Report

closer
The home of longitudinal research

A review of quantitative analytical training needs for users of longitudinal studies

Jane Maddock, Dara O'Neill, Meghan Rainsberry, Rebecca Hardy
CLOSER

October 2019

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
MRC Medical Research Council

Contents

1. Introduction 2

2. Review strategy 3

3. Results: Mapping existing training 5

4. Results: Consultation of key stakeholders:
Current users of longitudinal data 8

5. Results: Consultation of key stakeholders:
Supervisors of users/future users of longitudinal data 15

6. Results: Consultation of key stakeholders:
Senior academics 18

7. Recommendations and future directions 19

1. Introduction

The ESRC makes considerable investments in longitudinal studies which support the understanding of population trajectories over the life course in changing contexts. In the 2017 Longitudinal Strategic Review[1], it was noted that despite existing initiatives such as the National Centre for Research Methods (NCRM), Q-Step and the ESRC Doctoral Training Network, training capacity needs to be improved. This training is fundamental to building the skills required to use these longitudinal studies effectively and to their full potential. The ESRC have commissioned CLOSER[2] to conduct a review of these training needs.

The aim of this report is to review the provision of quantitative data analyses training for users of longitudinal studies and to provide recommendations for improvement to the availability and scope of training currently available. In this report, we define 'users' as students, academic and third/public sector users. The scope of this review is training provided at the post-graduate, early-mid career and continuing professional development level, and specifically training that is open to all users (i.e. not part of closed provision, such as enrolled MSc courses).

In addition to this report, the ESRC have commissioned further reviews as part of a comprehensive UK Training Needs Assessment, including an NCRM consultation on the wider training needs of the social science community in the UK.

[1] https://esrc.ukri.org/files/news-events-and-publications/publications/longitudinal-studies-strategic-review-2017/
[2] https://www.closer.ac.uk/

2. Review strategy

The activities for the review were two-fold:

1. Mapping existing training
2. Consultation of key stakeholders

Mapping existing training

A list of relevant training provision was created through desk-based research and consultation with the ESRC. It was not possible to create an exhaustive list of all training available. The following sources were searched for relevant training:

- National Centre for Research Methods and their associated partners:
    - University of Southampton (Centre for Applied Social Surveys [CASS])
    - University of Manchester (methods@manchester) / Cathie Marsh Institute for Social Research [CMI])
    - University of Edinburgh (AQMeN)
    - WISERD (Wales)
    - ARK (Northern Ireland)
    - Department of Methodology at the London School of Economics
- Cohort and Longitudinal Studies Enhancement Resource (CLOSER)
- Centre for Longitudinal Studies (CLS)
- Understanding Society
- Institute for Social and Economic Research, UK Longitudinal Studies Centre (ISER ULSC)
- Centre for Longitudinal Study Information and User Support (CeLSIUS)/ The Census & Administrative data Longitudinal Studies Hub (CALLS-Hub)
- UK Data Service (UKDS)
- ESRC and MRC funded longitudinal studies themselves
- UK universities

Consultation of key stakeholders

Current users of longitudinal data

We designed an anonymous survey to capture responses about analytical training needs from a wide range of current users of longitudinal data. The survey was advertised via relevant

mailing lists (e.g. AllStat, CLOSER, early career mailing lists), government contacts and Twitter. There were 304 responses.

Supervisors of users/future users of longitudinal data

On the 22nd May 2019 CLOSER, Understanding Society, CLS and the UKDS ran a free one day workshop giving an overview of longitudinal data and related teaching resources available to academics who teach and/or supervise students in quantitative social science subjects. There were 14 attendees from a range of social science disciplines who had varying degrees of knowledge about longitudinal data. In the afternoon session, CLOSER facilitated small group work (3 groups) to identify the key training needs and/or barriers for users of longitudinal data. Each group had approximately one hour to discuss six questions:

1. Why do/don't your students use longitudinal data?
2. What are the key analytical methods required for longitudinal data analyses and which ones do you think are the most difficult for new users of longitudinal data?
3. How would you advise new users to develop their skills / Where would you advise them to go?
4. Are there major gaps/overlap in training provision for longitudinal data analyses?
5. What do you think are the key barriers in accessing training?
6. What resources do you need as a trainer to effectively support your students with longitudinal data analyses/any recommendations for how training could be improved?

