More info on LIFECYCLE online:
lifecycle-project.eu

Task 1.3. Approaches and methods for designing, implementing and maintaining pregnancy and child cohort studies

Deliverable D1.3. Recommendations for approaches and methods for designing and maintaining pregnancy and child cohort studies.

Delivery date: June 30th 2021

Version 2.0

Authors: Anne-Marie Nybo Andersen, Marie-Aline Charles, Hazel Inskip

LIFECYCLE Project
Coordination and leadership:

Erasmus MC
Generation R (Na-2918)
PO box 2040
3000 CA Rotterdam
The Netherlands

Phone: +31 (0)10 704 3405
Email: lifecycle@erasmusmc.nl
Web: lifecycle-project.eu

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 733206 (LifeCycle).

Contents

I. Introduction

In this report, we aim to describe successful and less successful approaches and methods for designing, implementing and maintaining pregnancy and child cohort studies. This includes issues related to recruitment and retention, questionnaires and methods, collection and storage of biological samples, data management and ethics.

The purpose of this report is to document and highlight the dilemmas and necessary decisions that experience from the LifeCycle cohorts has identified. The aim is to provide guidance on a range of issues for those setting up and working on cohort studies. The report focuses on the experiences from types of cohorts that make up the LifeCycle consortium, but many of the topics discussed apply to cohorts of any type. This is not a comprehensive guide to conducting cohort studies, as there are many epidemiological textbooks that provide such guidance. However, the information here may complement the standard introductions to cohort studies, to assist those developing and running them.

II. Types of cohorts

Overall purpose of a pregnancy/child cohort

In principle, we have three types of pregnancy/child cohorts: life-course cohorts, cohorts specifically set up for studies of pregnancy and perinatal health, and cohorts set up for studying child health, often specific child health outcomes. The latter two types may develop into life-course cohorts, but ideally, the overall aim of a cohort should be a thread running through all decisions regarding the cohort, and it will inevitably affect the type of data collection that is given priority. Priorities have to be set, and all types of cohorts have their particular strengths and short-comings.

Life-course cohorts

Life-course cohorts endeavour to map at least part of the life-course, starting around birth. In LifeCycle and the EuroChild cohort work, the focus is generally on the early part of the life-course. The aim is to understand early life influences, from preconception onwards, on the health, development, wellbeing and social factors in the life of the individual. The growth in interest in the Developmental Origins of Health and Disease paradigm has led to many cohorts addressing influences during the first 1,000 days from conception to two years of age on later outcomes in childhood and beyond. Cohorts may combine biomedical and social science aspects, and cover a diverse range of exposures of interest, such as nutrition, lifestyle, genetics, epigenetics, social circumstances, parental physical and mental health and the wider environment.

Life-course cohorts cover a long time scale. This is challenging as there is a need to retain participants over many years and also to keep obtaining funding to maintain the cohort. Over time, the research focus may change. Questions that are important now may not have been considered when the cohort started, and thus crucial data may be missing. Further, the researchers who want to use data from cohorts with a long history may lack understanding of contextual matters around the time of data collection, which may impede the interpretation of the data. However, the long term focus does allow questions to be addressed regarding early life exposures on adolescent and adult health in a prospective manner and this is a strength of such cohorts.

Pregnancy/perinatal cohorts

Pregnancy and perinatal cohorts are largely established to address questions relating to pregnancy and birth outcomes or those in very early life. Follow-up is relatively short and keeping contact with research participants is not usually too difficult. Having obtained useful early life data though, there can be the temptation to continue to follow-up the cohort and then the study morphs into more of a life-course cohort, usually with rich pregnancy data.

Child cohorts for specific health outcomes

Some cohorts are established to consider particular health outcomes, though these have to occur commonly, as otherwise the cohort will need to be very large indeed to have sufficient statistical power. Common outcomes that can have been addressed by specific cohorts include childhood asthma, mental health and obesity. Such cohorts may be recruited in populations with a higher a priori risk of disease, e.g. children of parents with asthma, mental health problems or overweight. Again, such cohorts, with extended follow-up, can become life-course cohorts, and can assess other outcomes as the cohort develops.

Other types of cohorts

Many other types of pregnancy/child cohorts exist. Related to the LifeCycle work are cohorts that follow-up children born preterm or with particular congenital conditions or after Artificial Reproductive Technologies (ART). These various types of cohorts have a specific research interest attached to them, and they have been proved highly important. A key issue is, however, to have an appropriate comparison group, and it is advisable to consider establishment of a comparison cohort, ideally comparable on all other areas than the condition in focus and its risk factors.

Participants

The focus of a pregnancy/birth/child cohort is the child or the outcome from the pregnancy. Ultimately, it is the health, development and wellbeing of the child that is the subject of investigation, and she/he is the main participant. However, since the health of a child depends greatly on the environment in which it is born, raised and grows (cf. Rio declaration), a systematic collection of data that describes this environment is crucial (see chapters IV & VI). The environment of a child is partly determined by the parents, who also provide the genetic influences. Consequently, the parents are usually co-participants. Follow-up of the parents is tempting, as the cohort typically has plenty of information on the mothers and, to a lesser extent, fathers at baseline. It is, though, important to remember that these cohorts are selected on the ability to establish a pregnancy/get a child, which is both a biological and a social selection (on top of the selection resulting from being willing to participate in a research project).

The cohorts established in the 20th century were most often mother and child cohorts, for many reasons, including tradition and convenience. As the responsibility for upbringing increasingly is shared equally between the parents, we strongly recommend that both parents are involved in new cohorts.

Mother

For studies starting in pregnancy, the recruitment is invariably via the mother. She might be recruited before or during pregnancy. The same is in reality true for children recruited at birth as information on the mother is vital. Good engagement with the mothers is key to the success of the cohort. In cohorts for which the outcomes are during pregnancy (e.g. gestational diabetes and preeclampsia) or at birth (e.g. miscarriage, stillbirth and prematurity) the mother is the key focus of the research. Given the crucial role that pregnancy has on the development of the offspring, gaining data on the mother during pregnancy is important. Where data from before pregnancy can be obtained that is also beneficial. Given much data depends on maternal recall, the earlier the mother can be asked about her lifestyle, wellbeing, environmental and other factors of interest before and during pregnancy the better.

Father

Inclusion of fathers at the earliest possible stage is advisable. Paternal health and lifestyles are important predictors for health of the child, as well as paternal social and cognitive characteristics. Biological samples for genetic and epigenetic studies are needed from both biological parents. Information about paternal life style during pregnancy may even be important for the possibility to perform negative control studies of intrauterine exposures.

Step parents, rainbow families etc.

While the majority (but certainly not all) of children have co-habiting and/or married biological parents around birth, this, in many cases, does not last throughout childhood. Many children have one or, over time, even more mother or father figures, who are not the biological parents. Considerations have to be given as to whether to include step-parents and partners of the parents, as these may constitute important social, emotional and lifestyle environments for the child. For the researcher, this creates a complicated data structure, but since non-traditional families are increasingly common, and the fact that minorities may be offended if questionnaires do not reflect their circumstances, this has to be taken in to consideration from the beginning. One challenge is that cohorts will include children conceived using ART and this may mean there may be some doubt about the identity of the biological parents, for example in the case of egg/sperm donation; some parents feel this is very sensitive and even secret, and so care and understanding is needed by the researchers.

Fetus – Child

Research on the child starts *in utero,* at birth or soon after and then the child is the main focus thereafter. Access to the child is through the parents and apart from specific measurements and tests on the child (e.g. anthropometry, skin prick testing, biosampling etc.), it is the parents who provide all the information on the child through the early years. They also provide consent for the child’s participation (see chapter VII).

Multiple pregnancies and siblings

Recruitment in pregnancy means that decisions have to be taken about follow-up of twins and higher order pregnancies. It cannot be highlighted too much, that the clustered nature of data from multiple births needs to be taken into consideration in building the data infrastructure in the birth cohort. Cohorts need to be large to make follow-up of non-singletons worthwhile, as often they will get excluded from analyses due to there being too few of them to include. Children born as part of such pregnancies are different from singletons in many ways, and so they need to considered carefully in any analysis. Follow-up of twins and triplets places great demands on the parents as questions have to be asked about each child separately, thus doubling or tripling the time required to complete questionnaires and conduct measurements and tests. Birthweights of children born in higher order pregnancies are usually low and their *in utero* experience is very different from singletons.

In addition to considering higher order pregnancies, studies recruiting during pregnancy or at birth might include siblings as index children. Studies recruiting over more than a year will inevitably recruit women who have conceived very soon after the birth of the first child in the study. Consideration needs to be given as to whether to include the second and subsequent children or not, as the participant burden becomes high if the information collection has to be conducted separately on each child as they reach the appropriate age. On the other hand, sibling comparison within the cohort may provide important tests of causality.

Inclusion of twin, triplets etc. and siblings also requires consideration at the analysis stage. The mother is common to the all the children she has in the cohort, so clustering has occurred with the children being ‘nested’ within the mother. Multi-level analysis methods will be needed to assess the influences appropriately.

Many cohorts will, however, collect some information on siblings as they provide information on exposures for the index child. Commonly, the family structure is of interest, though identifying younger siblings has to happen as the index child ages. Older children in a sibship have different exposures to younger ones, and single children have different behaviours from those in large families where the siblings play and interact with each other. Information on numbers of older and young siblings is often collected, but this may be supplemented by particular factors of the siblings such as whether they suffer from allergies or not. Information on the family in which the child is brought up is important for many analyses.

Offspring of index child

As the index children grow older, a decision has to be made as to whether to follow-up their offspring. This is being done successfully in ALSPAC, a LifeCycle cohort, for example.(1) It is rarely possible with small cohorts, as, with attrition, the numbers available for following into the next generation become small. Generally, any such cohort has become quite a selected population by the time the index children become parents, and this needs to be considered in the analyses.

Some index children may start having their own children in their teens while others may not embark on parenthood until much later in life. The recruitment of the next generation takes place over decades, and so is a large undertaking. In the early years, any offspring recruited will be born to younger-than-average parents, and this needs to be remembered in the analyses. Exploring influences on teenage pregnancies, for example, could be seen as valuable, but such pregnancies are relatively rare, and only with a large cohort of index children is there likely to be enough offspring born to teenage parents to allow detailed study of influences on such pregnancies and their outcomes.

However, the wealth of information on the index child and their parents means that many pre-conceptional factors can be assessed in terms of the influence on the index child's offspring. It is worth remembering though, that information is only available from one parent of the next generation, unless by chance both parents are index children in the main cohort. Also, for roughly half the offspring cohort, the preconception information is for the mother and half for the father, thus complicating analyses, which often have to be analysed within parental types, thus reducing statistical power considerably. Many of the index children will have more than one child, and following-up all offspring can become burdensome. Nonetheless, where the offspring can be studied there are great rewards in terms of understanding preconceptional influences and such cohorts are extremely valuable, though currently rare.

Biological or social family

The issue of step-parents was discussed in the section above on fathers. However, the wider social family also influences on the child. When considering siblings, should information on half- and step-siblings be included? What about children in the family who are adopted or fostered? What about siblings (full, half or step) living elsewhere? The broader the inclusion of family members, the more complicated the cohort becomes, and tough decisions have to be made. It is worth deciding at the outset what the limits on the cohort will be, and justifying those decisions, bearing in mind the main objectives of the cohort.

Multigenerational cohorts

Offspring of the index children have been considered above, but some cohorts obtain information on previous generations. Grandparents of the index child are important influences, biologically and socially. Information on them can be gained from the parents of the index child, and this may be sufficient. Contacting all grandparents can be challenging. Again the aims of the cohort must be borne in mind in defining its scope.

The multiple types of participants and co-participants have implications for the data management that need to be considered very carefully in any cohort study (see Chapter VI).

Settings

Hospital-based

Recruiting from hospitals can offer advantages in that the study population can be approached in one or more distinct places. Pregnancy cohorts can often be recruited in hospital provided most women attend pre-natal care in a hospital setting. The feasibility of this will vary from country to country depending on antenatal care practice. Where management of pregnancies is largely in the community, women attending hospital are likely to be those where there is some concern about the pregnancy, and so the cohort would not represent the general population of women who are pregnant. In contrast, where almost all women attend hospitals for scans, for example, the cohort may be more representative.

Birth cohorts may also be recruited at the hospital in connection with birth, but again the question of representativeness applies. If many women deliver at home this would be a poor route to recruitment.

Examples of LifeCycle cohorts with recruitment based in one or two hospitals are Born in Bradford in the UK,(2) EDEN cohort in France,(3) Helsinki Birth Cohort Study in Finland,(4) and the RAINE cohort in Western Australia.(5)

City/Area-based

Many cohorts recruit from a local area, such as a city or a wider region. These provide the opportunity for creating local buy-in to the cohort with support from local authorities, media and other organisations. Local publicity can be conducted and participants will know each other and support for the cohort can snowball. The downside of this, and of hospital recruitment, is that many participants may move out of the area through the life of the cohort and that presents a problem for locally-based follow-up. While postal/online questionnaires can still be conducted with participants who have moved out of the area, more in-depth components of data collection (e.g. anthropometry, blood pressure measurements, DXA scans) require either the participant or the researcher to travel and this can add to the cost and participant burden considerably. Taking samples can be challenging too. Many LifeCycle cohorts are area-based, e.g. ALSPAC in Bristol and the surrounding area, UK,(6) GECKO from Drenthe(7) and Generation R in Rotterdam, both from The Netherlands (8) INMA in Spain, which recruited participants from seven areas,(9) the Northern Finland Birth Cohorts from, as the name indicates, Northern Finland(10, 11), the RHEA cohort from Heraklion in Crete(12) and the SWS from Southampton, UK (13).

National

National cohorts can be larger, but are usually only feasible in countries with strong state and population coverage in registers. Often there is a government drive to set up the cohorts, which provides great support for the cohort. Such cohorts have the advantage that they can track participants as they move round the country, though not necessarily if they move abroad. They generally link into national registries, which provide a wealth of data to complement those obtained directly from participants. Such cohorts tend to collect data mainly from questionnaires as obtaining direct measurements, particularly those involving specialist equipment in local centres, is challenging. Sometimes sub-studies focus on participants living in specific geographical areas who are likely to be able to travel to local research centres. Within LifeCycle, ELFE from France,(14) DNBC from Denmark,(15) and MOBA from Norway,(16) are all national cohorts. In addition, NINFEA from Italy (17) is a national cohort, but its recruitment and data collection are entirely internet-based, so this cohort is rather different from the others.

Trials

Increasingly, trials in pregnancy and young children are turning into cohort studies after the initial endpoint has been reached. Trials do tend to have strict inclusion criteria and that makes the study population less representative. Participants often are asked to adhere to quite demanding protocols and drop-out can be a problem. Also, if the intervention is effective, then the intervention group members are immediately unrepresentative of the population; sometimes only the control group is followed up as a cohort. However, using trials as cohorts does have advantages in that those recruited at the outset tend to be quite committed to research, but selection bias is probably a larger issue than for standard cohorts (though, of course, they also suffer from it). Within LifeCycle only one trial cohort has been included, namely CHOP, a trial of infant feeding formula that recruited participants from 11 sites in five European countries.(18) Other examples from elsewhere include cohorts from trials on women with obesity in London, the UPBEAT trial(19), women with overweight or obesity in Australia, the LIMIT trial(20) and a preconception supplementation trial in Singapore, New Zealand and the UK, the NiPPeR trial. Of course, all trials that become cohorts, if they follow up the intervention group, also continue to assess the longer-term effects of the intervention, as well as addressing other questions using observational methods.

