you want to look at rare outcomes
- You can immediately validate and replicate your findings – you find one outcome across a few cohorts but how does this compare across other cohorts – is it a cohort specific finding or can you replicate it?
- You can assess differences across countries – so looking at different treatments or policies that are based in different countries.
- Within RECAP preterm particularly, you can consider changes over time. With a range of cohorts starting from 1978 all the way up to 2011, you can look at how different treatment initiatives have changed over time.
- Although harmonisation is not easy and does take a lot of time, it is still cost effective when you compare it to the idea of establishing a new study.

Slide 4 - The RECAP Preterm approach to harmonisation

RECAP preterm was established to provide a platform for data sharing on an ongoing basis. Different work packages within RECAP had specific analyses planned but the overall aim was to provide a platform, methodology and guidance for ongoing use of the data sharing mechanisms that had been developed within RECAP. This was to ensure the usefulness of the work completed in RECAP preterm outlives the funding agreement.

With that in mind, the focus of the harmonisation work within WP3 of the RECAP preterm project has been to document what it is available across the 23 European cohorts, organise the information in a systematic away, as Deborah and Helen have told you and assist WP4 with making this data available on the platform. The also to provide guidance, both for the other work packages on RECAP preterm but also those who wish to use the RECAP preterm data in the future. This meant our approach was different to similar platform studies which has not come without it's challenges.

However, this leaves us with a lasting legacy providing hopefully everything people need to conduct harmonisations on the platform going forward.

When we started RECAP Preterm, it had been noted by people like Fortier et al and Rolland et al, that harmonisation studies up to this point were not providing sufficiently detailed information on the steps involved in harmonisation. So, we have taken this on board and have tried to provide the platform but also provide information and guides to help people conduct harmonisation going forward. And also, to facilitate reuse of harmonised data that other researchers have created on the platform already.

Slide 5 – RECAP Preterm harmonisation guidelines

We developed Harmonisation Guidelines to identify stages in the process of going from a research question to harmonised variable on the platform.

In all, we identified 15 steps from research question to the data having been harmonised on the platform ready for analysis.

We are not going to go through them in detail here as these guidelines are available for you to view and use yourselves.

Also, you will be seeing different sections of them in presentations from other groups involved in RECAP. The team from Porto will be demonstrating how the later steps of the harmonisation on the platform in their presentations and the ECR group videos will give you an example of a harmonisation which followed this procedure to go from an idea and a research question through to harmonised data and analysis.

Slide 6 – RECAP Preterm harmonisation guidelines

Here, we are just going to look at a few key points that you need to consider and be prepared for when doing a harmonisation using the RECAP Preterm platform. We are going to focus largely on Steps 1 to 4 which take you through identifying the cohorts you are going to include in your study and then, working through from your target variable to a variable that you can achieve with the available data.

We will then briefly consider steps 6 and 7, which are the starting steps in your harmonised variables coming to life.

Slide 7 – The importance of defining your research question
But, before, we look at the harmonisation guidelines I want to take you back a step – before you even start working on the harmonisation, as with any other research project, your starting point has to be your research question.

What question do you want to use your harmonised data to answer?

It is a common misunderstanding that you can take a pool of data, harmonise it and then see what question you can answer with the data that you have. In reality though, like in a project where you set out to gather data, it is your research question that will determine which cohorts you can include in your project and what variables will end up looking like.

Yes, by all means, take a look at the data that is available to help determine your research question but by the time you are coming to harmonise the data, you need to have a clear research question in mind. This research question might be affected a little by the reality of the data harmonisation but you need a clear destination as your starting point.

Slide 8 – What do you need to consider?

Many of the things you need to consider when starting a harmonisation study will be similar to in other research projects and determined by your research question –

Before you get down to the detail of the variables, you need to consider the bigger picture and think about the cohorts as a whole - in line with your inclusion and exclusion criteria. This is Step 1 of the harmonisation guidelines.

What is the population that you want to look at?

- So, what is your key exposure – do you want just Preterm births, studies that include very low birth weight, or are you including both?
- What time-period are you looking at? - Currently, the RECAP platform includes cohorts starting in 1978 all the way up to 2011. However, changes in antenatal and neonatal care may mean you want to select a particular time-period.
- Are you looking at variables relating to the perinatal period or follow up in child-hood or in adult hood?
- Regional considerations. E.g. do you want to consider country wide e.g. registry data, or smaller more localised cohorts? Are you taking into account differences in management of preterm birth, education etc across the different regions?
- Do you want to include cohorts that just look at live births or are you interested in still births and TOPs as well?

Slide 9 – What do you need to consider?