Senior academics

Finally, we consulted three senior academics who are involved in analyses of longitudinal studies and training provision.

3. Results: Mapping existing training

At CLOSER, we depict a typical training route for users of longitudinal data from learning that longitudinal studies exist through to advanced techniques and study management. From this we have identified three main categories of training to support the successful analyses of longitudinal data:

1. Data enabling, data manipulation and study-specific knowledge e.g. getting data ready for analyses, use of survey weights
2. Basic quantitative statistical methods (these may not be specific to longitudinal studies e.g. summary statistics, t-tests, regression)
3. Advanced quantitative statistical methods specific to longitudinal analyses e.g. multilevel modelling, structural equation models, missing data, causal analyses

We note that these skills are not always developed in a consecutive manner.

Over the last decade there has been considerable investment in quantitative methods training in the UK. In terms of structured training the ESRC, Nuffield Foundation and the British Academy have developed undergraduate quantitative skills through programmes like Q-Step. The ESRC have also developed initiatives such as the Doctoral Training Centres which provide a structured approach to developing quantitative skills at the postgraduate level. The type of training received at this postgraduate level depends on the discipline. For example, econometrics students tend to have more training in advanced statistical modelling than other social sciences. Therefore, beyond these formal stages of training on degree courses, there is still a wide range of training required to support and develop skills in all users of longitudinal data.

Our mapping exercise identified a non-exhaustive list of face-to-face short courses, workshops, and seminars as well as webinars, videos and online training resources to support all users of longitudinal data. Training tended to be based on a particular statistical method or in relation to a specific study.

Face-to-face training

The main source of information for face-to-face training for this review came from the ESRC's National Centre for Research Methods (NCRM) which was established in 2004. Previous assessments of research and training needs by the NCRM can be found on their website and they are currently undertaking an up-to-date review[3]. The most common categories of training covered by the NCRM were basic and advanced quantitative statistical methods. The basic analytical methods included introductory statistics (e.g. t-tests ANOVA) and regression models. London was the predominant location for these short courses which lasted between one and five days and incurred a fee.

The advanced methods included topics such as structural equation modelling, mediation analysis, multilevel modelling and survival analysis. These were predominantly held in London with institutions in other locations including Bristol, Manchester and Southampton also providing a number of courses. While these analytical methods can be applied to longitudinal studies that was not the main objective for most courses identified.

The statistical software used in training varied. While many identified training courses did not specify a software, the courses in the basic statistical methods category that did provided training mainly using SPSS, with a smaller number of courses using Stata and R. The courses providing advanced quantitative statistical methods mainly used Stata, with others using Mplus (for structural equation models) and a few providing training using SPSS, R and MLwiN.

Workshops/Seminars

Training for longitudinal data users is also provided in the format of free face-to-face workshops and seminars. These tend to be related to study-specific issues. For example, the UK Data Service (UKDS) provide a workshop to understand census microdata in Edinburgh and the Centre for Longitudinal Studies (CLS) provides a series of seminars on study-specific analysis relating to the birth cohorts in London. Understanding Society provide a range of workshops specific to handling their data. CLOSER provide a series of face-to-face seminars and workshops mostly addressing cross-study analysis and challenges, including data harmonisation and study-specific analysis. These workshops and seminars are advertised

[3] https://www.ncrm.ac.uk/publications/assess.php

through various media e.g. via the NCRM, CLOSER, CLS, Understanding Society, UKDS websites as well as through various mailing lists and Twitter.

Online/Videos

A number of training providers offer online materials to support analytical skill development. For example the Centre for Multilevel Modelling in Bristol was supported by the ESRC to provide free online training specific to analysing longitudinal data (Longitudinal Effects, Multilevel Modelling and Applications (LEMMA)). CLOSER have developed their open-access Learning Hub which provides guidance on the different statistical techniques applicable to longitudinal data analyses as well as guidance on the use of real-world data through a teaching dataset with worksheets developed in conjunction with CLS. CLOSER also make their in-person workshops available as webinars, disseminated through their own website and YouTube. The UKDS has recently developed 'Data Skills Modules' which provide online training for users of longitudinal as well as survey and aggregate data.