Historical cohorts

Sometimes cohorts are defined based on data obtained from historical medical and other records from hospitals or registers. Follow-up of the participants may be already well into adulthood by the time the cohort is established. If the quality of data on the pregnancies is good then some useful insights into the long-term effects of maternal prenatal influences can be obtained. Such studies avoid the long time span between birth and the development of adult chronic diseases. The disadvantage though is that the exposure data are limited to the information recorded at the time and this is not as extensive as that recorded in prospectively recruited cohorts. The only LifeCycle cohort of this nature is the HBCS in Helsinki, Finland.(21) Such cohorts have however been widely used in life-course research, where the interest is on adult health outcomes; LifeCycle, however, mainly, but not exclusively, focuses on childhood outcomes.

Timing of recruitment

Pre-conception cohorts

Recruiting a cohort before pregnancy is very challenging. No one, not even the women themselves, know when or even if they will become pregnant. Targeting women who are planning a pregnancy helps, but, given that up to 50% of pregnancies are thought to be unplanned, a cohort of women planning pregnancy is not representative of the population. Recruiting fathers before pregnancy adds to the challenges as women may change partners between recruitment and pregnancy. Ideally recruitment would take place in the months or weeks leading up to conception, but this cannot be planned. If women (with or without partners) are recruited from the general population then many women will not become pregnant in a reasonable time and the resulting cohort of children is a fraction of the number of women originally recruited. Such cohorts of women from the general population are very rare, but the SWS(13) is an example within LifeCycle. Generation R Next is recruiting its participants before pregnancy but is focusing on those planning a pregnancy, and is accepting some participants who are already pregnant at the time of recruitment. Cohorts such as ALSPAC, which are following up the subsequent generation have recruited prior to pregnancy for that generation, and have a wealth of data from one parent of the next generation, as noted above.

During pregnancy

Many cohorts recruit during pregnancy. The advantages of this are that the mothers are accessing medical and midwifery services and can easily be contacted. Their partners are often available too. Ideally, participants are recruited as early in the pregnancy as possible, but some women do not realise they are pregnant until the pregnancy is quite advanced, or they do not seek support until it is well established. If recruitment occurs later in the pregnancy, then preterm births can be missed, adding to the selection bias of the cohort. A large majority of the LifeCycle cohorts recruited their participants in pregnancy.

At or shortly after birth

Practices vary across countries and in some places recruitment in pregnancy can be difficult. Sometimes it is felt that mothers are more amenable to joining a research programme after they have delivered a healthy baby for whom they want the best in life. Recruitment following registration of the birth can be done through routine birth data collection, with appropriate permissions. For some cohorts recruited at birth, the recruitment is still via maternity provision. Within LifeCycle, the ELFE study in France recruited at birth,(14) and the CHOP trial recruited during the first eight weeks of life.(18)

1. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013;42(1):111-27.
2. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort Profile: the Born in Bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42(4):978-91.
3. Heude B, Forhan A, Slama R, Douhaud L, Bedel S, Saurel-Cubizolles MJ, et al. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. Int J Epidemiol. 2016;45(2):353-63.
4. Eriksson JG, Forsén T, Tuomilehto J, Osmond C, Barker DJ. Early growth and coronary heart disease in later life: longitudinal study. Bmj. 2001;322(7292):949-53.
5. Straker L, Mountain J, Jacques A, White S, Smith A, Landau L, et al. Cohort Profile: The Western Australian Pregnancy Cohort (Raine) Study-Generation 2. Int J Epidemiol. 2017;46(5):1384-5j.
6. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, et al. Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. Int J Epidemiol. 2013;42(1):97-110.
7. L'Abée C, Sauer PJ, Damen M, Rake JP, Cats H, Stolk RP. Cohort Profile: the GECKO Drenthe study, overweight programming during early childhood. Int J Epidemiol. 2008;37(3):486-9.
8. Jaddoe VW, Mackenbach JP, Moll HA, Steegers EA, Tiemeier H, Verhulst FC, et al. The Generation R Study: Design and cohort profile. European journal of epidemiology. 2006;21(6):475-84.
9. Guxens M, Ballester F, Espada M, Fernández MF, Grimalt JO, Ibarluzea J, et al. Cohort Profile: The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project. International Journal of Epidemiology. 2011;41(4):930-40.
10. Järvelin MR, Sovio U, King V, Lauren L, Xu B, McCarthy MI, et al. Early life factors and blood pressure at age 31 years in the 1966 northern Finland birth cohort. Hypertension. 2004;44(6):838-46.
11. Järvelin MR, Hartikainen-Sorri AL, Rantakallio P. Labour induction policy in hospitals of different levels of specialisation. Br J Obstet Gynaecol. 1993;100(4):310-5.
12. Chatzi L, Leventakou V, Vafeiadi M, Koutra K, Roumeliotaki T, Chalkiadaki G, et al. Cohort Profile: The Mother-Child Cohort in Crete, Greece (Rhea Study). Int J Epidemiol. 2017;46(5):1392-3k.
13. Inskip HM, Godfrey KM, Robinson SM, Law CM, Barker DJ, Cooper C, et al. Cohort profile: The Southampton Women's Survey. Int J Epidemiol. 2006;35(1):42-8.
14. Charles MA, Thierry X, Lanoe JL, Bois C, Dufourg MN, Popa R, et al. Cohort Profile: The French national cohort of children (ELFE): birth to 5 years. Int J Epidemiol. 2020;49(2):368-9j.

15. Olsen J, Melbye M, Olsen SF, Sørensen TIA, Aaby P, Nybo Andersen A-M, et al. The Danish National Birth Cohort - its background, structure and aim. Scandinavian Journal of Public Health. 2001;29(4):300-7.
16. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol. 2016;45(2):382-8.
17. Firestone R, Cheng S, Pearce N, Douwes J, Merletti F, Pizzi C, et al. Internet-Based Birth-Cohort Studies: Is This the Future for Epidemiology? JMIR Res Protoc. 2015;4(2):e71.
18. Koletzko B, Von Kries R, Closa R, Escribano J, Scaglioni S, Giovannini M, et al. Lower protein in infant formula is associated with lower weight up to age 2 y: A randomized clinical trial. American Journal of Clinical Nutrition. 2009;89(6):1836-45.
19. Briley AL, Barr S, Badger S, Bell R, Croker H, Godfrey KM, et al. A complex intervention to improve pregnancy outcome in obese women; the UPBEAT randomised controlled trial. BMC pregnancy and childbirth. 2014;14:74.
20. Dodd JM, Turnbull D, McPhee AJ, Deussen AR, Grivell RM, Yelland LN, et al. Antenatal lifestyle advice for women who are overweight or obese: LIMIT randomised trial. Bmj. 2014;348:g1285.
21. Eriksson JG, Forsen T, Tuomilehto J, Osmond C, Barker DJ. Early growth and coronary heart disease in later life: longitudinal study. BMJ. 2001;322(7292):949-53.

III. Recruitment and retention

Recruitment

Recruitment is the first major challenge for a birth cohort. It should be guided first by scientific choices, but practicalities and budget should not be forgotten.

Enrolment rate

Birth cohorts, along with many population studies, face a decreasing enrolment rate over time. In the UK, enrolment reached 95% in the 1958 Birth cohort, 68% in the Millennium cohort in 2000, and the Life study in 2014 had to stop because of too low a recruitment rate. This trend has been observed in many countries. Several reasons have been identified (1). Due to the proliferation of studies for research, but also for other reasons such as marketing and polls, potential participants are faced with a high number of requests and, on the other hand, may get the feeling that their participation is less and less worthwhile. Second, the decreasing willingness to participate in research may be a reflection of a more global decrease in social engagement. It may also be explained by a growing popular disillusionment with science and participants' belief that research results may not have a significant impact for their own life, or worse may have a negative impact. The Covid-19 crisis may reverse this feeling, but it is too early to observe. Finally, the increasing burden for participants, as studies become more and more complex, may also play a role.

Having said that, a key feature of cohort studies is precisely that they do not have to be representative, as long as the studies have sufficient variation in the exposures of interest. Comparisons within the cohort are generally valid, but it should be remembered that a non-representative cohort study cannot provide information on population frequencies of exposure and disease. Increasingly though, there are concerns about collider bias(2) due to the non-representative nature of cohorts and this needs to be considered. Indeed, collider bias may also affect representative cohorts, but this can be treated by adjusting for the relevant outcome risk factors.(3)

Burden to participant

This issue deserves particular attention. The best scientific study may just not be feasible if it is too complex. The National Children Study in the USA is an example of a sophisticated study that never started (4). Planning a feasibility study ahead of the launch of the main study is highly recommended. Enough time between the pilot and the main studies should be given to draw conclusions on the accessibility of the targeted population, the enrolment procedures, participation rate and cost of this phase of the study. Adding a qualitative component is of interest to understand participants' motivations and barriers. Participants in the feasibility study may be followed up as a pilot cohort to test follow-up procedures. If this is planned, the number of participants in this feasibility study needs to account for attrition, and additional recruitment may need to be considered as time goes on.

Information of participants about the study

The quality of the information given to the potential participants and the way it is delivered are crucial for successful recruitment. Sending an impersonal letter asking people to self-register to a study yields low enrolment rates. The best option is undoubtedly face-to-face contact, with highly motivated staff able to deliver precise information and answer questions. The potential participant should also have enough time to make his/her decision and not feel under too much pressure. Media campaigns during the recruitment phase are helpful. Co-construction of information documents and recruitment procedures with staff and representatives of the target population is recommended.

Local or national setting

The local or national setting of a cohort has a several implications. Recruitment on a local basis (city, district, region) is easier to organize due to a more limited number of actors who often know each other well. It also offers the opportunity to mobilize local funding in addition to national sources. Support from local authorities may help with practicalities and in communication with the population. Another great advantage of the local setting is the possibility to invite participants to a clinical centre and to perform sophisticated investigations, such as MRI. Last, but not least, local events can be organized for the participants to maintain their engagement and create a sense of community and bonds with the study team. Compared with those, the main point in favour of a national study is its visibility, which more easily draws the attention and support of national authorities and media.

Representativeness

Birth cohorts are population studies with multiple objectives. They are increasingly multidisciplinary due to the recognized interdependency of social and health outcomes. In social sciences, representative samples of the targeted population are widely used with the objective of estimating the proportion of the population exposed or affected to advise public policies and project costs. In public health, the ability to generalize research results to the targeted population is an important issue for their usefulness to the society. A representative study will always attract more political support and possibilities for funding.

Representativeness may be difficult to achieve for cohort studies as the participants are informed of the longitudinal nature of the research and have to commit themselves to be contacted again for follow-up sweeps. Some members of the population may be more reluctant to engage so extensively. Weighting procedures and/or margin calibration can correct for departures from representativeness or to transport the results from the study population to a target population. For weighting, a set of characteristics is needed to compare respondents and non-respondents, which may be challenging to collect on all non-respondents. In the French ELFE cohort, authorisation was granted to collect, and then store anonymously, birth registry information for non-respondents.(5)

Recruitment of a representative sample is more complex than recruitment of a convenience sample. A population list for the targeted population is needed, or a list of organisations from where information can be obtained, such as maternity units for enrolment at birth in countries where home birth is rare. Locally-based cohorts may be able to recruit pregnant women or births from antenatal clinics, or, for national cohorts, selected maternity centres may be used. For example, the Born in Bradford cohort (6) recruited women from the one maternity unit in Bradford Royal Infirmary, while the Avon Longitudinal Study of Parents and Children (7) recruited women as early in pregnancy as possible from routine antenatal and maternity health services in three health administration districts, as well as using local media and publicity. Some randomly selected organisations may, however, not be willing or able to cooperate easily or fully. Population lists of pregnant women are not usually available at national level, and attempts to sample pregnant women from housing lists had low response rates when this strategy was tested in the US National Children’s Study. Therefore, some cohorts start with a representative sample enrolled at birth or in infancy, using for example, child allowance benefit lists, if the benefits are granted to everyone. An alternative is seen in the Southampton Women’s Survey,(8) a cohort starting before conception of the child. It recruited young women who were not pregnant from the general practitioner lists in the city. Oversampling based on certain characteristics may be considered. This could be done for practical reasons (oversampling of large maternity units will limit staff need and be less costly) or when higher attrition rates are anticipated in some subgroups. In such cases, the use of sampling weights will be needed in all the analyses.

Recruitment should be enhanced by publicity wherever possible. National cohorts can access the national media more easily than local ones, but the latter can generate local knowledge of the cohort and the word can spread in the community. The Southampton Women’s Survey publicised the study widely in the local media as well has participating in activities to enhance recruitment such as having stalls at local events, and recruitment drives in shopping malls. The study team also took part in the local carnival, producing a float that processed with the carnival through the city, during which leaflets were distributed. Nowadays, use of social media provides another approach to publicity and enhancing recruitment, and this can be useful both for local and national cohorts.

On the other hand, recruitment of a convenience sample is much easier and has the great advantage that the study team is then working with motivated participants. However, it is then difficult to document self-selection and take account of potential self-selection bias in the cohort. Although this self-selection bias has often been ignored for cohorts, it is now more widely discussed in the literature (9, 10).

Retention

After the initial recruitment, retention is the greatest challenge for cohort managers. Comparing attrition between cohorts is complicated due to different definitions. We have collated in Table 1 the percentage of respondents to a follow-up survey when the children were 3 years of age according to the number of births, in some cohorts.

Table 1: Participation rate to a 3 year-follow-up according to the number of initial births or inclusions* in different cohorts

Cohort	**SWS (8)**	**ALSPAC (7)**	**EDEN (11)**	**ELFE (5)**	**Millennium (12)**	**Growing up in Ireland (13)**	**Growing up in Australia (14)**
Country	England	England	France	France	United Kingdom	Ireland	Australia
Recruitment **Year(s)** **N** **Time**	Local 1998-2007 3 158 Preconception	Local 1990-92 14 062 Pregnancy	Regional 2003-6 1 899 Pregnancy	National 2011 18 040 Birth	National 2000 19 552 9 months	National 2006 11 134 <12 months	National 2004 5 107 <12 months
Participation rate at 3-yr follow-up	83%	72%	72%	67 %	81%	88%	90% **

*inclusions for cohorts with enrolment after birth ** between 2 and 3 yrs

The attrition rate can be substantial, especially when enrolment took place during pregnancy or at birth. In all cohorts, retention is at its maximum in the first years of follow-up and declines thereafter, with the retention of more stable and engaged participants. A second noticeable drop in response rate usually occurs with the transition to adolescence and the increasing need to involve the children themselves in the response to questionnaires.

