Optimal Design of Cohort Studies for Maximum Learning

Keith Godfrey, PhD, FRCP

Animal studies have shown that early life exposures can induce developmental plastic responses, with major long-term consequences for a wide range of metabolic pathways relevant to human health. As a result well-designed cohort studies are needed to define which exposures underlie the link between impaired human fetal development and susceptibility to later noncommunicable diseases. The design of such studies needs to focus on core exposures and outcomes, taking into account the likely intermediary mechanisms. Collection of appropriate biological samples will enable definition of the epigenetic and other mechanisms that underlie developmental effects on noncommunicable diseases, aiding characterization of interventions. Such cohort studies need to be combined with a new generation of randomized intervention trials if we are to define the effects of maternal lifestyle, diet, and body composition on biological endowment.

Differences Between Observational Cohort and Intervention Studies

A strength of observational cohort studies is their temporal sequence, in which incident cases of disease or the development of particular phenotypes allows examination of whether a cause or risk factor precedes the effect. This reduces the risk of some potential biases, for example, lowering the risk of reverse causality or recall bias. A further strength of observational cohort studies is the breadth of analyses they can allow. New hypotheses can be tested after cohort studies have commenced and repeated measures can be ascertained over time. Cohort studies enable both multiple exposures to be related to one disease or phenotypic outcome, and examination of multiple phenotypic outcomes of a given exposure, enabling assessment of risks and benefits.

Cohort studies, however, have a number of weaknesses that need to be acknowledged. Compared with case control studies, they need to be sizeable for adequate power. Data ascertainment is often retrospective in phases, raising the possibility of recall bias, and their long duration makes them expensive and prone to changes in current practice and exposure over time. They are not useful for rare-disease outcomes and multiple testing increases the possibility of chance findings.

Compared with controlled interventions, observational cohort studies have the potential for biases of various types, including measurement, selection (particularly if losses to follow up are high), referral, observer, performance, and detection biases. Confounding is another critical issue, as it is likely that groups exposed and unexposed to a particular risk factor will differ in ways other than the risk factor under consideration. As such it is important to measure potential confounders, and undertake stratified and multivariate analyses that take these into account.

Despite their limitations, cohort studies can nonetheless give unique and important insights, informing both policy and the design of new intervention studies. Examples of insights arising from observational cohort studies undertaken by the MRC Lifecourse Epidemiology Unit include the following observations relevant to developmental influences on later health outcomes:

- The mother's prepregnant dieting behavior, health, and age influence the early trajectory of fetal development.[1]
- Many pregnant women are undernourished or vitamin-D insufficient, with lasting effects on offspring body composition.[2]
- Maternal obesity, gestational diabetes, and excessive pregnancy weight gain are common, with long-term effects on offspring adiposity.[3]
- The mother's dietary intake and parity are associated with offspring body composition.[4]
- The mother's diet influences fetal developmental plasticity and how she feeds her infant, and is itself influenced by education, smoking, and other children in the home.[5-7]
- Current infant feeding recommendations are likely to have long-term benefits, but are often not followed.[8]

Randomized Controlled Trials

Contrasted with observational cohort studies, strengths of randomized controlled trials include the unbiased allocation of subjects to treatment groups, and as a result confounders are generally equally distributed between groups. When such trials are blinded, the risk of most biases is low. In controlled experiments, with a clear time line in which intervention or treatment precedes outcome, it is easier to infer causality. While the evidence from randomized controlled trials is widely acknowledged as the gold standard on which to base policy, randomized controlled trials do have some weaknesses that often receive little emphasis. They are costly, particularly when the regulatory aspects are taken into account, and at times there may be ethical issues in relation to giving unproven treatments or, conversely,

withholding proven treatments. From the outset there is a selection bias, so generalizability may be uncertain, and randomization does not necessarily result in balanced groups. Attrition can be an issue and needs to be built into the power calculations. Also, highly controlled interventions generally test efficacy not effectiveness. Finally, it is critical to remember that these studies assess one "exposure" only, or a group of exposures, so given the cost and time involved they are generally not undertaken without a substantial evidence base accrued from cohort and other studies.