The NCRM have created online learning resources covering a range of statistical analysis techniques as well as hosting a repository from previous ESRC-funded training activities (ReStore) and signposting to external online content.

The longitudinal studies themselves provide a range of online introductory and training-related resources. For example, CLS provide a series of webinars including: introductions to the birth cohorts, special study-specific topics, and tutorials on data access. They also have a methods strategy online which provide users with guidance on best practices on topics such as dealing with missing data. Other resources and studies such as the Avon Longitudinal Study of Parents and Children, Understanding Society, CLS, CeLIUS, CALLS-Hub and UKDS provide online guides and video tutorials to support use and access of data they host and many also have a YouTube channel. Understanding Society and CLS have also developed extensive online training materials specific to use of their data, including teaching datasets, and provide and on-line help desk once a week to support users.

4. Results: Consultation of key stakeholders:

Current users of longitudinal data

Of the 304 respondents, 56% identified as a higher education researcher and 18% as a postgraduate student with the remaining coming from public sector (7%), third sector (7%), higher education professional staff (6%) and <5% from private sector, general public and undergraduate student. The majority of respondents (85%) had used a longitudinal dataset with 6% planning to use one in the near future.

(i) Key analytical skills required for longitudinal data

All three categories of training (i.e. data manipulation, basic quantitative statistical methods and advanced methods) were considered key skills required for the analyses of longitudinal data with data enabling/manipulation and advanced statistical methods scoring slightly higher. Other key analytical skills that were mentioned included:

- Understanding data collection processes and coding and how it impacts conclusions
- Software skills
- Panel data methods
- Data protection and sample management
- Linkage

(ii) Key barriers to successful analysis of longitudinal data

The top barriers for new users to successfully analyse longitudinal data were: a lack of specific analytical skills, inability to access or handle data, a lack of training provision for specific techniques and cost of training (figure one). Another major barrier identified was time for training. Other barriers mentioned were software literacy, physical access to data e.g. those held in a secure space, poor awareness of the complexity of longitudinal data and a mismatch between a supervisor's skills and time to support the postgraduate student.

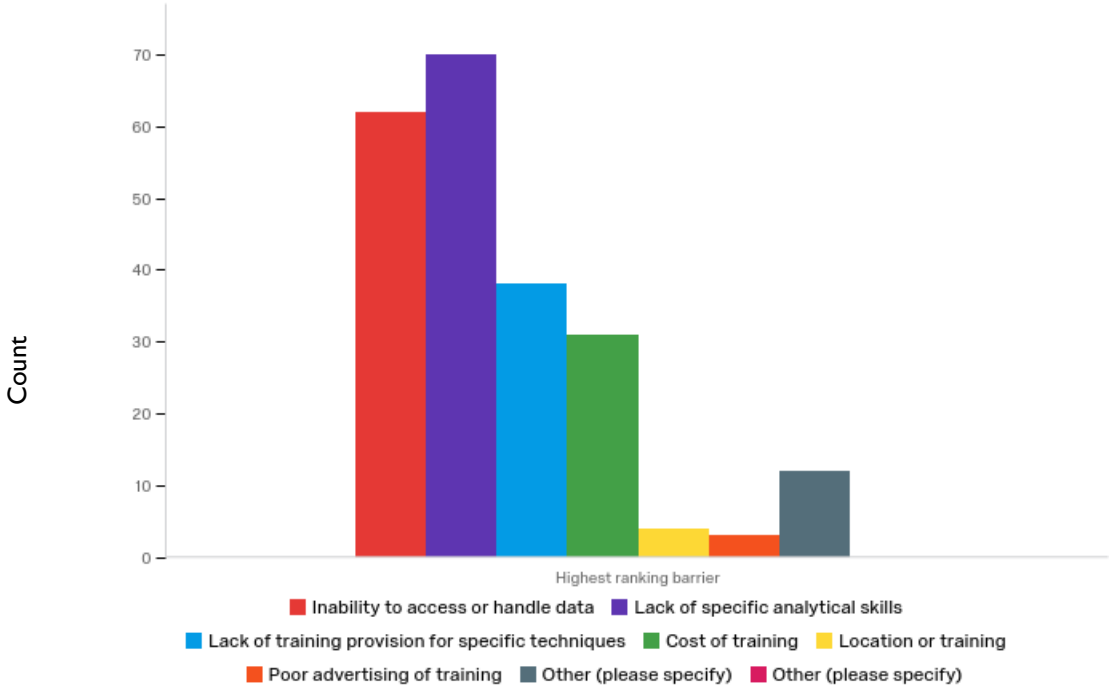

Figure one: Highest ranking barriers to the successful analyses of longitudinal data