Attrition occurs for two main reasons: 1) related to the participant: wanting to leave the cohort or non-response to study requests; 2) loss of valid contact details for participants.

Participants who live in deprived social situations or experience stressful life events encounter more barriers in taking part in research (e.g. language, remoteness from study centre, poor internet connection) and are more likely to withdraw from the study or to be non-respondents.

Participants' motivations

For the first situation, qualitative studies examining participants' motivations for staying in longitudinal studies are useful in analysing the reasons for attrition. In a recently published Australian study on a birth cohort now reaching 27 years of follow-up,(15) it appeared that regular participants mainly emphasised the individual and collective benefit of their participation in the study. In contrast, infrequent participants indicated that the obstacles to their participation (time, travel, etc.) were too great in relation to the perceived benefit. For these infrequent participants, their social or personal circumstances can increase the difficulties they face in participating. For them, any action likely to remove certain barriers is likely to increase their participation. This study also highlights the influencer role of other family members in maintaining participation, such as the second parent or the grandparents.

Over time, it is important to maintain the feeling about the collective benefit of participating in the cohort. It is generally well stressed at enrolment in the cohort but demonstration of this collective benefit may take time and participants may get disappointed. Frequent updated information for the participants through newsletters, websites and media even for small achievements is important, as it provides the opportunity to remind participants about the general goals of the study and to help them understand how the research is proceeding. Release of study results or events in the media has potentially the strongest impact but it is sometimes difficult to control the content and to meet certain requests from journalists, such as when they want to contact participants. Social networks offer new opportunities for communicating with young parents and adolescents but need to be adapted to the target population. One has to keep in mind that the popularity of any given social network changes fast.

Objectively, participating in a cohort has few individual benefits. Getting updated information on certain topics, free clinical examination or receiving results of some biological tests are among those that can be quoted. However, for the latter, biological tests performed in a research setting have, most of the time, not yet established personal benefit and for this reason are usually not sent back to participants (unless they ask for it). Emphasizing personal benefits in order to increase participation may also induce some self-selection of participants and therefore is not recommended.

Sending back some personalized results is an option. In the 5-year survey of the ELFE cohort (5), they tested whether, in the letter announcing the survey, inclusion of personalized results from a previous wave and the average of all responses had an effect on participation. A newsletter with or without personalisation (on time spent in front of screens) was therefore randomly sent to around 1000 families who had participated irregularly in previous surveys. Extreme situations were excluded before randomisation. There was no significant effect on survey participation and the observed trend was even marginally against personalisation.

Relationship with cohort management team

There is a long-term relationship between participants in a cohort and the study management team. As such, it requires trust. Trust is established through the honesty of the information communicated to the families about the general goals of the study, about the announcement of each data collection wave and about the use of the data collected. Renewing consent for any new data collections, and offering opt-out options for sensitive parts of questionnaires or sample collections are elements that reinforce the participants' feelings that their freedom is taken into account and respected within the cohort. The guarantees given about data protection are also very important. They can be given through information and documents, but also in the way the data are collected (for example, using identification numbers rather than personal identifying information in questionnaires). Research institutions in Europe have ethics committees to review research protocols and consent forms. Studies have to comply with the General Data Protection Regulation (GDPR) that has been implemented in the European Union since 2018. Using legal documents and formulations to comply with these official regulations is mandatory. However, very few participants take the time to read legal documents in detail and therefore it is important that the main points are summarised either by an interviewer or in a more friendly information document.

It also easier to build trust if the participants know the management team. All ways to humanize the relationships should be considered. In local studies, the families can meet the staff in person and they will appreciate seeing the same staff over years or meeting new staff who are introduced to them. Of course, enthusiasm and motivation of the staff is a key point for this relationship. In national studies, this is rarely possible but putting faces to the names of staff remains important. Presenting the team in information documents and newsletters and explaining their role in the study can help. Mail, email and phone hotlines should be provided to participants to enable rapid and personalized responses.

Involving participants themselves in the study procedures through a group of representatives who can be consulted for advice or tests is useful, and is also a means for building trust. However, it has to be kept in mind that parents and young people volunteering for such groups are self-selected and do not represent all the opinions. This can be improved if care is taken when the group is constituted to represent diverse social and familial situations.
Building a relationship with the children and adolescents, who progressively will become the main respondent for the cohort, is of major importance. Information documents designed

specifically according to their age, and use of their favourite social media, birthday cards and small gifts are among the possible means to achieve this.

Withdrawals

Withdrawals will of course occur. Participants in research have the fundamental right to withdraw at any time and with no need for justification. For research on children, the right to withdraw has to be respected even if only one of the parents asks to withdraw. Once the child has reached an age at which it can be involved in discussion about participation (around age 6 years) their assent should be recorded. If the child declines even though the parent has provided consent, then the child’s wishes need to be respected. To keep records, it is good practice to ask participants to inform or confirm their request to withdraw by writing/emailing/calling the cohort management team. The contact with the study management team offers the opportunity to clear up any misunderstanding or to propose a lighter survey protocol. It is not unusual for parents who do not confirm in writing, to agree to participate in a subsequent wave. Whenever appropriate, it is important to remind participants about the possibility of skipping a particular data collection wave without withdrawing completely from the cohort.

Non response

Cohort management teams have developed a number of strategies to limit non-response. A review of the studies that evaluated the methods used to reduce attrition in population cohorts (16) identified three types of strategies: incentives for participating in the survey (in the form of gifts or a sum of money); reminders to non-respondents; and other methods such as reducing the length of questionnaires. In randomised studies, the use of inducements increased the response rate from 2% to 13% depending on the study, with, when it comes to money, a greater effect with greater incentives. There was also some evidence that proposing incentives upfront of a survey may reduce the cost of the interview, by limiting the number of unsuccessful attempts to reach the participant. Targeting incentives to participants with low response rates to previous surveys may offer the best cost-benefit ratio, but raises some ethical issues. Payment may be particularly important for engaging the children as they become older and as their parents are less involved in encouraging them to take part. For many years, payment to participants was discouraged but, as participation rates have reduced, such methods are being used more widely, though the payment is often in the form of vouchers.

Using different survey recall strategies can also be very effective. Most studies use it in a hierarchical manner starting with the less costly and labour intensive methods, such as reminder e-mail or SMS and ending with more intensive methods such as proposing other collection methods. Participation increases with greater numbers of reminders, and the optimal number is a compromise between cost and yield. In longitudinal studies, an excessive number of reminders may compromise participation in the next survey wave. Reminder methods that facilitate participation have to be preferred. For example, it has been shown that resending a postal questionnaire is more effective than sending a reminder letter. In telephone surveys, the number of calls determines the response rate. Usually, most of the participants can be reached with a limited number of calls, but as many as 50 calls may be needed to reach each of the last 10%.

Loss to follow-up

Except in countries with national population registries such as in Scandinavia, participant tracing is a regular activity of cohort management teams and appropriate budget should be allocated to the task.

Considerable attention should be given at enrolment in the cohort to obtain and validate contacts using all possible means (postal address, phones and e-mails of the two parents). Many families of young children move often, and changes of mobile phone numbers or inactive e-mail addresses are more frequent than one might have thought. Contact details of other family members could also be collected in case the participating family cannot be reached anymore during follow-up, but these family members should be informed about this. Grandparents are more likely to stay at the same address than the family, so can be useful contacts.

Annual newsletters or birthday cards are good ways of reminding families to update their contact details if needed. Local cohorts can also use field tracking at the participant’s last address. Post and phone providers, public listings and forwarding address services can also be used to trace participant, but nowadays increasingly citizens ask for their contacts not to be disclosed. Social media searches can also be used (17) but a balance has to be found between respecting people’s privacy and the reduction of loss to follow-up, which may well differ according to the cultural context. It may be possible, with ethical approval and consent, to use school, GP or health insurance registries to update contact details, but again there are variations across countries in the acceptability of this practice.

Dealing with attrition

Although every measure should be taken to limit it, attrition is inevitable in longitudinal studies.

In a country such as Denmark where there are many registers, a number of associations were studied both in the general population and in the population of respondents to the Danish Birth Cohort at the 7-year follow-up survey (18). This study is informative for many birth cohorts in

European countries because the main factors predicting attrition, especially those related to social status, are often similar across cohorts. For associations between several pregnancy-related outcomes and heath conditions at 7 years, it showed modest over- or under-estimates of the odds-ratios depending on the health events studied, generally from 3 to 10%, but up to 30% in one analysis.

However, the articles cited above are limited to the study of health events and the authors could only study the effect of factors for which external information was available. The conclusions might be less reassuring for outcomes that are more strongly differentiated socially. Specific statistical methods for analysing cohort data to adjust for attrition is increasingly being recommend. These methods involve calculating specific inverse weights for the selection of the analysis sample or imputing missing data, including the event of interest.(19, 20) In any case, being able to quantify departure from representativeness of cohort data before or after weighting using external sources from the targeted population will always be useful.

1. Galea S, Tracy M. Participation rates in epidemiologic studies. Ann Epidemiol. 2007;17(9):643-53.
2. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. International Journal of Epidemiology. 2018;47(1):226-35.
3. Richiardi L, Pearce N, Pagano E, Di Cuonzo D, Zugna D, Pizzi C. Baseline selection on a collider: a ubiquitous mechanism occurring in both representative and selected cohort studies. J Epidemiol Community Health. 2019;73(5):475-80.
4. Kaiser J. The Children's Study: Unmet Promises. Science. 2013;139:133-6.
5. Charles MA, Thierry X, Lanoe JL, Bois C, Dufourg MN, Popa R, et al. Cohort Profile: The French national cohort of children (ELFE): birth to 5 years. Int J Epidemiol. 2020;49(2):368-9j.
6. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort Profile: the Born in Bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42(4):978-91.
7. Boyd A, Golding J, Macleod J, Lawlor DA, Fraser A, Henderson J, et al. Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. Int J Epidemiol. 2013;42(1):111-27.
8. Inskip HM, Godfrey KM, Robinson SM, Law CM, Barker DJ, Cooper C, et al. Cohort profile: The Southampton Women's Survey. Int J Epidemiol. 2006;35(1):42-8.
9. Hernan MA, Hernandez-Diaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004;15(5):615-25.
10. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. Acta obstetricia et gynecologica Scandinavica. 2018;97(4):407-16.
11. Heude B, Forhan A, Slama R, Douhaud L, Bedel S, Saurel-Cubizolles MJ, et al. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. Int J Epidemiol. 2016;45(2):353-63.

12. Connelly R, Platt L. Cohort profile: UK Millennium Cohort Study (MCS). Int J Epidemiol. 2014;43(6):1719-25.

13. Williams J, Murray A, McCrory C, McNally S. Growing Up in Ireland: Development from birth to three years. (Infant Cohort Research Report No. 5). Dublin; 2013.

14. The Longitudinal Study of Australian Children. Annual statistical report 2010. Australian Institute of Family Studies; 2010.

15. Costello L, Dare J, Dontje M, Straker L. Informing retention in longitudinal cohort studies through a social marketing lens: Raine Study Generation 2 participants' perspectives on benefits and barriers to participation. BMC Med Res Methodol. 2020;20(1):202.

16. Booker C, Harding S, Benzeval M. A systematic review of the effect of retention methods in population-based cohort studies. BMC Public Health. 2011;11:249.

17. Schneider S, Thomas G, Burke-Garcia A. Facebook as a tool for respondent tracing. Survey Practice. 2015;8(2):1-8.

18. Greene N, Greenland S, Olsen J, Nohr EA. Estimating bias from loss to follow-up in the Danish National Birth Cohort. Epidemiology. 2011;22(6):815-22.

19. Cumming J, Goldstein H. Handling attrition and non-response in longitudinal data with an application to a study of Australian youth. Longitudinal and Life Course Studies. 2016;7(1).

20. Seaman SR, White IR, Copas AJ, Li L. Combining multiple imputation and inverse-probability weighting. Biometrics. 2012;68(1):129-37.

IV. Questionnaires and data collection methods

Introduction

Cohort and longitudinal data present challenges due to the data collection taking place over a long time, in some cases entire lifetimes. Some cohorts started many years ago and the variety of methods of data collection available has expanded considerably. Most analyses of cohorts of the type included in LifeCycle would invariably draw on data collected in the early waves, so an understanding of the methods of data collection used over time is important for contextualising the data and assessing their quality. This chapter describes methods used for collecting data in cohorts, with a focus on the areas of particular relevance to the types of cohorts that engage with LifeCycle.

Questionnaires

Methods of administration

Self-administered questionnaires: Historically, questionnaire data were all collected on paper, and some still are. Questionnaires are either sent to participants directly, or they could be completed by a participant while in a clinic, for example while attending a routine antenatal appointment. In more recent years, online questionnaires have become much more widely used, with an acceleration in their use as a result of the COVID-19 pandemic. This is called computer-assisted web interviewing (CAWI). The participant receives a link to the questionnaire and completes it online in their own time. Questions can be skipped depending on previous answers and this can help the participant and reduce time. Widely used software include Qualtrics and REDCap.

Interviewer-administered questionnaires can be collected on paper, and that was how they were traditionally done. They can either be done in person at the participant's home, or in a research or clinic setting, or conducted over the telephone. Increasingly, though, electronic methods are often preferred. Two main types of computer-assisted interviews are used: Computer-assisted personal interviewing (CAPI) in which the interviewer meets the participant face-to-face and records the answers on a laptop, tablet or other device that holds the questionnaire. Computer-assisted telephone interviewing (CATI), as the name suggests, is conducted over the telephone with the answers being recorded by the interviewer into a computer system.

Interviewer-administered data collection methods can allow for more complex data collection. For example, detailed questions on areas such as employment history, educational attainment and household composition may require careful questioning and interviewers need skill to identify the correct category for each participant's answers. Birth cohorts, almost without exception, measure the size of the children as they develop, and reports of length/height and weight from the mother are not sufficiently precise for good quality analysis. An interviewer who meets the participant face-to-face can take the appropriate equipment to the home for extensive anthropometry measures, or the measurements can be made at a clinic. A research interviewer may also need to conduct other procedures, particularly for health-focused cohorts, such as: measuring blood pressure, collecting blood or other samples, skin prick testing, spirometry, cognitive function measures etc. A clinic/research centre visit is required for measurements that require large equipment for scanning such as dual x-ray absorptiometry (DXA), magnetic resonance imaging (MRI), ultrasound etc. or other procedures involving monitoring with complex equipment.

Advantages and disadvantages of different methods of administration: Postal questionnaires have advantages in that they require less staff time. They also allow the participant to complete the questionnaire in their own time. However, there is no method for checking the questionnaire as it is completed, sections can be omitted, and rebellious participants can give ridiculous answers. Errors can occur, and checking back with the participant can be time consuming when inconsistencies are noticed. They also have to return the questionnaire via the post, and, increasingly, with most communication nowadays being electronic, a journey to post the questionnaire back can deter some participants, particularly younger ones. Postal questionnaires have the disadvantage of excluding those who have literacy problems and are difficult for those who suffer from visual impairments.