Key Elements to Consider When Designing and Interpreting Observational Studies

Study design is critical in observational cohort studies of pregnancy nutrition and later health outcomes, as shown in Fig 1 below.

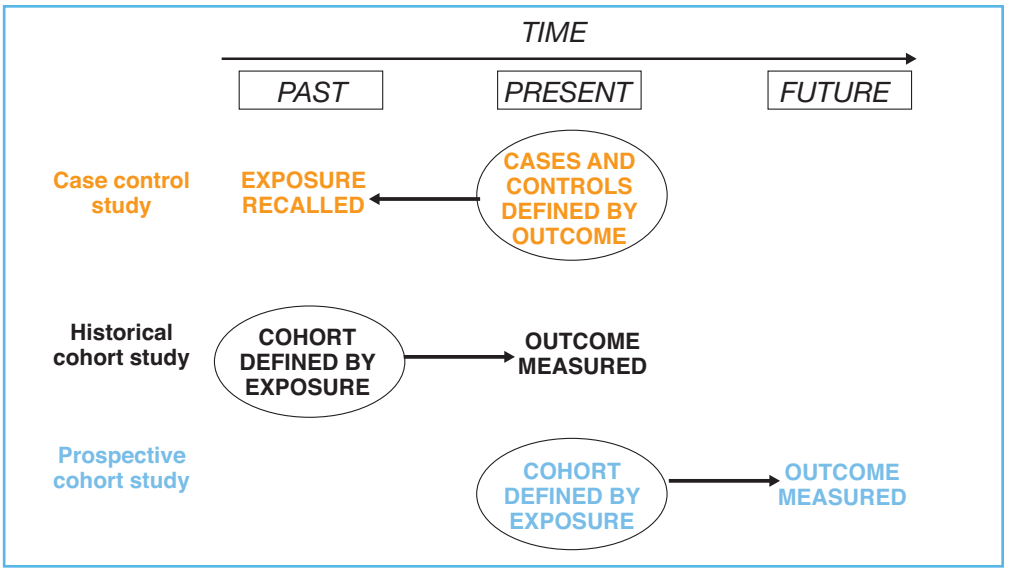

Fig 1. Key elements to consider when designing and interpreting observational studies.

While case control studies are valuable for rare outcomes and much less costly than other designs, retrospective recall of exposures is notoriously subject to bias and in isolation from other evidence provides a weak basis for recommendations.

In historical cohort studies, the cohorts are defined by a past exposure and outcomes assessed in the present time. Advantages include their lower cost and

results are available quickly, as soon as follow-up has occurred. Historical cohort studies are particularly useful if the disease has a long induction time, such as type 2 diabetes and coronary heart disease. Disadvantages of such studies are that exposure information may be of poor quality or incomplete, people or survival effects may be difficult to trace, and there is a risk of recall bias if any aspect of the exposure is recalled. Important examples of historical cohort studies include the Hertfordshire and Helsinki Cohort Studies,[9,10] which have provided pivotal early evidence in the field of developmental programming.

In prospective cohort studies exposures are defined in the present time and outcomes assessed in the future. Advantages of such studies include the capacity to ascertain detailed, good-quality exposure information and to collect biological samples enabling definition of biomarkers of later risk. Disadvantages are that they are expensive and can be prone to drop-outs. Investigators may have to wait a long time for hard outcomes and the relevance of intermediate outcome measures, such as childhood adiposity, may be uncertain. However, recent follow-up studies are starting to provide greater certainty; for example, follow up of 5000 Native Americans for 24 years showed that childhood obesity is associated with doubling of adult mortality before age 55 years.[11]

In designing prospective early life cohort studies it is important to realize that the "ideal" cohort study may be a flawed concept. There is a need to focus around the core hypotheses being addressed, as collection of quality data on all possible exposures and outcomes imposes substantial participant burden, increasing the risk of drop-outs and attendant biases. Such studies should commence in early pregnancy, or ideally preconception, as critical periods are often earlier than generally thought. New prospective cohort studies can certainly take advantage of methodologies developed for previous studies, but it is critical to appreciate that data collection needs to be context- and age-specific. This is particularly true for dietary and physical activity data; for example, in our Southampton Women's Survey[12] different dietary assessment tools had to be developed and validated for the various lifecourse stages at which we have assessed diet.[13] Exposure assessments should combine questionnaire and objective measures, for example, using accelerometry to measure physical activity and serum micronutrient concentrations to assess nutrient status.