(iii) Best modes for delivering training

Short courses were clearly identified as being the best mode for learning the analytical techniques required for longitudinal data analysis (figure two). This was followed by mentor or supervisor-led, self-directed online training and face-to-face seminars. Books and papers, user groups/advice forums, documentation and guides, hands on practice, and learning through data application with supervisor support where mentioned as additional important training modes.

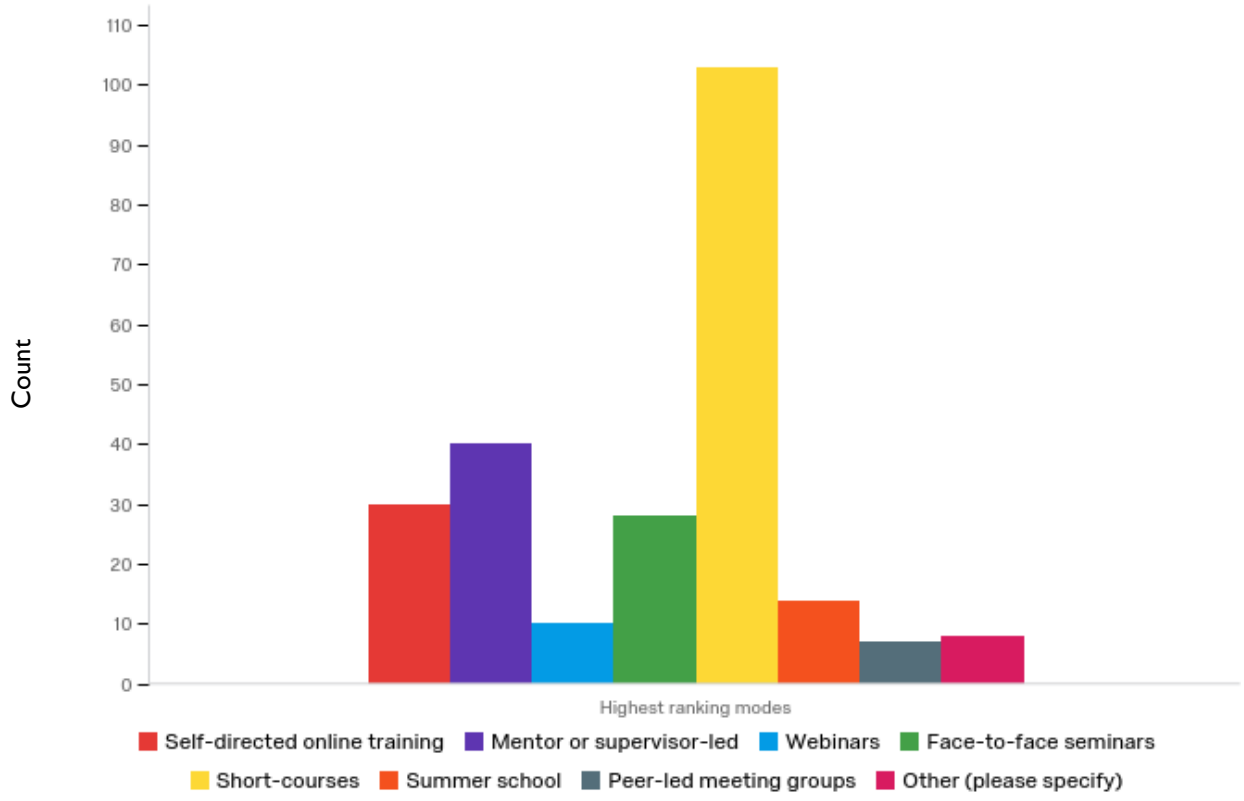

Figure two: Highest ranking mode for delivering training

(iv) Advertising of training

Most people found out about training provision through internet searches (22%), word of mouth (19%) and university/institute website (17%) (figure three). Of the 'other' category, the platforms mentioned included: Twitter, AllStat mailing list (note that this survey was distributed via Twitter and the AllStat mailing list), through the respondents' own company and through society websites (e.g. Royal Statistical Society and Society for Longitudinal and Life course Studies).