Online questionnaires get round some of the disadvantages of postal versions as the questionnaire can be designed so that answers outside the possible range cannot be accepted, and logical checks between answers can be included. This can reduce error in comparison with a postal questionnaire. Such methods have improved over the years, making it easier for participants to answer the questions in an easy and logical flow. As with postal questionnaires, those online cannot be accessed by participants with literacy problems. They also exclude participation from those who do not have access to the internet, though, in high-income countries, this now applies to few people and mainly to those who are of older age. However, those with hearing impairments may find postal/online questionnaires easier than being interviewed by a researcher.

Self-administered questions may be more acceptable to participants when sensitive questions are being asked. Having an interviewer present may inhibit those who are embarrassed about particular issues in their lives.

Interviewer-administered questionnaires are more costly than postal or online versions, but enable an interaction between the interviewer and participant. As noted above these allow more complex issues to be addressed, and measurements and other procedures can be conducted at the same time, with measurements all being recorded in the questionnaire. Asking participants to visit research centres can affect response rates as some will not want to travel and some are unable to do so. If large distances are involved, this can place a great burden on participants. Nonetheless, for collection of some types of data, there is no alternative. In cohort studies, there is often a need for participants to attend many visits over time, and attrition can be a problem.

Design of questionnaires and data collection

Designing questionnaires is both a science and an art. The designer must focus on how the questionnaire will be perceived by the participant. Thus involving participants is key. Many cohorts have participant panels who advise on each wave of the cohort. Participant involvement can make a major contribution to good data collection. However, a word of caution is that such panels are formed of those who are keen and engaged in the cohort and they tend not to include those from more disadvantaged backgrounds. Bearing in mind how the less engaged and those who are more disadvantaged would perceive the data collection methods is also vital and the participant panel’s views need to be interpreted carefully. Having said that, participant views should be sought on all aspects of data collection that are planned, including the choice and acceptability of topics.

A useful guide was produced some years ago by Stone(1), and although it relates to paper questionnaires, it includes a checklist for designing a questionnaire that is broadly generalisable, as follows:

1. Decide what data you need
2. Select items for inclusion
3. Design individual questions
4. Compose wording
5. Design layout
6. Think about coding
7. Prepare first draft and pretest
8. Pilot and evaluate
9. Perform survey

These are useful points to consider in any questionnaire.

Choice of specific questions/methods

Using established validated questions wherever possible is usually recommended. However, there may be a wide choice. For example, there are many questionnaires that assess mental health and wellbeing; a systematic review in 2015 identified 60 different scales for assessing wellbeing.(2) Choosing the most appropriate for the population under study needs care, and discussion with topic experts is vital. Notably, general overviews often do not include scales for specific conditions or situations that might be widely used; for example, the Edinburgh Postnatal Depression Scale is widely used in the LifeCycle cohorts for measuring postnatal depression and would be a better choice for that particular disorder than a more general depression scale.

For cohort studies, the choice becomes even more complicated. There is a need for age- and stage-appropriate measures. Measures for children are usually validated for certain age groups, and, for particular issues such as wellbeing, are rarely valid throughout the entire age-range into adulthood. This presents a challenge when wanting to examine trajectories through the life-course as the measurement scales may have to change and so may not result in an outcome measure that is consistent across ages. For instance, some result in binary measures, others categorical, and some are continuous but with varying ranges. Careful thought needs to be given to forward and backward compatibility across age groups as the cohort progresses.

With a focus now on analyses combining data across cohorts, it is well worth considering the measures used most commonly by other cohorts. Using the same measures is a great advantage when it comes to harmonising data prior to meta-analysis. However, this is challenging and it has been the subject of much discussion in recent years. Should there be a strong recommendation to use a specific scale for a particular topic? That sounds sensible, but it is not straightforward. Cohorts are usually set up to address particular hypotheses and have foci of interest; there is a contrast between cohorts set up with a health remit and those under a social science umbrella, and their priorities for data collection differ. The questionnaires preferred by topic experts are often lengthy, as experts want comprehensive data. But, if that topic is only of peripheral interest to a particular cohort, it may be too long to include and so gets omitted. In such situations, it might be valuable for the cohort to have a shorter questionnaire, albeit not as comprehensive, but one which gives a general overview that is useful for some analyses, and may be valuable as a measure of a confounding variable.

New validated questionnaires present a challenge to researchers working on cohorts. Should the new measure be used even if it differs from those used in previous waves of the cohort? There is a tension between using out-moded methods to allow for compatibility across waves and updating the methods as appropriate, but then struggling to analyse across the life-course. The challenge of long-running cohorts is that their data are always out-of-date. A birth cohort in which the offspring are now aged 30 years contains data on early childhood that do not represent experiences of children now. However, the strength of cohort data is in allowing an understanding of the exposures over the life-course that led to outcomes at particular ages or their trajectories over time. No cross-sectional study can give such insights. Bearing this in mind is important when considering which questionnaire to use in any particular data collection wave. This is not just a problem for questionnaire design, as it can apply to all types of data collection. For example, imaging methods with newer equipment will give better more accurate measurements than could be obtained in previous waves, and refinements to laboratory methods can result in improved measurement accuracy or a new measurement altogether, which can be hard to relate to previous results.

It is important not to over-emphasise the reliance on validated questions and questionnaires. The suitability of questions in the particular context of the cohort needs to be assessed. It might be better to derive a new questionnaire than to rely on a validated one that does not address the issues well in the cohort under consideration.

Choosing appropriate options for answers

Making it as easy as possible for participants to answer the questions is vital. If they are asked to choose between vague categories, they may not answer the question at all. So if asked "How often do you take exercise that makes your heart beat fast and makes you breathless?", answers such as: never; occasionally; sometimes; or often, are too vague and it is better to give options such as: never; less than once a week, one to six times a week; once a day; more than once a day. It is important that categories do not overlap. For example, asking people to indicate their age where the options include 20-25 and 25-29 years, make it difficult for people aged 25 to know what to answer.

Sometimes categories are hard to define, and numerical or visual analogue scales can be helpful. Questions such as "How satisfied do you feel with your life at the moment?" could be answered on a scale of 1 to 10 where the participants are told that 10, is the best possible life for them and 0, is the worst possible life. Alternatively, they might be asked to mark on a line drawn on the questionnaire where the left-hand side of the line is the worst possible life and the right hand side is the best possible.

Complications can occur when differing units of measurement are widely used. In the UK, for example, some people think of their height in metric while others still use the old imperial measure of feet and inches. Participants who are being ask to report their height, need to be provided with the option of using either metric or imperial.

Translation

Translation of questionnaires into different languages is far from straightforward. In a multi-cultural context, there is a need to ensure that all cohort members can understand the questionnaire. In some countries, there is more than one national language, so there is not even one starting point. Questionnaires are often designed in English, due to the widespread use of that language, but most cohorts will need to have questionnaires in various languages. Many established questionnaires have versions in a variety of languages that have been validated and these are helpful. But what if there is no validated version for the languages needed for your cohort? Validation of a questionnaire is no small task and there may be no 'gold standard' against which to validate it anyway. Internal consistency and validity can though be checked. However, few cohorts have the resources to validate all the questionnaires they may wish to use. The best that may be possible is translation and back-translation to ensure the meaning is retained. Testing for comprehensibility in the target population is vital to ensure that participants do not misunderstand the questions. This is a major challenge for those working in populations that speak minority languages, but it affects most cohorts to a greater or lesser extent and requires great care and effort. The style of the language into which the questionnaire has been translated must also be appropriate for the age of the participants who will be completing it.

Validated questionnaires may not exist for a particular cohort, or those that exist may not work well when translated. Sometimes the language does not seem appropriate for the level of education of the cohort, or, for example, the use of American English may not be appropriate in the UK or Ireland. Nuances of validated questionnaires may be lost when translated. There is no point in using a validated questionnaire just because it works in other cohorts, without assessing its use in your own cohort and its various languages. It may be that a new questionnaire is required.

Prompts for participants

Interviewers may use prompt or flash cards to provide extra information to aid participants to understand and answer questions. For example, in dietary questionnaires, when asking about consumption of foods from within a food group, cards containing example foods contained within the food group can be useful. Pictures of portion sizes, or examples of particular foods can also act as prompts to participants. Whenever the answer requires a choice of a number of categories then a card containing those categories can be helpful, rather than have the interviewer read out the list and expect the participant to remember and then choose the correct one. Such cards can usefully be used for questions about topics such as employment, education, perceived general health, wellbeing, diet, housing etc.

For self-complete questionnaires on paper or online, the categories and their explanations are built into the questionnaire. Pictures can also be useful as they make the questionnaire more engaging to participants. However, questions with complicated instructions are not easily answered by participants, and, where such questions are necessary, thought needs to be given as to whether the questionnaire should be interviewer-administered.

Context

The context in which the study is being conducted will need to be considered in the development of the data collection methods. Cultural differences mean that certain questions may be acceptable for one group, but not for another. Sensitive questions, for example, on sexual activity or childhood abuse, can be particularly difficult, but harder in some cultures than in others. A particular challenge is collecting data on diet due to the variations between and within countries. A validated dietary questionnaire used in one location with a particular cultural/ethic group is unlikely to be suitable for another, due to the differences between their dietary practices. Food diaries can be helpful in this situation but they present an enormous burden on the researchers in terms of coding, and are not suitable for participants with poor literacy skills.

Flow of the questionnaire

The order of questions in a questionnaire needs careful thought. It is unwise to start with sensitive topics and the trust of the participants needs to be developed, particularly for interviewer-administered questionnaires. The more sensitive issues need to be absorbed in the central part of the questionnaire; ending on them is not recommended. There is a pattern to the questionnaire that is worth considering, so that the sensitivity gradually increases and then falls away again. In cohort studies, developing the trust of the participant is important and familiarity can help. Thus, it can be helpful to keep the order broadly the same from wave to wave. Similar opening questions in each wave can make the participant feel comfortable and on familiar ground, and that can help their engagement. However, it is important that boredom does not set in, and so variation can be helpful.

For self-completed questionnaires, particularly those online, it may be wise to have the most important questions near the beginning, as participants may give up part way through, though this is only appropriate if the key questions are not too sensitive. A fine balance needs to be struck.

Length of questionnaires

Participant burden is a major challenge in cohort studies. Lengthy questionnaires can be off-putting. The stamina of the participants needs to be considered. Interviewing parents with small children can be difficult, and shorter questionnaires may be required than for older participants. Those who are frail or have mental health issues may only be able to tolerate short bursts of engagement. If data collection is face-to-face then everything needs to be collected at once as the cost of travel and the time taken to do so for the interviewer or the participant means that many separate visits are unworkable. For online questionnaires though, some participants may prefer one long one that they work through, while for others short questionnaires sent separately over a period of time may be more manageable. Understanding the needs and wishes of the cohort participants is vital and their views need to be sought. With all questionnaires, the participant needs to be informed about the approximate length of time it will take to complete. With paper questionnaires, it is fairly self-evident to the participant approximately how long the questionnaire is, but this is not so clear online. Some systems show the percentage of the way through the questionnaire, but people vary in the time they take. Clear guidance at the front is necessary. The ability to save partially completed questionnaires is good, as it is better to have some data than none, but as noted above, consideration of the questions to put first needs care.

Check questions

With questionnaires, we are relying on the participant to provide accurate results. Some questions can be challenging for people to answer. For example, being asked how often one eats particular foods is not easy for anyone to answer precisely. After asking about specific foods, it can be helpful to ask about broad categories, such a fruits, vegetables, fish, meat, snacks etc. and then query answers that do not seem broadly to match those for the individual foods. This is probably only possible if an interviewer is present to conduct the check. While checks can be incorporated in online questionnaires, participants get weary if their answers are challenged too often.

Testing and piloting

An important step in developing a questionnaire is getting others to complete it and comment on any challenges or difficulties they have with it. This can start with testing within the research team, but after that it is vital to test it among people who are similar to the cohort members. This may be done with a cohort participant group or with individuals who are similar to the cohort members, in terms of age, gender, ethnicity etc. Testing and piloting questionnaires at the outset and then making improvements can add considerably to the quality of the data collected.

Other types of data collection

Qualitative data collection

Increasingly, studies are including elements of qualitative data collection. This may be as simple as asking open questions to which the participant gives a text response, and these then need coding and analysing. Where resources allow, detailed text responses can provide rich insights into the participants' health, wellbeing and opinions. Use of such questions in a large cohort can, however, lead to a large amount of work, and, in such cases, it may be better to use quantitative methods by providing options for the answers. For example, a question such as "How is your health today?" could lead to lengthy essays as answers, but a simple categorisation of options such as: Excellent; good; fair; poor; or very poor, would save a large amount of time spent coding later on.

Much more extensive qualitative work may be conducted in which samples of the cohort have in-depth interviews or are asked to be part of focus groups. Specific issues relating to cohorts are how to select participants for qualitative data collection. Should different participants be included at each wave to broaden the representation, or should a panel of participants be consulted at each wave? The danger with the latter is that the group may shrink over time and need supplementing, but that then alters the balance. The nature of the issues to be explored will likely dictate the approach to be taken, but it needs consideration, and qualitative research expertise is required to conduct such studies properly.

Newer technologies

Questionnaires are an old approach to data collection, though still very useful. Richer, potentially more accurate data can be obtained with the use of newer methods. A whole approach to data collection known as ecological momentary assessment (EMA) collects data in real time and can be done in different ways.

Increasingly smartphone applications (apps) are being used for frequent data collection. This can be done passively, in which the app records constantly in the background, for example activity, sleep, movement, social networks, but do require the participant to keep the phone with them at all times. A recent example of this is apps for contact tracing for COVID-19 in which the Bluetooth technology in the phone is used to assess distance between people and the time they are together. If one person tests positive all those who were in close proximity for at least a specified time can be contacted and asked to isolate. Active use of apps requires the participant to record data in the app. This, for example, could include the food eaten during the day, or recording moods and feelings. However, active apps do require significant participant engagement and motivation, and long-term participation, as in cohort studies, can be burdensome. Another coronavirus example, is a successful COVID Symptom Study app in the UK in which more than 4 million participants recorded each day whether they had taken a coronavirus test, had received a vaccine, and whether they felt physically normal or not – if not there were further questions about symptoms. This was used to understand transmission of the virus and, particularly during times when testing capacity was limited, allowed a greater understanding of infection rates defined on the basis of symptoms. The clever part of that app was that, after a few basic questions for registration, on most days only two or three questions needed to be answered, so it was not too burdensome.

Wearable technologies are another approach to data collection. Probably the most widely used in cohort studies are for accelerometry to measure physical activity over a period of days. The participant wears an accelerometer that records their movement in real time. Most of these are similar to a watch but others can be worn round the waist or on the leg. At the end of the measurement period, the accelerometer is removed and the data downloaded by the study team. Other such methods can be used for tracking sleep, blood pressure and heart rate, for example. Many smart watches and fitness trackers of the type widely worn by members of the public conduct this type of monitoring, though these have rarely been validated as being sufficiently accurate for research purposes. Also, silent monitoring is to be preferred; smart watches and fitness trackers show the participant their results and can lead to modification of behaviour, whereas those suitable for research do not reveal any information to the participant in an attempt to measure, as closely as possible, their usual behaviour. Sometimes the first few days of monitoring need to be removed as participants try to improve their behaviours as they know they are being monitored; this usually settles down after a short period.