Advances in metabolomics and in epigenetics have brought a greater emphasis on collection of biological samples for intermediary biomarkers. Different technologies require particular sample collection and processing protocols, so it is important to be as specific as possible regarding the purpose for which such samples are being collected. Even more important, however, is the choice of outcome measures.

These should include detailed assessments of body composition, including validated measures of fat mass, and not simply body mass index. We have reported, for example, that a shorter duration of breastfeeding was associated with greater adiposity measured by dual-energy X-ray absorptiometry at age 4 years, but there was no association with children's body mass index (Fig 2).[8]

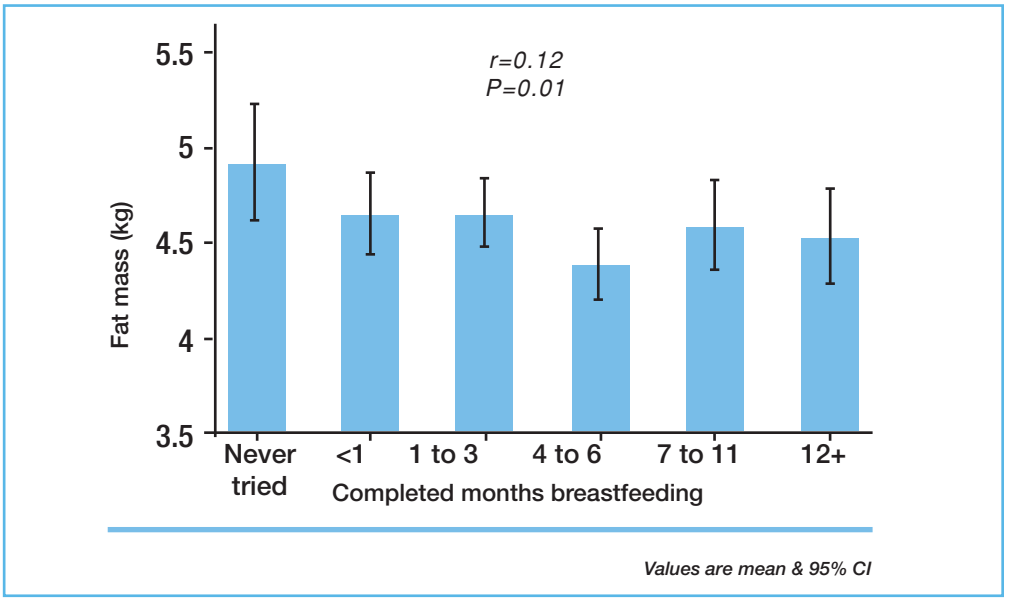

Fig 2. Shorter duration of breastfeeding associated with greater adiposity age 4 years.[8] No association with child's BMI.

Source: Robinson SM et al. Variations in infant feeding practice are associated with body composition in childhood: a prospective cohort study. *J Clin Endocrinol Metab*. 2009;94:2799-2805.

Preferably, outcome measures should take into account likely underlying mechanisms, such as assessing accentuated responses to a stress challenge as a link between early development and later cardiovascular risk. Wherever possible, outcome measures should be relevant to later health. A good example, with respect to cardiovascular disease, is to determine carotid artery intima media thickness, increasingly linked with later disease, rather than simply measuring blood pressure.[14]