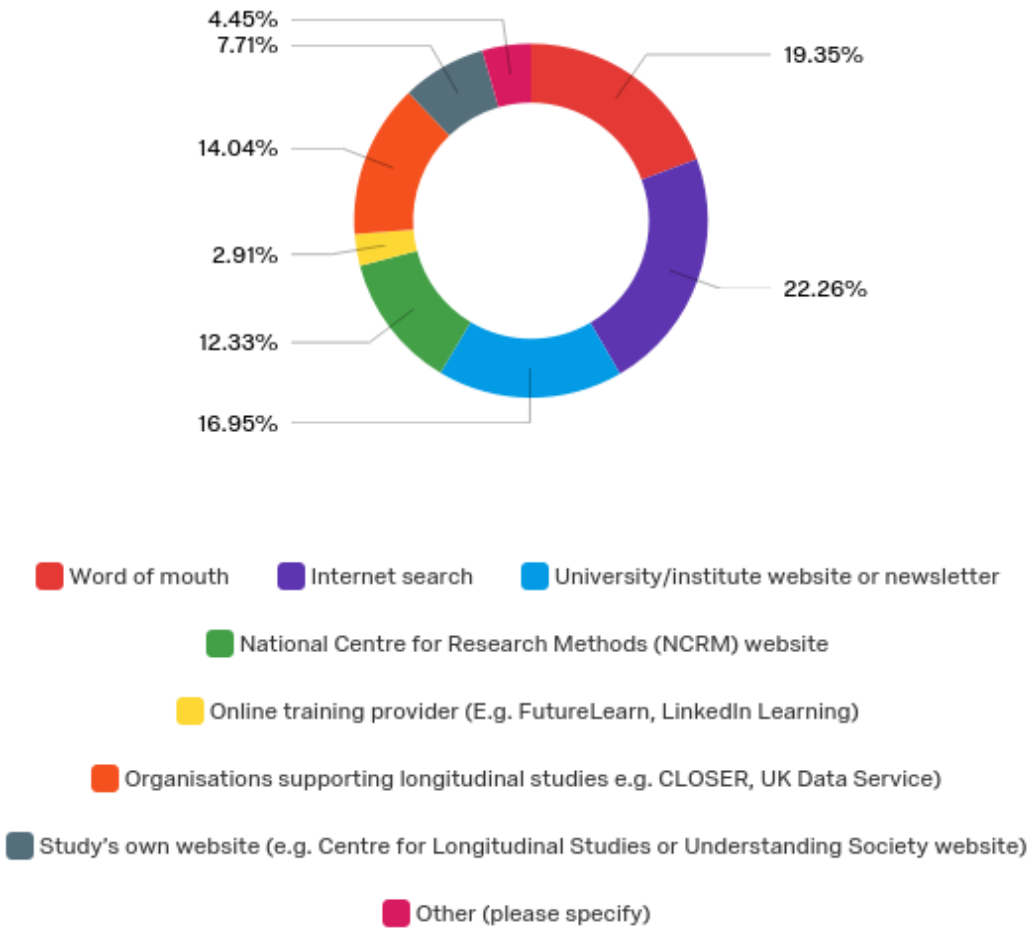

Figure three: Response to advertising of training provision

(v) Gaps/overlap in current training provision

Seventy nine percent of respondents felt there were no significant overlaps in training for specific topics for longitudinal training. However, an equal number of respondents felt that training gaps remained in other topics. These respondents were given the opportunity to provide a text response as to where they think the gaps are. The following major themes were identified:

- Lack of training in specific analytical techniques directly applicable to longitudinal data e.g. causal analysis, growth curves, missing data, multilevel models
- Lack of training/support for data handling/manipulation