Social media data

There is a constant drive to improve data collection by finding new ways of obtaining information about participants. One example is social media monitoring. With appropriate consent in place this enables insight into the political views of participants, their mood as assessed by the way they engage, the time at which they do it and whether they are passive or active, as well as insights into their pastimes and interests and their social networks. Analysis of such data is not simple, and the data are extensive, so data management is challenging, but they are a rich source of information on groups that engage with social media.

Data from routine sources

Increasingly cohorts are being enriched with data from routine sources. Such information includes data from health services (e.g. hospital admissions, medication prescription, disease registers), educational attainment, crime and justice data, employment information, environmental monitoring, deprivation measures and data in other public repositories. Obtaining consent for this is vital and the bureaucracy involved in obtaining permissions to access the data can be onerous, so considerable commitment is required to obtain the information. For cohort studies, however, these can be very valuable. Retrospective data can fill in gaps in information that was not collected in earlier waves, while snapshots can provide outcome data for particular analyses. Ongoing flagging of participants in these routine sources can allow for regular updates at times when the most recent data are required for particular analyses.

Data linkage

Data linkage is vital within cohorts; if the waves of data cannot be linked then the whole cohort is undermined. Participants are given ID numbers, but those alone are rarely sufficient. If they are mis-typed the linkage cannot occur. Other identifiers are needed, such as names, or dates of birth, to confirm the linkage, but once confirmed should be deleted from the data.

Linkage to external data sources might be crucial for longitudinal studies, but does require identifiable information and a legal framework, as well as a political will to allow it. This is an evolving issue, and there is a need to find a sensible balance between issues of data protection and privacy and the research and public health needs. When linkage is allowed then the issues are complex to ensure correct matches; managing the identifiers required and ensuring consistency across waves is vital to ensure good data that can be analysed across the life-course.

CLOSER Resources

CLOSER is an organisation that aims to maximise the use, value and impact of the UK's longitudinal studies, both at home and abroad. While based in the UK, much of the work they do is of a general nature and the resource is useful for those involved in cohorts worldwide. The organisation has held various workshops and produced a range of reports about data collection and questionnaires for use by those involved in cohort design and management. The full set of reports is available at https://www.closer.ac.uk/resources/

Particular reports of relevance are:

Mixing modes and measurement methods in longitudinal studies

Overview of bio measures in longitudinal and life course research

New technology and novel methods for capturing health-related data in longitudinal and cohort studies

The use of new technologies to measure socio-economic and environmental concepts in longitudinal studies

The future of data collection in longitudinal population studies: during and after COVID-19

The following three reports focus on dietary, physical activity and cognitive measures. They describe measures used in specific cohorts and longitudinal studies within the CLOSER consortium but provide useful overviews of the issues involved in collecting data on these topics.

A guide to the dietary data in eight CLOSER studies

Physical activity across age and study: a guide to data in six CLOSER studies

A guide to the cognitive measures in five British birth cohort studies

1. Stone DH. Design a questionnaire. British Medical Journal. 1993;307(6914):1264-6.
2. Lindert J, Bain PA, Kubzansky LD, Stein C. Well-being measurement and the WHO health policy Health 2010: systematic review of measurement scales. Eur J Public Health. 2015;25(4):731-40.

V. Collection and storage of biological samples

General purpose

Biological samples are undertaken in most birth cohorts with health outcomes. They have two main purposes:

1) Contribution to the understanding of the pathophysiology of diseases: Human biology is composed of fine-tuned systems that respond to each other. The precise description of their dysregulation often needs tests or access to tissue that are not doable or affordable in a birth cohort. However, the omics era is bringing new potential. Indeed, some combination of genetics, epigenetics, transcriptomics, metabolomics in easily accessible fluids and tissue may soon allow a more precise approach of biological dysregulations and yield new potential for cohorts in assessing pathophysiological processes.

In any case, when these dysregulations can be assessed by biomarkers easily measured in accessible body fluids or tissues, the longitudinal nature of birth cohorts offers a unique opportunity to identify early biological dysregulations and the succession of anomalies that ultimately lead to overt disease.

2) Measures of biomarkers of population health: There are two distinct types of biomarkers
(i) Biomarkers of biological dysregulations: In addition to the understanding of pathophysiology as described above, their measurement is useful to assess their predictive value for diseases alone or in combination with clinical signs. In large birth cohorts, the predictive value of biomarkers can be compared across subgroups in real life situations. This knowledge forms the basis for targeted prevention.

(ii) Biomarkers of exposures: A wide range of external exposures are either directly incorporated in the human body or have detectable consequences on human biology. Environmental pollutants, tobacco smoke, food intake, and radiation are examples of exposures that can be assessed with biomarkers.(1-3) Their measurement provides objective assessment of exposures that complement data obtained by questionnaires. In cohort studies, their assessment aims to decipher their role in pathological processes. In pregnancy and birth cohorts, whether these exposures can alter early development is a central question.

However, collection of biological samples adds cost and complexity to the design of a birth cohort. The cost needs to be envisaged not only for the collection of samples but also, and largely, for the long-term storage of the samples if they need refrigeration or freezing, which is often the case.

In the remainder of this chapter, we will briefly cover some of the main points to consider when planning biological samples collection in a birth cohort.

Type of samples to be collected, when and from whom

The type of samples that are usually collected in birth cohorts are:
- invasive samples: venous blood, capillary blood spot
- non invasive: placenta, cord blood and tissue, maternal milk, urine, meconium and stools, saliva, hair, nail, teeth, exhaled breath condensate, induced sputum

The samples are collected to get information on different developmental periods:
- For the prenatal period, most birth cohorts collect samples from the mother during pregnancy and/or at birth, and the cord at birth. Sampling of the placenta after birth is also commonly conducted to get insights into the exchanges between the mother and the fetus during pregnancy, and to understand how maternal exposures affect placental development and physiology.

- Assessment of preconceptional exposures is of value in understanding their potential role in the very first phases of development after conception and during embryo development, notably through epigenetic alterations. It is difficult from the maternal side unless the cohort has started before pregnancy and samples have been collected close enough to the estimated conception date. Preconceptional exposures may be easier to approach on the paternal side with biological samples from the fathers if they can be collected early in pregnancy, with the assumption that his lifestyle and health status has not changed much. The same assumptions for samples collected early in pregnancy are more difficult for the mother as the pregnancy state induces a rapid change in her metabolism and may also quickly induce lifestyle changes. There are some exceptions. For example, some measurements performed in hair reflect long term exposure, and if collected in the first three months of pregnancy and segmented can provide information about maternal pre-conceptional exposures.

- For the post-natal period, biological sampling of the child can be undertaken at any postnatal age to provide information about child development, health and post-natal exposures. Of course, non-invasive samples are preferred at young ages. Maternal milk has a special status as a sample from the mother used to assess post-natal infant nutritional intake and exposure to other compounds present in breast milk.

The choice of the type of samples to be collected in a cohort depends on three main factors: (i) the objective of the cohort, (ii) the setting of the contact with the child and parents and thus the feasibility of sample collection, and (iii) the budget.

For cohorts with visits organised in clinical centres, collection of all type of samples can be envisaged depending on objectives and budget. For cohorts obtaining data remotely or through home visits, some samples can be easily collected by the participants themselves at home if they are given clear instructions and accessible procedures. They have to be subsequently sent by post to a storage centre. Hair, nail clippings, and teeth are among the easiest as they can be stored and sent at room temperature. Blood spots are increasingly used for the same reasons but they are less acceptable for children at young ages. Other samples such as saliva, stool, and urine can also be considered but need to be stored in a freezer and sent quickly after collection. The duration of non-refrigerated transportation time will limit their use to analysis of molecules that remain stable under these conditions. For stools, there are sampling kits that contain preservative liquids that limit sample degradation for a few days.

The type of sample affects the type of biomarkers that it will be possible to measure. Blood and urine have the advantage of providing information on a large variety of biological process and on contaminants and their metabolism. However, especially in blood, the concentration of some molecules vary according to nutritional status and chronobiological rhythms, and are secreted with pulsatility or have short half-life, and these conditions needs to be controlled to interpret results and compare between participants.

Blood is also used for DNA extraction from white blood cells. DNA is a stable molecule. On the other hand, RNA used for transcriptomics is quite fragile. It needs specific sampling tubes with preservative and has to be frozen fast at low temperature (- 80°C) for long term storage. Hair, nail, and teeth are mainly used for assessment of exposure to contaminants, with the advantage of reflecting average exposure during long periods. For example, it is considered that the more proximal cm of the hair contains average exposure during the past month while long hair can represent average exposure for several months, covering the whole pregnancy for example. Baby teeth form during fetal life and depending on the teeth can reflect exposure during pregnancy or early childhood. However, analytics are less developed for these samples than for those used in every day medicine. They are performed in few specialized laboratories and are usually expensive. Nevertheless, they are developing fast and now extend to other molecules than contaminants, such as for example steroids for the hair, with the same advantage of integrating secretion over long period of time.

Saliva samples are non-invasive and offer several possibilities. DNA can be extracted from white blood cells and also from epithelial cells found in saliva, but the yield is much lower than with blood. Numerous biomarkers present in blood can also be found in saliva (hormones, antibodies, interleukins).(4) More recently with growing interest about the influence of our microbiota on human health, saliva has also been used to assess its specific microbiota.

Meconium has been used to screen for markers of exposure to drugs, alcohol, tobacco, and some pesticides. With the development of metabolomics, this fluid is now being investigated for new biomarkers (5). Microbial DNA has also been found in meconium and is supposed to reflect early fetal gut colonization (6).

Collection of faeces in infancy and at older ages has been performed in some recent birth cohorts to understand early gut colonization by microorganisms, its determinants and how it affects later health.

Lastly, birth is unique as it offers the possibility of the non-invasive collection of two tissues, placenta and cord (including cord blood). The advent of epigenetics has reignited interest in their collection. Epigenetic marks are tissue specific, and, for placenta and cord, may serve as biomarkers of alteration induced by exposures during pregnancy.

Sampling and processing

The first step in all sampling procedures is getting consent from the participant. For children, consent from the parents or legal representatives is necessary until the age of their legal majority. Depending on the type of sample and the analysis planned, the ethics committee may require explicit consent from both parents, or accept that only the parent present at the time of the sampling will sign the consent form if the other parent has previously been informed and given enough time to express opposition to the sampling. The latter situation is by far the easiest to operate in practice.

Clear information to the parents, and child as soon as he/she is able to understand (from about 6 years), precedes signature of the consent form. Use of assent forms for children who are able to understand the issues is helpful to ensure that the child is willing. A particularity of biological sampling for cohorts is that part, if not all, of the samples, are stored for later analyses, the nature of which is not known at the time of the sampling. This has to be clearly explained to the participants. The broad objectives of the sampling needs to be given to participants, as well as the procedures for informing them more precisely of the projects that will later use their sample and the way they might indicate if they do not want to participate in a given project. Usually, the broad scope of the use of samples is provided so that participants can understand the range of investigations that might be conducted. Some studies, for example, have specified to participants that no analyses will be conducted for specific disorders that can be diagnosed from a single sample.

It is beyond the scope of this chapter to detail sampling procedures for all of the samples that have been presented in the first paragraph. They are changing with time as new material and techniques become available. We just aim to present a few points that are important to consider:

- As many of the analyses that will be performed are not usually known at the outset of a cohort study, it is useful to plan different samples and processes to be able to accommodate as many analytical techniques as possible. For example, for blood, it is usual to store plasma, serum, whole blood, and buffy coat. Establishing a lymphocyte line collection as a source of DNA or to study cellular phenotypes may also be considered, although the extra cost needs to be anticipated.
- The down-side of the above statement is that more complex sampling and processing procedures are prone to more sources of variability. Quality controls have to be planned to maintain homogeneity over time and across technicians and centres, if several are involved.
- Time between sampling, processing and storage needs to be kept as short as possible. However, depending on the conditions of the sampling (e.g. sampling at home, births during nights or week-ends), some delays may be unavoidable. It is important to keep track of all pre-analytical conditions related to timing, transportation and storage conditions as they may preclude some analyses. In any case, it will be useful to take them into account when results are analysed to reduce the induced variability and increase the power of statistical analyses.(7)
- The measurement of environmental contaminants requires specific precautions to ensure that the sampling, processing and storage of material will not be a source of contamination. One helpful approach is to store some shadow samples containing just distilled water that then undergo the same processing steps as the biological samples. They may be used at the analytical step to check the absence of contamination induced by the processing of the sample.
- Repeated samples over a day or several days subsequently pooled or, for urine, 24h collection, can also be considered for measurements of molecules with a short half-life in the body. Taking such repeated samples or sampling over a longer period, lowers intra-participant variability and increases the power of statistical analyses.
- Accurate identification and tracking of the samples throughout the process is vital. The use of pre-printed labels resistant to humidity with bar codes that can be digitally read has become a standard.

Biobanking

Samples collected for a cohort will be stored for a long time as most of them will not be used before enough events of interest have occurred during follow-up. Storage in -80°C freezers or in liquid nitrogen refrigerators are recommended to prevent degradation of the sample over time. Samples are often aliquoted in small quantities and stored in several subsamples. The aim is to be able to perform different types of measurements from the same sample without thawing and refreezing the sample, as the frost-defrost cycle can alter some molecules. With the progress of analytical procedures, samples as small as a few hundred microliters are enough to perform many measurements in plasma or serum. To be on the safest side, some investigators opt for a duplication of their biobank in two separate locations.

It is important that samples are stored in a professional and secure biobank that will guarantee their integrity over time and will be able to retrieve them with efficiency. The quality and security of the associated information system is also vital. There is little point in storing samples if they cannot be accurately and easily retrieved when needed for analysis. Also, flexible databases for identifying and cataloguing samples are needed that allow for interrogation about the number of different types of sample for specific groups that would be available for analysis.

As stored samples are a finite resource, the decision to use them requires some strategic thinking. It may be wise to preserve a fraction of samples from each participant for many years of storage. Analytical techniques that need small amounts of biological material should be preferred. 'Omics that offer large potential for analyses in relation to many outcomes are good choices. However, techniques are improving fast and there is always a trade-off between waiting for the best analytic performance versus analysing the samples quickly and gaining useful scientific insights. Analyses based on very old samples may have less public health relevance in current times. Maintaining a biobank is costly and the worst outcome is that the biological sample collection is underused.