References

1. Godfrey KM, Crozier S, Inskip HM, Robinson S, Barker DJP. Fetal gender and maternal health and age are associated with fetal size in early pregnancy: findings from the Southampton Women's Survey. *Early Hum Dev.* 2007;83 (suppl 1):S77.
2. Javaid MK, Crozier SR, Harvey NC, et al. Maternal vitamin D status during pregnancy and childhood bone mass age at nine years: a longitudinal study. *Lancet*. 2006;367:36-43.
3. Crozier SR, Inskip HM, Godfrey KM, et al. Weight gain in pregnancy and childhood body composition: findings from the Southampton Women's Survey. *Am J Clin Nutr*. 2010;91:1745-1751.
4. Reynolds RM, Osmond C, Phillips DIW, Godfrey KM. Maternal BMI, parity and pregnancy weight gain: influences on offspring adiposity in young adulthood. *J Clin Endocrinol Metab.* 2010;95:5365-5369.
5. Haugen G, Hanson M, Kiserud T, Crozier S, Inskip H, Godfrey KM. Fetal 'liver-sparing' cardiovascular adaptations linked to mother's slimness and diet. *Circulation Res*. 2005;96:12-14.
6. Crozier SR, Robinson SM, Borland SE, Inskip HM, SWS Study Group. Dietary patterns in the Southampton Women's Survey. *Eur J Clin Nutr.* 2006;60: 1391-1399.
7. Fisk CM, Crozier SR, Inskip HM, et al. Influences on the quality of young children's diets: the importance of maternal food choices. *Br J Nutr.* 2011;105:287-296.
8. Robinson SM, Marriott LD, Crozier SR, et al. Variations in infant feeding practice are associated with body composition in childhood: a prospective cohort study. *J Clin Endocrinol Metab.* 2009;94:2799-2805.
9. Syddall HE, Aihie Sayer A, Dennison EM, Martin HJ, Barker DJ, Cooper C. Cohort profile: the Hertfordshire cohort study. *Int J Epidemiol.* 2005;34: 1234-1242.
10. Eriksson JG, Forsén T, Tuomilehto J, Osmond C, Barker DJ. Early growth and coronary heart disease in later life: longitudinal study. *BMJ.* 2001;322:949-953.
11. Franks PW, Hanson RL, Knowler WC, Sievers ML, Bennett PH, Looker HC. Childhood obesity, other cardiovascular risk factors, and premature death. *N Engl J Med.* 2010;362:485-493.
12. Inskip HM, Godfrey KM, Robinson SM, et al. Cohort Profile: the Southampton Women's Survey. *Int J Epidemiol.* 2006;35:42-48.
13. Marriott LD, Inskip HM, Borland SE, et al. What do babies eat? Evaluation of a food frequency questionnaire to assess the diets of infants aged 12 months. *Public Health Nutr.* 2009;12:967-972.
14. Polak JF, Pencina MJ, Pencina KM, O'Donnell CJ, Wolf PA, D'Agostino RB Sr. Carotid-wall intima-media thickness and cardiovascular events. *N Engl J Med.* 2011;365:213-221.

Q & A

Q: I would like to ask about cut-off points—for example, vitamin D less than 25 nmol/L or energy below 2150 kcal. You say there are even more gradations and perhaps a greater range. How do you interpret studies in that context?

Dr Godfrey: This question goes to the heart of what we are doing—the grade effects across the range of exposure. That might be the case for some exposures. Or are there effect thresholds? Typically in the first layer of analysis we tend to group subjects into quarters or fifths of the distribution. But those categories are not biologically driven; they are statistically driven. And you are right that we need biologically driven models. Typically those require a much more substantial number of subjects to be secure thresholds. To model the data, they often need to be combined with other study designs to say, for instance, that we know there is activation of the parathyroid hormone access in relation to vitamin D below a certain level.

In cohort studies, I think that we can make only a certain degree of progress, but it is critical to undertake those analyses. There is huge disagreement, particularly between the US and Europe, about what an appropriate vitamin D threshold should be. And 25 nmol/L was just one of the thresholds.

I hope the randomized controlled trial (RCT) we are undertaking will allow us to define thresholds with greater certainty. We do not think it is ethical to randomize subjects who are clearly vitamin-D deficient. So 25 nmol/L is an intermediate level of vitamin D. The RCT design does allow us to tackle those thresholds, at least to a degree.

Q: It takes a lot of effort to collect every little piece of data on every subject, and so we have missing cells when, for instance, a child wriggles during the dual-energy X-ray absorptiometry (DXA) or somebody forgets to write down the height or whatever. Could you speak about your views on the imputation of data and the limits of that when doing major analyses?