- Lack of real-life example of the whole research process from start to finish
- Lack of 'messy' datasets when learning
- Lack of ongoing support and/or mentorship at all career levels
- Lack of software literacy for longitudinal analysis
- Lack of support for mid-career training

The following quotes were taken from two respondents to this question:

"Very little for data handling and data management, which in my case had to be self-taught/trial and error for nearly 6 months for my PhD research. Most analytical training (either in short courses or in summer schools) begins with datasets that are already cleaned and synthesised (which is necessary for short courses to learn the fundamentals of modelling); however, there is very little face to face training and assistance available for people to practice synthesising raw data from UKDS files and then ensuring that everything is adequately coded and cleaned for analysis. This was a HUGE part of my work when I had to work with the [data], which had different coding structures and names, for example. Analytical modelling techniques (for the most part) are covered relatively well in short courses and online, but it might help to have resources available for analysts to practice and ask questions of experts on this important data management element of long analysis."

"Intermediate stuff. There's loads of webinars etc on "how to download [data] from the UKDS", then incredibly advanced techniques, but not clear how to get from point a to b..."

(vi) Recommendations

Ninety six respondents provided recommendations on how training for longitudinal analyses could be improved. Figure four shows a frequency-based word cloud of the key terms mentioned. The main themes that emerged from responses to this question were:

- Additional resources to support training particularly for freely available online content in which users can work at their own pace
- More training specifically targeted at longitudinal data
- Increase advertising/awareness of current training
- More training/support for initial data handling

- Ongoing support including forums and refresher courses
- Make training more affordable and provide funding opportunities to support this
- Increase the number of locations and frequency of short-courses
- More training of the full research process from data collection to dissemination with real-life examples/messy datasets

An important point was that training the trainer is often overlooked. Recommendations were made for more resources targeted at mid-career as a way to support their own development and to enable them to supervise early-career colleagues effectively.

Figure four: Key words for recommendation on how training to analyse longitudinal data could be improved

The following quotes were taken from a number of respondents to this question:

- *"Better advertising (e.g., at conferences), courses focusing on analysing multiple longitudinal studies"*
- *"More online materials to support initial access to data and how to handle data"*
- *"More short-courses offered at reasonable prices & at regular intervals. A lot of courses are really expensive"*
- *"Free courses, more courses, bursaries…to attend, site that has a complete list (as far as possible) of relevant courses"*
- *"[Training] would be deeply appreciated by many researchers at faculty level so that they can adequately support their postgrad students"*

5. Results: Consultation of key stakeholders:

Supervisors of users/future users of longitudinal data

The main themes emerging from the group discussion of the 14 attendees during the workshop were:

(i) Initial access to the data

While the main objective of the discussion was on analytical training needs for new users of longitudinal data, it was evident that a fundamental step is the knowledge that the data exist and how data can be accessed. This lack of knowledge was applied to supervisors as well as students. It was stated that major barriers to accessing data were fear and a lack of confidence from both parties in terms of how to handle longitudinal data and a perception that secondary data analyses may not be as worthwhile as collecting primary data in terms of addressing a novel question. It was suggested that more "selling" of the benefits of analysis of existing longitudinal data beginning at the undergraduate and MSc level would be beneficial. Another major barrier was the perceived transition from theoretical quantitative methods to application in a longitudinal dataset. This barrier was noted even for those students and supervisors who are proficient with advanced statistical methods.

(ii) Current available training

The training needs for new users of longitudinal data are discipline-specific and dependent on their starting level, and it can be time-consuming to upskill. It was recognised that the key analytical methods required included basic statistical methods (e.g. descriptive statistics, t-tests, regression models) and more advanced methods (e.g. mediation analysis, structural equation modelling, handling missing data and multilevel models). There was a perception that while there is an array of training available for the basic methods, the gaps in provision are wider for more complex methods that are specific to longitudinal analyses. It was also noted that while SPSS tends to be the software of choice for basic statistics, more advanced users rely on Stata or R. This can be a barrier when trying to upskill for both the student and supervisor and

may require students to be competent in more than one package. Participants noted a major gap in current training for data-manipulation and understanding the data structure for specific datasets. The worked examples and online training resources provided by Understanding Society were cited as being extremely useful.