1. Huhn S, Escher BI, Krauss M, Scholz S, Hackermuller J, Altenburger R. Unravelling the chemical exposome in cohort studies: routes explored and steps to become comprehensive. Environ Sci Eur. 2021;33(1):17.
2. Brennan L, de Roos B. Nutrigenomics: lessons learned and future perspectives. Am J Clin Nutr. 2021;113(3):503-16.
3. Anderson RM. Cytogenetic Biomarkers of Radiation Exposure. Clin Oncol (R Coll Radiol). 2019;31(5):311-8.
4. Chojnowska S, Baran T, Wilinska I, Sienicka P, Cabaj-Wiater I, Knas M. Human saliva as a diagnostic material. Adv Med Sci. 2018;63(1):185-91.
5. Peng S, Zhang J, Liu L, Zhang X, Huang Q, Alamdar A, et al. Newborn meconium and urinary metabolome response to maternal gestational diabetes mellitus: a preliminary case-control study. J Proteome Res. 2015;14(4):1799-809.
6. Silverstein RB, Mysorekar IU. Group therapy on in utero colonization: seeking common truths and a way forward. Microbiome. 2021;9(1):7.
7. Mortamais M, Chevrier C, Philippat C, Petit C, Calafat AM, Ye X, et al. Correcting for the influence of sampling conditions on biomarkers of exposure to phenols and phthalates: a 2-step standardization method based on regression residuals. Environ Health. 2012;11:29.

VI. Data management

Introduction

Ultimately, the value of a cohort is measured in terms of the quality and accessibility of its data. Those analysing the data need to understand how the data are structured, how they were collected, what issues there were during collection, the context and rationale for the cohort, full documentation of the data processes such as audit trails in relation to the data cleaning, and comprehensive metadata and details of access procedures. All these features take considerable effort to manage, document and make available to others. However, they are vital if a cohort is to have widespread use, even by successive generations of the local research team. This chapter outlines some of the major issues to consider in relation to data management.

Structure of data

The structure of data is important to consider before any data are collected. Cohorts can be complex. Even at their simplest, the types of cohorts within LifeCycle collect data on one child, the mother, and usually the father, thus giving data on triads who need to be linked to each other. More complexity arises when siblings and multiple births are included. Data on each of the mother's pregnancies needs to be related to the correct child (or to more than one in the case of twins and triplets), and so the structure needs to work for the data analysis. There is, in effect, a nested structure of child, within pregnancy, within mother. Adding to that is the need to consider data on fathers, and ensuring that the father's data are linked to the correct child. An example family structure is given in the figure below. This is actually quite a simple structure. It could be even more complicated if either or both of the fathers had children in the cohort with other women as the mothers. Furthermore, when considering the upbringing of Child 1 and Child 2, Father 2 may be their stepfather and could therefore be at least as or

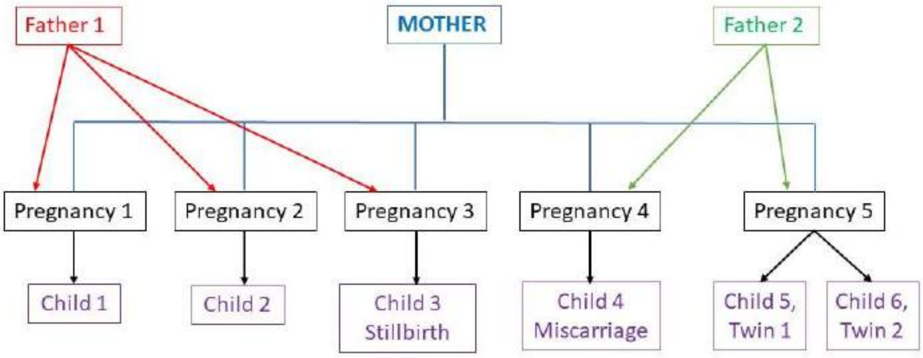

Figure 1: Example of a family structure

more influential in terms of the social circumstances and behavioural factors affecting the child than the biological father, so linkage to the stepfather might be important too. There may be different stepfathers who influence the child at different ages. If the cohort staff plan to collect data on all these individuals, then appropriate linkage must be established at the outset with individual ID numbers and links to the pregnancy, mother and father ID and the options for stepfather and even step-mothers too where necessary. Establishing the structure at the outset is vital, and drawing figures such as the one shown above can be helpful in developing the way in which the data can be set out, so that all the linkages are correctly in place. This is a vital step, for anything other than the most simple of cohorts.

Documentation

All too often documentation is overlooked and then done retrospectively. Few researchers relish this task; producing papers and bringing in grant funding are seen as higher priorities. However, cohorts last for many years, and those involved in the cohort at the outset are rarely still involved decades later. The information is often in their heads, but then gets forgotten. Later generations of researchers struggle to find out what happened in earlier phases of the cohort. The legacy left by the founders of the cohort is only of value if others can understand what was done.

Two main aspects need documenting. Details of the data variables and file structures are vital but so is the context in which the data were collected. Planning and structuring the documentation at the outset can save considerable time later on, even if the structure is modified as the documentation grows.

Data

Every variable collected in the cohort needs to be documented. Largely this will be done through the metadata (see below). However, the structure of the data needs to be described too. For cohorts that collect relatively little data, it might be possible to keep all the data in one file, but for most this is far too unwieldy. If the data are collected in waves, then different files may be used to store data for each wave (e.g. birth, 1 year, 2 years, 3 years, 5 years etc, depending on the data collection points). Some data may be collected in an on-going way, perhaps intersecting and linking with routine data, and those data may need to be kept separately from the planned waves. For example, heights and weights might be extracted from routine health clinic visits with some children having many measurements and others few. Although the health service may have target times for children to be measured, in reality they will be measured when they attend a clinic, and the timing will be variable. Cohorts focusing on illness episodes may need to keep those records separate from the planned waves too, as they will not apply to all children at the same time point. Giving thought early on as to how the data can be managed and documented is important for avoiding problems later.

Context

The context and rationale for the cohort needs to be described. Are there geographical boundaries, and what was the ethnic mix of the target population at inception of the cohort? How were participants recruited? How do the characteristics of the cohort differ from the target population? Such details may be documented in a cohort profile paper of the type published in various epidemiological journals, most notably the International Journal of Epidemiology. Publishing such a paper is important for getting the details of the cohort into the public domain, and assisting in maximising its use for research.

Each dataset needs accompanying documentation about how and when the data were collected, which cohort members were included in each wave and which had extra data collection on particular topics. Were some cohort members specifically excluded at particular times and if so why, or is the fact that they have no data in any particular subset of the data simply because they refused to provide it?

Each cohort will have different contexts and reasons for decisions taken about inclusion and exclusion, and these will need to be clearly described. This is vital for interpreting analyses and understanding the associations between variables that are identified. Selection bias is a concern for cohort studies,(1, 2) and it increases as the cohort follow-up time extends and some study members drop-out. Collider bias(3, 4) is a concern in analyses and understanding the context of the cohort helps to characterise the extent of such biases.

Types of data

In any cohort study, data come from a number of sources but need to be stored in a central accessible way. Some data are simple and require relatively little onward processing, while others are much more complex and need detailed attention. The issues are summarised here but more detail on data collection is provided in Chapter IV above.

Paper questionnaires

Traditionally, this has arguably been the most common method of data capture. Questionnaires are posted or given to cohort members for self-completion, or a researcher interviews the participant and records the answers on the paper questionnaire. Interviewer-administered questionnaires have advantages over self-completion in that the researcher can provide additional explanations to participants who do not understand the questions or when questions are complex and need prompt cards. In addition, the researcher can ensure greater completeness of the questionnaire; when participants complete the questionnaire themselves they may omit answers to questions that they do not understand, and some may have fun making up silly answers. However, the cost of interviewer administration is markedly greater than using self-completion, which is a considerable barrier to its use. Thus, self-completion of questionnaires by participants may be preferred.

Electronic questionnaires

Increasingly, electronic methods are being use for data capture. In days when access to computers and the internet was less widespread, this started with interviewers inputting participants' answers into a laptop or hand-held device. In clinic settings, this was sometimes done online and gradually this increased to use in participants' homes, but good access to the internet is needed. Now such methods are in widespread use with large amounts of data being completed online in clinic and home settings. Even more widespread now is for participants to self-complete a questionnaire online, with links to the questionnaire being sent out by email or text, representing a large reduction in costs. Some postal communication may still be required, however, or indeed for interviewers to visit participants' homes. In terms of inclusion, it is vital that provision is made for those who lack the appropriate equipment, do not have good internet access or find use of computers difficult, as otherwise selection bias is exacerbated.

Laboratory results

Laboratory results from biological samples taken from participants are usually sent in an Excel file or other form of spreadsheet. They need to be brought into the computer system used by the cohort and linked appropriately. Preparation for this is needed. The laboratory results must include an individual ID, which needs to be recorded on the sample before being sent to the laboratory. Bar coding or other digital identification methods can be used. The date that the sample was taken needs to be recorded somewhere; this may not be used by the laboratory and so not feature in the file sent to the researchers. Other information about the sample may also need recording and so a separate paper or online questionnaire may need to be completed at the time the sample is taken. For example, characteristics such as date and time the sample was taken, time since last meal, ambient temperature, date of last menstrual period, pulse, etc. might be needed for interpreting or adjusting the laboratory results.

Clinical measurements

Clinical measurements are usually recorded electronically or on paper by the person who takes the measurements. Examples include anthropometry, blood pressure, bio-impedance measures, spirometry, sleep laboratory measures, cognitive function and psychological assessments. Sometimes measures are repeated and then averages, medians, minimums or maximums, as appropriate, are derived at a later stage for use in analyses. Some types of equipment produce results directly and these need to be outputted into appropriate data storage.

Images

Imaging has increased rapidly over recent years, particularly in clinical studies. Examples include fetal ultrasound scans, magnetic resonance imaging (MRI), dual x-ray absorptiometry (DXA) scans, and cardiovascular imaging. These all provide electronic images that require considerable storage, for which budgeting is required. Furthermore, they need processing. The images themselves cannot be used in data analysis directly, so expert input is required in extracting measurements from the images. Some imaging systems provide measurements alongside the scans. For example, DXA scans provide measures that include fat and lean mass and bone density, but expert assessment of each scan is required to ensure that the image has been captured appropriately. For example, if the participant moves during the scan, the image will be fuzzy, resulting in inaccurate measurements that need to be discarded. Most, if not all, imaging methods require expert input for interpretation and appropriate coding.

Qualitative data

The simplest qualitative data arises from text responses to questions about opinions, politics, quality of life etc. These need coding in some way to be usable in the analysis. Such text responses are discouraged in large cohort studies as they present a large coding burden on the researchers, but may be essential to address a particular topic. More formal qualitative research may be conducted in the cohort, though usually will only involve subgroups of participants, due to the time and effort involved. Interviews or focus groups need to be conducted by researchers, and these are usually recorded, transcribed and coded. The coding is a time-consuming task, even when appropriate software is used. However, qualitative research within a cohort can provide insights that quantitative work cannot, and its use can enhance the understanding of actions and thinking within the cohort enormously. It complements the quantitative work in a valuable way. The original recordings tend to require large amounts of storage, but they may be deleted once transcribed, not least as ensuring anonymity is hard when voices can be recognised by others.

Linked data

Cohort datasets can be enhanced by including data from other sources. This requires detailed consent from the participants, and there are usually various bureaucratic and legal procedures that are needed before linkage to external data sources can be performed. In order to conduct the linkage, personal identifying details need to be transferred to the data owners of the external dataset. These details need to be quite extensive and the more information that can be sent the greater the chance of correct linkage. Some countries have mandatory identity numbers and that helps linkage greatly, though providing some extra information for confirming the link is advisable in case the number has been written incorrectly. In other countries such universal numbers are not available and more data items are required. Types of data that contribute to good linkage include current and former names, dates of birth, marriage and death, dated addresses and medical practitioner name, along with other information that may help the link to specific databases (for example, names of schools attended would be useful for education data linkage but not for health data). However, the fact that so much information needs to be transferred leads to concerns about potential breaches of confidentiality that do not arise in quite the same way from internally collected data. Nonetheless, increasingly, such linked data are being acquired by cohorts and they enhance the data considerably. Thus, information on hospital admissions, health visitor records, primary care records, medication prescriptions, data from disease registers, educational attainment and employment records, for example, can be brought in to enhance the cohort. While some of this information could be obtained from the participants, recall can be a problem, so obtaining data in this way can be useful. However, they need to be treated with caution as well. Routinely collected data are not obtained with research in mind, and quality can be compromised. For example, routine height and weight measurements may be collected simply to see if there are problems with the child's growth, and approximate measures may be sufficient, but they are not obtained using the more rigorous protocols that are normally applied in research.

Genetic and omic data

Vast amounts of data are now generated for studies involving genetic, epigenetic, transcriptomic, proteomic, metabolomic and other 'omic data. They require bioinformatics support and large computer storage. Management of such data is a specialist skill and beyond the scope of this summary, but any cohort that accrues such data will need to ensure that large computer storage can be accessed, and that expertise is available to curate, manage analyse the data appropriately. Bioinformatics expertise will undoubtedly be required.

Data cleaning

An important step in data management is the cleaning and validation of data. Ensuring errors in the data are removed at an early stage can save considerable time later on.

Paper questionnaires

Data from paper questionnaires can be scanned in, or the data typed in manually. Scanning has its challenges as handwriting can be hard to interpret correctly and experience has shown that while seemingly a quicker solution, the potential for error is great and large amounts of checking are required, which can be more time consuming and expensive than often anticipated. The alternative of manual entry into computer files is time consuming too, of course, but can be efficient when done by skilled data entry personnel. It is good practice to have two different people inputting the data, and then the two datasets are compared and differences are agreed mutually or by a third person, with reference to the original questionnaire.

Consistency and logic checks

Checking of the data is an important step prior to analysis. With electronic data capture, checks can be built into the system to prevent answers that are outside a plausible range. They can also ensure that certain questions are only answered where appropriate. For example, only if a participant says that they smoke, will they then be asked the question about how much they smoke. If participants have completed questionnaires themselves on paper, such checks cannot be incorporated and inconsistencies need to be identified and decisions made about setting data as missing, or how to modify the data depending on what other information is available to justify the choices made. Online questionnaires can build in detailed checks but errors can still occur. For example, two variables can be within range but together do not make sense; An adult height of 2m and weight of 50kg are each plausible, but not possible together. In cohort studies, the sequence of data might point to errors, such as a child whose height reduces from one visit to the next, even if the height recorded at each visit is plausible for a child of that age. Other types of data need cleaning too. Impossibly high or low laboratory results should be queried, and clinical and imaging measurements need checking.

Qualitative data collected within questionnaires in the form of text answers to questions will need careful examination and usually a coding schedule is needed in order to classify the various possible answers provided. Qualitative data from interviews and focus groups are not 'cleaned' as such, but transcriptions need to be examined and any statements that seem odd, need to be checked against the original recording. Fidelity of the transcription is very important. Qualitative data are often managed outside the core data management for the cohort, containing, as they do, data on smaller subsets. If repeat qualitative data are collected, then appropriate linkage is required between the data collections for longitudinal analysis. For mixed methods analyses though, it is important that those participants in the qualitative work can be linked to the core quantitative data.

It is vital to ensure the correct recording of ID numbers for each set of data collection in a cohort study. Errors in ID numbers ruin the linkage across waves of the data and affect any longitudinal data analysis.