Dr Godfrey: Sure. There are statistical methods now for imputing missing values in long cohort studies, although they are not perfect. More important is to take care to get as few as possible wriggling children in our DXAs. We take steps to help keep them still, such as ensuring that they are fed before they go into the machine. For older children, we spend hours sorting out decent DVDs and other materials to keep them still during DXA. By doing these things, we end up with missing data in 3% of the neonates and in about 6% of 4-year-olds. You are right, however; they can be systematically different.

Q: I recently saw criticism of some research reports because the authors used imputed data, so I am trying to bring that issue to the forefront. As we have more and more of these longitudinal studies, I think that we will have to agree on the limits of use of this statistical maneuver. What is your thought on this?

Dr Godfrey: That is right. Imputed data can introduce bias, especially when the exposure or the interest is specifically linked to missing data.

Q: In these cohorts, you collect a lot of information from the mother and from the child. How important might it be to collect information from the father as well? Dr Hanson showed some results that suggest that the influence of the father might be relevant as well.

Dr Godfrey: I think it is more important than we imagine. We have been doing the best we can to get this information, but we need to do more. Part of the problem is that a small but important group of women has had bad experiences with the baby's father after conception, and they do not want us to ask him anything. The rest have surprised us. I thought, for instance, that the fathers would be pretty apathetic and not give us blood. We get blood from about 85% of the fathers. But you are right; ideally we should have preconception information on the fathers.

Q: One of the biggest trends in nutrition research in this country in the last 2 decades has been the advent of what has been called epidemiologic nutrition. Large cohorts such as that in the Nurses Health Study have been used to make definitive statements about what we should and should not eat in terms of outcomes such as cancer. I have heard very little from the scientific community, which should be pushing back and saying, wait a minute, we cannot make those kinds of statements from these data. What are your thoughts on that?

Dr Godfrey: I agree that we need to be cautious about utilizing data from even big cohorts like that in the Nurses Health Study. On the other hand, the alternative is not great either, particularly in disorders such as cancer with long induction times. Honoring the evidence from experimental studies or from case control studies is arguably worse. If evidence from the EPIC cancer studies in the UK and Europe and the Nurses Health Study in the US gives us similar messages, we probably have a reasonable basis on which to form policy. My links with policy makers suggest that if we do not provide them with any evidence, they just make policies in the absence of evidence. So we must do the best we can with the information we have, but with caution.

Q: I was going to ask the same question, but from a different perspective. I think one of the reasons we get into trouble with observational studies linking nutrients with outcomes is because people fish their data. This is tempting when we have so much information. We want to be creative and push the envelope, but then we run a risk. Right now we are looking at vitamin D. But during my career we also have studied vitamins A, B, C, and E. So far observational studies have almost never proven the truth about these vitamins. When you begin a study, what process do you go through to decide the questions you are going to ask?

Dr Godfrey: One reason we have made some progress is that we have links with experimental work in animals that allow us to "nail" the effects of a particular nutrient, group of nutrients, or bands of nutrients. I also think that fishing is important, but the fish have to be explicit. Fishing is how discoveries are often made. Take, for example, epigenetics. One reason we have made progress is we have a breadth of studies at our disposal at which we can look and replicate, or not.

But you are right. Vitamin A and lung cancer science is littered with examples of observations that have not stood the test of time. But where there is a clear biological basis for an effect, such as that of vitamin D on animal development, regulators for the food industry are increasingly seeing such evidence as important to the assessment of claims.

Q: You recently published a paper describing following mothers' weight. I do not remember how old the children were, but I think they were 16 years old. Are you going to continue to follow the mothers into midlife? You enrolled them when they were pregnant, but by following them into midlife you may learn a lot more about what happens to those mothers and about what is going to happen to those children. This seems like part of the lifecourse approach.

Dr Godfrey: We would love to do this, but it comes down to money. We have put in bids for such a follow up, but we have not made the case and we have not secured the money.

Q: Do you think that one of the ways to improve causality association is to use randomization? A genetic marker linked to the exposure of interest and association to the outcome would strain our beliefs of causality.

Dr Godfrey: I think this approach has unacknowledged weaknesses. For example, if we have a developed pathway, genetic polymorphism, the route through which we have levels may be slightly different from the route through which the general population gets to high levels. This may be a poor example, but I think that care still needs to be taken with respect to randomization.