Key barriers in accessing face-to-face training were cost, location and time. Quality online training content was mentioned as a way to overcome all these barriers, with University of Bristol's LEMMA, the CLOSER Learning Hub and analyses-themed or study-themed YouTube channels cited as examples of useful resources.

Promotion of training provision was also discussed. The majority of the participants have used 'word of mouth' to find out about training. A central resource that could signpost supervisors to quality training was advised. While the NCRM already provides such a service, the majority of participants at this workshop were not aware of it.

The role of informal and self-directed training were identified as making a significant contribution to upskilling for both students and supervisors. Knowledge from supervisors, mentors and peer groups were cited as important sources in understanding study-specific data and analytical issues e.g. how and when to use survey weights.

(iii) Recommendations for improving training

Some key recommendations to improve training for new users of longitudinal data from the group discussion were:

- Promote the potential of longitudinal data from an early stage (e.g. speaking to undergraduate and MSc students from a range of disciplines)
- Provide training and resources (including funding) specifically aimed at mid-career researchers who will be supervising new students
- Provide more training in data manipulation/data management. These would be most useful if specific online content was created for each study
- Create 'messy' teaching datasets so that new users can get a realistic view of the data structure and what is required to clean data

- Develop online training materials that address different topics which document the complete research process: idea formation, data access, data management, results, dissemination
- Further develop free training content
- Make resources available in multiple statistical software
- Increase the promotion of existing training available through activities such as departmental newsletters, statistical support units, pre-conference workshops
- Establish a mentoring system for users of longitudinal data for both supervisors and students

6. Results: Consultation of key stakeholders:

Senior academics

All academics agreed on CLOSER's categorisation of training: (1) data enabling and data manipulation (2) basic quantitative statistical methods (3) advanced quantitative statistical methods specific to longitudinal analyses. All of the academics specified that data enabling and data manipulation are particularly challenging for new users and it was noted that if someone becomes skilled in training categories one and two they are more likely to be motivated to use more advanced methods. It was also mentioned that methods for longitudinal analyses, e.g. mediation, are becoming more complex and it is difficult for one person to be highly skilled in every method.

In terms of training currently available, the academics noted that there are courses available via NCRM, online resources and various summer schools. However these are more frequent for cross-sectional rather than longitudinal analyses and there is a gap in training provision for data enabling and data manipulation. An additional gap was also noted for harmonisation and cross-cohort analyses.

When guiding their own students, these academics direct them to relevant textbooks and attendance at intensive summer schools. However they noted that courses are costly, not offered very frequently, have travel constraints and get filled up very quickly. They also mention that good supervision and training on the job were key components for new users of longitudinal data. These new users need to build an understanding of how to deal with 'messy' datasets and how to document and log their code and analysis. The importance of using 'real-life' examples of research and helping researchers to understand how data was collected and coded was emphasised. Lack of staff resources and time were identified as key barriers for the successful provision of quality training. The necessity for much greater input and support from funders was highlighted as a particularly important requirement in addressing current challenges in the training provision for users of longitudinal data

Part 7: Recommendations and future directions

Using the information gathered in this report, CLOSER makes the following recommendations to improve training for longitudinal data analysis:

i. Develop a central resource dedicated to longitudinal data analysis training, building on current initiatives and signposting to existing resources (e.g. longitudinal training advertised through NCRM).

ii. Develop a coherent and comprehensive training pathway for longitudinal data analyses from data discoverability through to advanced analyses.
    - Establish a complete pedagogical framework for longitudinal data analysis training
    - Provide more 'real-life' examples of the whole research process using existing data
    - Develop new training on data handling and data manipulation, identified as a major gap in this pathway, including the creation of 'messy' datasets based on existing data that can be used for training

iii. Improve training for staff supporting early career researchers.
    - Increase training provision for mid-career researchers/supervisors so they can effectively support early-career colleagues in the use of longitudinal data
    - Develop a mentoring scheme to provide early-career researchers with study-specific support

iv. Remove barriers to training access.
    - Deliver training across multiple formats such as interactive teaching resources and events offered both face-to-face and online
    - Increase the provision of open-access training materials
    - Reduce cost of in-person courses and concurrently provide more funding opportunities to attend training courses at all levels

A coherent funding strategy is essential to maximise the impact of these recommendations.