New variables

Most analyses require some variables to be created. These may be derived variables using a combination of the information collected, or there may be a need to create categorical variables alongside continuous ones.

Derived variables

Some variables will be derived within waves of data collection. Commonly, for example, whenever height and weight have been measured we would calculate the body mass index (weight (kg)/height(m)^2). We may wish to use height and weight separately, but in addition have the summary BMI measure. Variables may need to be derived from data collected at different time points. For example, the mother's age at birth of the child will need to be obtained from the mother's date of birth and the child's date of birth. Duration of breastfeeding may need to draw on multiple waves of data as the child may be assessed at various time points in the first two years of life and the date breastfeeding stopped may appear in different data collection waves. Gestational weight gain requires measures from before and during pregnancy, and these may have been collected at different time points. Other derived variables draw on external datasets. One example is the derivation of height and weight z-scores which can be derived using an external standard (for example the WHO growth standards(5)). More complex, is deriving nutrients from dietary data, which may have been acquired from food diaries or food frequency questionnaires. Reference to external nutrient tables for foods is required to derive the nutrients in each food consumed by the participant and then totals for each participant for each nutrient needs to be obtained.
It is wise to derive variables as they are needed and store them in the cohort data. Different researchers do not then need to recalculate them, and, importantly, the same method for deriving them is used in all analyses.

Categorical variables

While it is generally considered better to use continuous variables in statistical analyses, where they exist, rather than to categorise them, there are certain categories that are widely used. Sometimes specific analyses within categories might be required, and for that the categories need to be derived. Sometimes a relationship with a continuous variable may not be linear and use of categories may be helpful. Body mass index (BMI) is commonly grouped into categories of underweight, normal weight, overweight and obese, and the obese category can be further divided. BMI does not have a linear relationship with health, with both those who are underweight and those who are overweight or obese having poorer health than those who are of normal weight. Thus examining the different categories separately may be useful.

Age groups may need defining, often in five-year groupings, but wider ranges can be used. We might wish to categorise some clinical measures into diagnostic categories such as hypertensive or not. Birthweight is often categorised into low birthweight or normal birthweight, while gestational age is used to define those who were born premature. A combination of birthweight and gestational age is used to define small for gestational age and large for gestational age thus providing derived categorical variables. Sometimes categorical variables need to be collapsed into fewer categories. Data on ethnicity may be collected in detail but the result is very small numbers of participants in some ethnic groups and careful combining of the categories should be considered for particular analyses. In a predominantly white cohort, the main grouping used might end up being ‘white’ and ‘non-white’, for example, for most analyses. This is not entirely satisfactory but may be necessary for analysis purposes.

Metadata

Metadata is the term used for data that describe the dataset. Each variable has to be documented, the values it can take need to be specified, and units of measurement recorded. The metadata are vital for anyone wanting to use the data, as without them the variables have little meaning.

Data structure

The structure of the data needs to be clear to anyone who wants to access the data. In a cohort study, it is most unlikely that all the data can be stored in one large file, so sections of the data will be in different files. There may be separate files for different waves of the data, and within those for different aspects of the data collection, for example, data from a questionnaire, clinical measurements, and laboratory results all collected as part of the same wave. If data are collected from routine data then structuring them in a meaningful way is necessary. Naming the files appropriately and providing guidance on the structure will help users of the data enormously.

Variable names

Historically, variables names in most software had a maximum length of eight characters. This rule has been relaxed in recent years and now variables can be much longer, though they cannot contain spaces. It is worth thinking carefully about a naming strategy for variables. Names that are too long can become unwieldy; it may be tempting to use the question from a questionnaire as the variable name such as: “Would_you_say_that_your_health_is excellent_good_fair_or_poor?‟ However, names get truncated in software and this would become difficult to distinguish from other questions that start with “Would_you say_that_‟. In contrast, names such as var1, var2, var3 etc. are unhelpful as they give no clue as to what the variable contains. If data on the cohort are collected at specific ages, then being able to identify the collection wave from the name can be helpful. One strategy is to use a prefix letter or combination of letters to identify the wave. Thus, names of all variables recorded at recruitment might start with the letter ‘a’ and those from the first follow-up with ‘b’ etc. Note that some statistical packages, such as Stata, do not accept variable names that start with a number, and hence the suggestion to use letters. Numbers can be used at the end of the variable name though. Much data collection is repeated across waves, so consistent naming of the same variable is helpful. Thus, aheight, bheight, cheight etc. could be used to indicate the measurement of the height at wave 1, 2 and 3 respectively. Numbering at the end of the name could be used too, such as height1, height2, height3, but this presents challenges if multiple measurements are made and stored in the data. Height might be measured three times and the average then used in analysis. If the three measurements are given names height1, height2 and height3 and the average is heightave, a prefix indicating the wave is easier to interpret than adding another number at the end (i.e. aheight1, aheight2, aheight3 and aheigtave, rather than height11, height21, height31 and heightave1 for the first wave). The age at which the child was measured could be an alternative indicator. Once variable names have been used in an analysis, it can be difficult to change them for subsequent analyses, as it causes confusion; old analysis programs can no longer run successfully without changing the variable names in the program, and that presents a problem for replicability of analyses. Even if the old variables are kept in early versions of the cohort data, clear documentation is needed to show how the old variables map to the new ones. Different cohorts will have different requirements and different methods of data collection, but a consistent, helpful and clear naming strategy is worth developing at the outset.

Variable labels

An important part of the metadata is the labels for the variables. These need to describe the variables in such a way that anyone using the data can understand what the variables represent. Clear labelling needs to include the data wave in which the variable was collected (assuming the data were collected in waves), a description of what has been measured/recorded, and, for continuous variables, the unit of measurement that has been used. The latter is vital as interpretation of the variable depends on understanding the units in which it has been measured. For categorical variables, it can sometimes be helpful to include in the variable name the number of categories that the variable can take. This is particularly useful when there is more than one variable that describes different categorisations of a variable. An example is breastfeeding duration, for which there might be two variables. The first, say bf3gp, could be in three categories of never, less than 6 months, and more than six months, while the second, bf6gp, could be much more finely categorised into six categories of never, 1-2 months, 3-4 months, 5-6 months, 7-11 months, and 12 months or more.

Value labels

Categorical variables all need to have labels for each specific value the variable can take. For many analyses, categorical variables need to be numeric and not text but the value labels provide the text detail. So the variable for sex at birth of the child might take the values 1 and 2 rather than ‘Male’ and ‘Female’ as text, but the value labels would link the label ‘Male’ to the value 1 and ‘Female’ to the value 2. Omitting the labels could cause serious confusion and erroneous analyses if the values were thought to be the other way round. Value labels may simply be ‘yes’ and ‘no’ for a binary variable, such as whether the mother smoked in pregnancy or not. Labels that are more complex may be needed for multiple categories, such as the breastfeeding categorical variables described above. Value labels should be as succinct as possible, so they can be printed out in a table easily when the data are tabulated. However, they need to include the necessary detail too, so a balance has to be struck, and giving thought to clear value labels is recommended.

Documentation of data collection methods

As well as documenting the variables and the structure of the data, it is helpful if the methods used for collecting the data can be described. Were the questionnaires self-completed or interviewer-administered? What machines or devices were used for particular clinical measures? What processing has the data undergone? Given that any data collection requires ethics committee or institutional review board approval, including details of the approvals can be helpful too. Generally, good documentation of the processes can provide helpful insights to anyone analysing the data, and can assist in ensuring that meaningful analyses are conducted.

Data storage

Where and how the data will be stored needs to be considered. Storage must be secure yet enable access for those who are authorised.

Dedicated servers

Servers within the study building with access limited to those in the building can provide good security, as firewalls can be firmly in place so remote access is not possible. However, while security is important, researchers do need to access the data. The COVID-19 pandemic has shown that such systems become challenging if staff are required to work from home, and data need to be copied onto individual laptops. The need for security may be great if the data are highly sensitive or if there are promises made to participants that need to be honoured. In non-pandemic times dedicated servers have their attractions, but may prove too restrictive as working environments change. However, one area where it may be particularly important to have local storage for the data is in relation to the personal contact details of the cohort members. These have to be kept up-to-date to ensure that cohort members can be contacted about new data collection waves, but they are highly confidential and need to be stored under tight control, with access limited to a small number of people on a need-to-know basis.
The dedicated servers need to be backed-up frequently and ideally a recent backup needs to be stored in another physical location in case of fire, flood or other disaster damaging the building that houses the servers.

Institutional servers

Most cohort data are stored on institutional servers. These present challenges too as the security is often not as tight as might be imagined. Institutions have large computing teams who have access to all servers and yet are not connected directly to the cohort management group. Large institutions require people, such as IT staff, to be able to access their servers and so their firewalls may not be as tightly controlled as one would like for storing personal data on cohort members. This is particularly important to consider in relation to the personal details of cohort members, and caution is needed if such data are placed on large institutional servers. Most institutions have a backup policy, but this should be checked before storing cohort data on the servers to ensure that the backups are conducted adequately and frequent enough.

Cloud storage

Increasingly, data are being stored in clouds. These are run by external organisations that put large amounts of money into their security, and they are an attractive option. However, the fact that they are hosted by large well-known companies does mean that hackers can see them as a challenge and are more likely to target them than a local server; if the security is tight enough though they will fail. As with institutional servers, putting the cohort members contact details in a cloud may not be advisable, due to the potential risks of them getting into the wrong hands. Cloud storage is usually backed up by the cloud provider, but the details of this should be checked to ensure that the backups are adequate.

Data repositories

Data repositories exist that hold data from many studies, and some funding bodies require data to be stored in these. However, sometimes only parts of the data can be deposited, due to consent and ethics requirements, and the full dataset may well be retained locally, or in a cloud, for access by the local research team and external users with appropriate permissions.

Software

Data are usually stored in commercial software. MS Access and SQL provide flexibility and allow the data to be exported into other software. Use of Excel is not advised: storing data in spreadsheets can be problematic, not least because data can be sorted separately from the ID and a dataset can be destroyed very easily this way.

Commercial databases are increasingly being used for clinical trials and some may be appropriate for cohort data. However, they need careful examination before buying an expensive licence or purchase.

Once data processing is complete, the data may be stored in analysis software such as SPSS, Stata or R. Data can be moved between these and other packages, but they are easier for statisticians and other data analysts to use than the database software that holds the raw data.

Audit and version control

Accountability is important in research, and it is vital that a full audit trail is available that can demonstrate changes made to the data at any stage. This might be during initial cleaning, or much later when inconsistencies have emerged during analysis. Changes made during cleaning need to be documented in the databases; the original values may need to be reinstated if the amendments are found to be incorrect, or values thought to be out of range are actually valid. Different versions of the analysis datasets will be created as amendments are made and new variables calculated and all versions need to be retained. A strict procedure for version control is needed. Analysis programs that were run on version 1 may not give the same results if run on version 20, and if version 1 no longer exists then the results cannot be replicated.

Data protection

Cohort databases hold large amounts of personal information on individuals, so data protection is vital. Some data are sensitive and extreme care is needed in thinking through where the data are stored and what controls are in place. Storage has been discussed above. An important point in data protection is separating the contact and identifying details from all the data that have been collected and are used in analyses. In cohort studies, accurate and up-to-date contact details need to be kept in order to be able to contact the participants about subsequent waves, but the analysis data must be kept separately from these. Access to the contact details should be limited to a small team of people who need to see the details. In terms of access to the data, researchers must also honour the commitments made to study participants about who can see their information.

GDPR

The General Data Protection Regulation was implemented across Europe in May 2018. All countries that are covered by the regulation have to comply with it, and no data can be sent to places in other countries that do not have equivalent standards. The GDPR builds on previous Data Protection frameworks in many countries and largely formalises best practice that has been adopted in research for some years. Compliance with GDPR is mandatory, and liaison with a Data Protection Officer is advisable to ensure that the rules are being followed.

Firewalls and general data security

Wherever data are stored there is a risk of illegal access to them. Research organisations are not usual targets for hackers but complacency is not advised. Care is needed to restrict the access to authorised individuals. In the Data Storage section above, the choice of storage was discussed, but a further important consideration is the degree of protection offered by the server holding the data. Firewalls are used to limit access, but on servers for large organisations, many people may have access, and further restrictions may be needed. Access to cloud storage is controlled by large multi-national organisations and as noted above, they may be more likely to be a target for hackers than local physical storage. Particular attention needs to be given to where the identifying and contact details of participants are stored. Thinking about data security at the outset is advisable, consulting with IT experts who understand these issues.

Anonymisation/pseudonimisation

Cohort databases hold large amounts of data on individuals. Even when all the known identifiers have been separated from the data, individuals could be identified from the data held. Large genetic databases hold data that can identify an individual uniquely, but more broadly, it is possible to identify someone likely to be in the cohort from knowing information about their social circumstances, health behaviours and preferences, their body composition, educational attainment etc. Putting multiple data items together means that individuals are more likely to be identifiable. Thus, no data can be strictly anonymous, and pseudonymisation is the best that can be done. Being aware of this is important as it should influence the way cohort researchers approach the use of the data and how the data protection and security is managed. Cohort data provide a hugely important resource that must be used responsibly and with due care and consideration for the cohort participants. Handing on data to others needs to be done with due diligence. Ultimately, once data have left the control of the original researchers, they can no longer guarantee their security. Thus, agreements need to be in place about the use of the data and whether they can be handed on to others or not.

As noted above, some cohort data are placed in data repositories that are widely available to researchers. They have protocols in place to minimise data breaches, but given the possibility of identifying cohort members from the datasets, careful consideration is needed in deciding which data can and should be deposited.

Data sharing

Data sharing has grown rapidly over the past couple of decades. The firm view is that data collected with public or charitable funding should be analysed as extensively as possible to maximise the value of the cohort. However, the whole area of data sharing is in a sense the opposite of data protection and these two priorities need balancing; limiting the numbers of people accessing the data enhances the protection of the data, but restricts its widespread use. Funding organisations have different requirements for data sharing and the rules have changed considerably over time. Most now require data sharing and have requirements in place, which might involve depositing data in a repository. Cohorts that were started many years or decades ago may not have the consent in place for widespread data sharing, which makes responding to funders’ requirements very challenging.

Agreements

In order to transfer data from one organisation to another, some form of data sharing agreement is needed. Different organisations have different requirements but each cohort needs to develop its own form of agreement that those who receive the data will sign. Legal teams in universities can make this very complex, but, ideally, each cohort develops a standard agreement that is acceptable to all and is reviewed, and updated as necessary from time to time. Each will have differing requirements that must be accepted by those receiving the data. Most, if not all, cohorts have access committees that need to review any data request before an agreement can be made.

Trust

Trust is important in research. Generally, the research community has a good reputation, and there have been few breaches of confidentiality or harms to cohort members. Ultimately, no agreement can prevent data breaches, and once data have leaked, they cannot be retrieved. Penalties can be made and disciplinary action can be taken against those responsible, but that is too late for the cohort members whose data have been released. So, we inevitably rely on trust. Training of researchers is very important, and it is vital to reinforce frequently the importance of confidentiality and honouring the trust the participants have placed in those holding their data. The attitude and culture of the organisation is, ultimately, what leads to the protection of the data it holds or receives from others.