You also need to think about methodological issues in the original studies:

- Is the timing of recruitment important to your study? Do you need mothers recruited when they were pregnant, or are you starting from neonatal care, or even discharge from neonatal care?

- Is how the data was collected important e.g. via Questionnaire, Interview, Observation, Physical Assessment etc
- Is the use of a certain standardised measure important for your analysis?

Each of the cohorts/registry studies on the RECAP Preterm platform is different – you need to spend some time familiarising yourself with the cohorts, which fit with your research question and which don't before you move on to look at the study data in more detail.

Slide 10 – Cohort metadata on the RECAP platform

This is where the metadata on the platform that Helen was describing comes in. We have gathered together as much information about each of the cohorts studies as possible to enable you to make informed decisions as to which cohorts will fit with the research that you are planning to carry out.

I am sure you will also hear more about this in the presentations by Porto team, WP4.

Slide 11 – Ideal versus reality
Harmonisation is very much a top-down meets bottom up approach.

The top down approach - You start with your research question and your conceptual knowledge of the area and say right, in order to answer this question what variables do I need? If I was setting out to do this study from scratch, how would I define that variable? And as Deborah mentioned previously, definitions are very important in harmonisation, they need to be clear and exact. This is step 2 of the harmonisation process and, tempting as it is, to just dive right into the data, it is important that you pause and identify what your ideal variable definition would be.

Ordinarily, you would go away and devise a method to collect this data. But, in harmonisation, the data has already been collected for you. So, in this instance you need to start to compare what your ideal variable would be, to the reality of what is available within the data that has been collected. This is the bottom-up element and is considered in steps 3-6 in the harmonisation guidelines which are going to look at now.

Slide 12 – How to find the variables – schema

So how precisely, does the bottom up approach work? How do you identify your variables?
The RECAP platform, when it is finished, will have over 23,000 variables available on it.

By using the schema and mapping work that Deborah told you about, you can pull from the platform the variables that are in your area of interest. So you start by identifying the module you are interested in, then you look at what themes are within that module, then this gives you a list of domains that are withing those themes. You pick which domains you are interested in and use the platform to search and pull out all of the variables that you are interested in.

If you look at the wiki we conceptualised a process to help you assimilate all of the information that is available on the platform, to narrow it down to look at precisely what variables are available in relation to the area that you want to look at.

Once you have the variables, it is a case of sifting through them and seeing which variables across which cohorts, can you use to reach your harmonised variable.

Slide 13 – How to find the variables – schema

You also need to consider any standardised measures that may be relevant to the subject area. Again, these are easily identified using the schema. You look at the instrument module, then theme and domain and that will put out a list of the standardised measures.

Slide 14 – Down to the detail – getting an achievable variable – things to consider

Once you have identified the variables/standardised measures that you think are relevant, you then need to consider which of the available variables can be harmonised to achieve your target variable.

This is done simply by spending time looking at the variables that are available – how did the cohort collect that particular information, what are the different definitions of that variable across the cohorts?

You need to think - can you use the variables that are available to create your ideal definition? Most likely not in the first instance, then there are something that you can consider -

- If the target definition is not achievable for many cohorts with simple re-categorisation of the data then consider:
  - the number of cohorts that have data that could be harmonised – would this be enough?
  - whether other variables from a cohort could be used to create a variable that is harmonisable e.g. calculating lengths of time from dates etc.
  - the next best variable that you could create which would include more cohorts e.g. by reducing the number of categories or by broadening the definition.
  - can more or alternative variables be created? E.g. one detailed variable with your ideal definition but then another broader variable to include more cohorts
  - whether advanced statistical techniques could be used to allow the inclusion of more cohorts.
- Consider how useful a broader definition would be, but this will need to be considered in relation to your research question – the definition still needs to be precise enough to answer your research question, but also broad enough to work with the data that is available from the cohorts.

Slide 15 – Ideal versus reality

As I said before, harmonisation is a bottom up versus top-down approach, so it ends up being iterative. You start with your ideal definition, the top-down approach, but then you look at the reality of the data, the bottom up approach.

You consider what you can achieve, but you can't get quite what you want so you revise your ideal definition. Then, you look again at the data and see if is achievable with the data that you have available. If not, then you revise your definition again.

It might be that you come up with something that you think is achievable but then when you check with the cohort, you misunderstood their data or their definition and so you need to revise your ideal definition again.

You keep doing this iterative process until you finally reach a variable that is as close to your ideal variable as you can get but takes into account the reality of the data that is available.

You need to constantly be assessing your variable definition in the context of your research question, so for some questions you will need a strict definition (which may mean excluding studies) but for others, you can relax the target definition and include more studies.