Data requirements for federated data analysis

Federated data analyses are quite new. Each cohort's data are stored on a server in their organisation and the data are never transferred elsewhere. Analyses are conducted through firewalls with restrictions inbuilt to ensure that no data on individuals can be seen by the external analysts. Only summary results can be obtained. Conducting federated data analyses is a major plank of the LifeCycle programme. This is an important advance, which cuts out the need for data transfer from one organisation to another, and allows data analyses to be conducted by anyone in the world. The fact that researchers analysing the data cannot see data on any individual cohort member dramatically reduces the risk of data confidentiality breaches.

Agreements

Despite the lack of transfer of data between organisations, it is currently still felt important for data access (rather than transfer) agreements to be put in place. Data can only analysed by external researchers after appropriate access credentials and security are in place. The agreements are for particular research projects, which are tightly defined. Authorship of resulting papers is agreed (in terms of the numbers of authors from each cohort) and details of the tables/variables required are listed. The agreements in use in LifeCycle are evolving.

Keeping them as simple as possible will ensure the success of federated data analyses; if they become as complex as data sharing agreements then the value of federated data analyses, given the challenges of conducting them, may not be realised.

Computing hardware

One of the complications of federated data is that dedicated hardware is required. Each organisation needs to have an appropriate server with firewalls set up correctly, and the relevant software installed and updated as necessary. This all requires expertise that may not be easily available in the organisation. The issues will become simpler over time, no doubt, but they do present challenges to federated analyses currently.

Harmonisation

A key feature of LifeCycle has been the drive to harmonise data.(6) Federated data analyses cannot be conducted unless every cohort has the data in a similar format, with identical variables names and categories. Harmonised data allow for individual participant data (IPD) meta-analysis rather than pooling analyses done within each cohort, and that generally is a major advantage. However, caution is needed if the harmonisation proves difficult to do well, as the analyses will lose the refinement that exists when many of the cohorts, but not all, have detailed categorisation of key variables.

Harmonisation is a complex process though that needs to be done carefully. Sometimes only partial harmonisation is possible and this needs to be noted, and will impact on IPD analyses. For example, a cohort that collected data on employment in the following categories: employed, unemployed, student, homemaker or unknown, cannot fully harmonise to a variable that requires employed and self-employed to be distinguished. Continuous variables need to use the same units of measurement so calculations may be required, and scrutiny of ranges and checking of the data are important steps.

Within LifeCycle, the harmonisation has been performed by each participating cohort and this has been an extensive process. The details are described in a recent publication.(7)

Summary

In summary, data management is a vital part of the life of a cohort. Without it, the value of the cohort will be reduced, and the findings may not be credible. Before starting any data collection, careful planning of the way in which the data will be collected, structured, stored, organised and shared is strongly advised and can save large amounts of time later on.

1. Howe CJ, Cole SR, Lau B, Napravnik S, Eron JJ, Jr. Selection Bias Due to Loss to Follow Up in Cohort Studies. Epidemiology. 2016;27(1):91-7.

2. Nohr EA, Liew Z. How to investigate and adjust for selection bias in cohort studies. Acta obstetricia et gynecologica Scandinavica. 2018;97(4):407-16.

3. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. International Journal of Epidemiology. 2018;47(1):226-35.

4. Richiardi L, Pearce N, Pagano E, Di Cuonzo D, Zugna D, Pizzi C. Baseline selection on a collider: a ubiquitous mechanism occurring in both representative and selected cohort studies. J Epidemiol Community Health. 2019;73(5):475-80.

5. de Onis M, Garza C, Victora CG, Onyango AW, Frongillo EA, Martines J. The WHO Multicentre Growth Reference Study: planning, study design, and methodology. Food and nutrition bulletin. 2004;25(1 Suppl):S15-26.

6. Jaddoe VWV, Felix JF, Andersen AN, Charles MA, Chatzi L, Corpeleijn E, et al. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. European journal of epidemiology. 2020;35(7):709-24.

7. Pinot de Moira A, Haakma S, Strandberg-Larsen K, van Enckevort E, Kooijman M, Cadman T, et al. The EU Child Cohort Network's core data: establishing a set of findable, accessible, interoperable and re-usable (FAIR) variables. European journal of epidemiology. 2021;36(5):565-80.

VII. Ethics

There are many ethical issues that need to be considered in cohort studies. Any researcher has to conform to ethical principles as outlined in the Declaration of Helsinki. Moreover, all research has to receive approval from a recognised ethics committee or institutional review board. This report will not consider issues that relate to all types of research, but will focus on additional dilemmas and concerns that arise in cohorts such as those participating in LifeCycle.

Informed consent

Few types of research on human participants do not required consent. These largely only occur when the data are completely anonymous to the researcher. In cohort studies this is impossible as, in order to follow-up participants, identifying and contact details are required. Good practice dictates that the contact details and the data for analysis are kept separate, but nonetheless the research team know the participants and complete anonymity is impossible. So, a cohort study can only be conducted with informed consent from the participants.

Consent at recruitment

At recruitment to the cohort, the researchers do not have a clear idea how long the participants will be followed up. Each wave of data collection requires funding and that is not guaranteed at the outset. Indeed, the timing of the waves may change, depending on when funding is secured. This presents a challenge in wording the consent form. What exactly are we asking participants to do, how often will we be contacting them and for how long will they be involved in the study? None of these questions have answers at the outset. One approach is to recruit participants for the recruitment wave, making it clear that the aim is to see them at further time points, but that at each wave consent is sought again. The participants are not committed for life, and are given the opportunity to opt out at any stage and this must be made clear to them. Ideally, from the cohort point of view, retaining participants in the cohort even if they do not participate in particular waves is ideal, rather than participants dropping out permanently at the first wave in which they decide not to take part. The relationship with the participants is crucial, and while the informed consent is vital, the trust between researchers and participants is at least as important. Participants commit to a cohort and many are extremely dedicated to it, so researchers must never breach any commitment made to them.

Consent on behalf

At the outset of most LifeCycle cohorts, the primary contact is the mother, or possibly the father or guardian. The mother provides the consent, though it is not only consent for herself but also for her fetus or child. Through childhood, an adult has to provide consent for all information taken on the child.

Assent

Assent from the child is important. When the child is very young, the researchers need to ensure that the child is not unduly distressed by the data and sample collection procedures and be prepared to stop if necessary, even if a parent/guardian has provided consent. As soon as the child is old enough, obtaining written assent is advisable. Thus, the researcher engages with the child and talks through the procedures that will happen in a way that is understandable to the child. A simple child-friendly form will be completed by the child acknowledging that they have had the explanation and they are willing to take part. The child signs the form. Assent procedures clearly cannot be introduced until the child is old enough to read and write and to understand all that is involved. The age at which this can happen will vary across countries depending on the education systems in place but introducing written assent as early as possible helps ensure the child's understanding and involvement in the cohort, and builds up trust. Particular care needs to be taken when a mother is very keen for the child to take part, but the child is unwilling, and this needs delicate management. No child should be made to take part simply because a parent wants them to do so and has provided consent.

Child age when confirmation of consent

One of the challenges cohort studies face is when the child moves into adulthood. At a certain age, the child can provide their own consent. This is usually around the age of 16 years but will vary between countries with different practices. After many years of the mother providing consent, this suddenly changes to the child. The child has never provided formal consent before. This leaves open the question about whether retrospective consent is needed from the child for all the data that had been collected in the past without their explicit consent. It is a dilemma for cohort studies. While such consent could be sought, inevitably there will be members of the cohort with whom contact has been lost or they have dropped out and do not want further contact. What should happen to their data, for which consent was only provided by the mother? Generally, the practice has been that data can be retained even though consent was provided by the mother rather than the child. The challenge arises about consent for linkage to external sources for on-going provision of routine data. Once the child has reached the age at which they can provide their own consent, such providers will not continue to hand over data for which consent was only provided by the mother. This issue needs to be considered when deciding which external sources of data are sought. There is little point in getting the mother's consent for criminal justice data on the child from birth as, unless the cohort is extremely large, few members of the cohort will enter that system before they reach the age at which their own consent would be required. Even some medical data linkage sources may not be worth pursuing for a small cohort, as for example, relatively few children will have hospital admission data on any particular disorder of interest. The transition to the child providing their own consent is one that is exercising the cohort community currently, and changes may well occur to the procedures over the coming years. Clearly, the mother's consent for the data on early childhood will have to be sufficient as if consent to retain the data past the age of consent is required from the child, then large amounts of data would have to be destroyed and the cohort would become unviable. This would make running cohort studies almost impossible.

Long-term consent

Cohort data are generally kept for many years, and often longer than for other research projects. Participants need to be aware that their data may be kept for a long time. Indeed, where consent has been provided for linkage to external sources, information may be provided from those sources even decades after the original consent. It must be borne in mind that the participant may have consented to linkage to, for example, mental health records, when they were perfectly healthy, but might become less willing later in life if their mental health has deteriorated. This is a challenge for cohort studies as obtaining renewed consent every few years is impossible due to drop-out from the cohort. Sensitivity is needed here, but the important principle of ensuring that analysis datasets contain no identifying data must be remembered throughout. Also, all researchers need to commit never to try to identify individual participants from the dataset, and sanctions should be in place if they do. This concern is particularly an issue for locally-based cohorts where study participants may be known to individual researchers, and indeed, some members of the research team may be cohort participants.

Withdrawal

Participants have a fundamental right to withdraw from the research. This needs to be made explicit in the consent form from the outset, and within each consent form that the participants sign. However, defining withdrawal from the cohort is not easy. The most common form of withdrawal is for participants to say that they no longer want to be contacted again for further waves. This is relatively straightforward to manage, simply by ensuring that no further contact is made with the participant.

More complex is when participants want their data to be removed retrospectively. This rarely occurs but can arise, not least if the participant has been upset by the research team in some way. Keeping the goodwill of participants is vital in a cohort study, but participants can inadvertently get upset. The data can be removed from the latest datasets but it is a major task to remove records from each and every wave of the data. Once it is done, then all releases of data for analysis would not include those participants. However, removing the data completely from everywhere is probably impossible. Other researchers may have datasets for analysis, and liaising with all those who have access to the data asking them to remove one participant would be a mammoth operation. Also data are likely to exist on backups, and going through all backups would be another major challenge. Thus complete retrospective withdrawal becomes impossible. Furthermore, data cannot be withdrawn from analyses that have already been published. Removing data from datasets on which analyses have previously been conducted means that the results can never be replicated.

The other challenge is who can insist on the withdrawal. If a mother requests retrospective withdrawal of the data but the child does not, then separating data from the mother from that from the child is far from straightforward. For example, is birth weight a variable for the mother or for the child, or maybe both?

Use of cohort data (research ethics)

All use of the data has to conform to the latest data protection regulations. Throughout Europe the requirement is to comply with the General Data Protection Regulations known widely as GDPR.

Cohort data are not just available to the cohort researchers. They have invariably been funded by public or charitable finance and it is important that the maximum use is made of them for the public benefit. Data sharing has become the norm and is usually a requirement of funding bodies. It is unethical to retain data such that others cannot use them for bone fide purposes, as that means that the participants have spent time and effort providing data that are underused. Participants give their time willingly, often for free or for a small cost or voucher, and cohort researchers have a duty to put them to best use.

In 2016, the FAIR Guiding Principles for scientific data management and stewardship were published.(1) These argued that data should be managed under four core principles: Findability, Accessibility, Interoperability, and Reusability. These principles are described in detail in the reference and are widely available. They place an onus on the researchers to make their data widely known in an accessible form that others can use. Detailed meta-data need to be prepared for each cohort so that external users can identify whether the cohort contains the data they need for their analysis. This means that each variable needs a description and the values that it can take are documented. Moreover, there needs to be a system to help external researchers search for the variables they wish to use. This might be through a computer search or in discussion with the cohort researchers, but obscure systems breach the FAIR principles. Everything must be conducted in a spirit of openness so that the data can be as widely used as possible and that everyone has equal access. The demands of the FAIR principles are high, and many cohorts are still working to achieve them, but access to datasets is improving rapidly as a result of the demands that need to be met under the principles. In LifeCycle these issues have been taken very seriously.(2)

Interaction with participants

Participants are not just there to provide the data. They are vital members of the cohort team in the sense that they need to be included and kept engaged. Ideally some cohort members will be included in discussions about future waves of data collection and often provide valuable insights into the way in which interactions with the cohort should take place and the types of issues that need to be considered in the research programme. Not all participants can be involved in this way but good communication and engagement with the entire cohort is vital for the continued success of the research. Most cohorts produce newsletters at intervals that are sent to study participants, and keeping a website up to date is important. Other engagement activities such as cohort parties can be planned, depending on the size of the cohort and the ages of the participants, particularly if the cohort is locally-based. For children in cohorts, competitions can be organised with small prizes, and cohort members can be encouraged to engage with the cohort and comment on the way they see the research going. Whatever approaches are used, they must be done sensitively and in an inclusive way, remembering that once a cohort member has become alienated, they are unlikely to engage with the cohort ever again.

Access to own data

Under GDPR, participants have a right to know what data are held on them. All data must, of course, be held with consent of the participants, but they may not remember exactly what was collected or know about medical tests that have been conducted. Most cohorts do not have the resources routinely to feedback results of medical tests to participants, though must do so if the participant requests the information directly. This can be difficult as measurements made for research are not necessarily appropriate or interpretable in the context of an individual, and the feedback needs to be provided with involvement from a knowledgeable clinician who can encourage the participant to seek further help if necessary. It is a tricky area in cohort study management and this needs to be handled carefully, with appropriate personnel available to provide the data in a meaningful form to participants, as required.

Influence of research

Participants who engage with research are changed by it. The very fact of asking people questions may alter their subsequent behaviours. For example, asking people detailed questions about their diet may focus their minds on the poor quality of their diet and they might enact change, which will impact on later waves. This is a worry for cohort studies, but, in general, many trials have shown how hard to get people to change their behaviour and the effect may be short-lived. More concerning is if any of the research upsets the participants.

Asking sensitive questions has always been challenging and sometimes can have a negative effect on participants. Care is needed in the questionnaire design to make sure that the questions are appropriate and asked sensitively. Questions about stillbirths or neonatal deaths, for example, can be very distressing, as can questions about mental health issues. However, in contrast, some participants welcome the chance to express their distress to researchers or put the information in writing. Careful thought about such questions and discussion about their use is certainly advisable, preferably involving members of the cohort or people of a similar age. The risk of upsetting the cohort members is an ethical issue that needs to be considered at every stage.

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016;3(1):160018.
2. Pinot de Moira A, Haakma S, Strandberg-Larsen K, van Enckevort E, Kooijman M, Cadman T, et al. The EU Child Cohort Network's core data: establishing a set of findable, accessible, interoperable and re-usable (FAIR) variables. European journal of epidemiology. 2021;36(5):565-80.