But, this process is important as when it comes to writing up your study, or sharing your harmonised data with other people, you need to be able to explain how you got to the definition that you finally used. What would you have liked your variable to look like, why wasn't that possible, and what did the variable end up looking like in the end?

Slide 16 – e.g. of iterative process – Gestational Age

So this is an example of iterative development of a definition.

Please note, this example is not necessarily a full accurate representation of how gestational age is measured within the RECAP cohorts but is rather an illustration of how you may have to revise a planned variable definition.

So, looking at Gestational Age. You determine your ideal variable, based on your knowledge of the area, and your research question, your starting, ideal variable definition is:

Gestational age in completed weeks and days of gestation - based on early pregnancy ultrasound dating.

However, when you look at the definitions of gestational age within the cohorts you see they have used different definitions for gestational age, e.g. some of have used dating based on pregnancy ultrasound, while others have used last menstrual period, or estimated at first prenatal consultation. So, you decide to broaden your definition to cover all of these so your new definition becomes:

Gestational age in completed weeks and days determined by the obstetrical team caring for the pregnant woman.

Then, you look at the definitions again and note that there are variations between the cohorts on whether they have recorded gestational age in weeks and days, or just in weeks, or even used categories, so recorded 36 weeks or more, 32-36 weeks and less than 32 weeks. You could create 3 possible definitions now with different numbers of cohorts included –

Gestational age in completed weeks and days determined by the obstetrical team caring for the pregnant woman – 5 cohorts => 5 cohorts can be harmonised.

Gestational age in completed weeks – 2 cohorts => 7 cohorts can be harmonised.

Gestational age categories – 1 cohorts => 8 cohorts can be harmonised. Your categories for this will be determined by the categories used by the one cohort that used categories. For the others, you generate a variable from the number of weeks.

At this point, you need to refer back to your research question and see which is useful for you – how useful is it to have a broad definition which includes more cohorts? Are you better with sticking with less cohorts and using weeks and days or are you better using completed weeks? Or, is that last cohort that only includes gestational age in categories is key to your analysis? If so, you go with the gestational age in categories.

Slide 17 – Don't forget the cohorts

Talk to the cohorts, check out any assumptions you have made, ask them if they have any other variables that aren't on the platform that you could reach that variable with?

Remember, the cohorts know their data the best. Use that knowledge and keep them involved.

Check your understanding of the variables with them – we think we can do this to make your variable – do you agree? Have we understood your data correctly? Is there anything we are missing?

Slide 18 – What is a harmonisation dictionary?

Finally, once you have come to a list of variables and definitions that achievable for the cohorts and will answer your research question, your work moves back to the RECAP Preterm platform and you create a harmonisation dictionary.

So, what is a harmonisation dictionary?

A harmonisation dictionary is simply, a data dictionary (as Deborah described) but this one contains the information about the variable that you will be creating from your harmonisation. So this is the information about your target variables that you have identified through the iterative process of comparing your ideal with the reality till you arrive at a variable that you can achieve.

It includes the same information as in the data dictionaries - the variable name, label, a detailed description/definition of the variable, unit and value type. Like for the cohort data dictionaries, we also mapped these new variables, to the correct section of the data schema, so making the variables searchable within the data platform for people coming along behind you.

Once your harmonization dictionary is complete, it will be uploaded to the platform and as the data is harmonized by the cohorts, their data will tie back to your harmonization dictionary.

WP4 will take up the story and talk you through the practicalities of conducting a harmonisation on the platform.

Slide 19 – Next steps in the harmonization journey

Other key summer school sessions to watch

- Work Package 3
  - Metadata catalogue
  - Anatomy of a data dictionary and the development of a data schema
- Work Package 4
  - Harmonising data across multiple studies
- ECR group
  - Harmonisation for the EPT demonstration project

Links

- Harmonisation guidelines -

https://gitlab.inesctec.pt/wp4-recap/wp3/-/wikis/guidelines

Slide 20 – Our experiences

To sum up, here are some top tips based on our experiences.

- The process of harmonisation is as important as the results you get out of it.

- Record keeping is key to enable others to see how you did the harmonisation – to save duplication of effort and enable sharing of variables across studies.
- Don't underestimate the time this takes
- Cohort leads are your most important source of information through this whole process – work with them, ask them questions, they are sharing their data with you because they are interested in your work, involve them and make sure you use their data correctly.
- Beware of making assumptions.

Slide 21 – Thank you

Thank you for listening. In case of questions email me or the wider team at the University of Leicester and we will be happy to help.

Slide 22 – References

Thank you for your time.