How to study cancers with a possible intrauterine etiology

Jorn Olsen

Department of Epidemiology, UCLA School of Public Health, University of California, 650 Charles E. Young Drive, Los Angeles, California 90095, USA

(Correspondence should be addressed to J Olsen; Email: jo@ucla.edu)

Abstract

The concept that many chronic diseases has an early, even fatal, etiology has inspired funding agencies to support large and long-term follow-up studies starting as early in life as possible. These cohort studies provide new opportunities for studying childhood cancer using data that are less biased than those from case-control studies. However, these studies have to be coordinated to reach sufficient sample sizes and a number of novel ethical concerns have to be solved.

Introduction

Several observations indicate that the time period of organ development has a profound impact on health; not only health as recorded shortly after birth, but also much later in life (1). A number of studies indicate that a shortage of nutrition could lead to insulin resistance as the fetal response is to slow down growth in order to protect brain development and adapt to an expected extrauterine environment with a shortage of food. When this shortage of food is being replaced with an abundance of food, overweight, and disease following overweight, may be a problem (2, 3). Other studies have shown long-term health effects of fetotoxic exposures, infections, and pregnancy complications (1).

Hormonal exposures may influence cell growth, cell division, and cell maturation. Toxic exposures may induce specific lesions; infections may lead to cross-reaction of antibodies that may destroy fetal tissue. All these ‘effects’ need not manifest themselves before late in life and many of these probably need additional hits before a disease occurs in childhood or adult life (4–7).

Hormonal exposures, natural hormones, or environmental pollutants with hormonal activities (hormonal disrupters) have been the suggested causes of breast cancer and cancer of the testis (5, 6). Evidence for an effect of in utero levels of hormones is still circumstantial, except for DES and vaginal cancer (6). Trichopoulos’s landmark paper (7) has been a great inspiration to many, but so far, it has been difficult to find data that allow a critical test of the hypothesis. Birth weight is apparently associated with breast cancer risk, at least according to some studies (8, 9) and is correlated with estrogen levels during pregnancy. However, high birth weight seems to be correlated with many cancer types, not just breast cancer. The link need not be the hormonal exposure as such, but simply the number of cells at birth (10, 11).

The best documented risk factors for childhood cancers are infections, mutations in fusion genes, and a high birth weight (12–14). Other potential risk factors may be a low intake of folic acid, certain pesticides, and intake of coffee or smoking during pregnancy. Almost all studies which provided data for this field of research have been the case-control type due to the rarity of...
the disease. Rather, most studies have been limited and based upon recall, making exposures during pregnancy almost impossible to detect. No studies have taken a possible multi-stage disease development into consideration in the analysis and the data collection (15). The possible set of causes is probably different at different phases of disease progression and one step in the disease development may affect environmental exposures and biomarkers, making reverse causation a possible explanation for many reported associations. It is believed for diseases like leukemia and testis cancer that the first steps in disease progression starts \textit{in utero}, driven by hormonal factors or mutagens. Being born with carcinoma, \textit{in situ} or critical mutations may make these children susceptible to other sets of exposures for onset of the disease. The time has come to use the new options the cohorts provide (16).

It has always been known that the time period of organogenesis and fetal development is important for the health of the newborn child. Recent studies on animals and humans have supported this view. It is a time period with substantial developmental plasticity. The fetus is capable of adapting to an unfavorable uterine environment and prepares itself for the extraterine life based upon its intrauterine experience. A shortage of nutrition is adapted to by slowing down fetal growth, perhaps in order to maintain brain growth. Insulin resistance may be the feed-back mechanism to slow down growth and further allows the child after birth to rapidly accumulate fat in the extraterine world. The prediction from the intrauterine life was a shortage of food in the external world and under these conditions it is wise to accumulate fat when there is the opportunity. If the prediction becomes wrong by food shortage being replaced with ‘food pollution’, the programming becomes counterproductive and we see epidemics of obesity and diabetes. To study this sequence of events requires longitudinal data collected prospectively. Several intrauterine settings may have ‘programming’ effects like hormones, infections, toxins, and these exposures may affect the different organ systems and the development of cancer, probably dependent on the timing of exposure. The problem is that many of these putative causes cannot be reconstructed back in time, making case-control studies too biased or simply impossible to conduct with reasonable valid data. Exposures to environmental pollutions and infections may leave biological ‘fingerprints’, but in general, these biomarkers will not provide information on the timing of exposure, which is expected to have crucial importance. To get this information, we need longitudinal data collected from pregnancy as early as possible, at best, even before conception. Such data may become available due to the international efforts in setting up large-scale follow-up studies related to hypotheses like, ‘fetal programming of adult diseases’ or ‘early origins of chronic diseases’ or whatever label investigators have chosen to give this broad research concept.

In many places new pregnancy/birth cohorts are being established. Pregnancy/birth cohorts from Europe are described at the website: www.birthcohort.net. The most ambitious project is called ‘The National Children Study’ (www.nationalchildrenstudy.gov). It aims to cover 100 000 newborns and their parents in the US. The costs are estimated to be around 2.7 billion dollars, but the project has not yet been funded. Studies of similar size (but not of similar budget or richness of data) have been performed in Denmark (www.bsmb.dk), Norway (www.birthcohorts.net), and China. Taken together, these and other cohorts provide unique opportunities for longitudinal studies on childhood cancers and the value of these cohorts will only increase with time if follow-up can be maintained. Prospectively collected data (almost in ‘real time’) are expected to be much less biased than data collected retrospectively after the onset of the disease under study, years after the etiologic time window. More importantly, the timing of exposure in the follow-up studies may or may not be available and the timing is expected to be of crucial importance.

The cohorts that have been established or are presently being established usually cover many research topics. Researchers collect data on multiple exposures and have to capture several end points during follow-up. Data have to be based upon a compromise between what is wanted and what it is possible to get, and must take into consideration that ‘the best is the enemy of the good’. Even when studies are set up with substantial community support, there are limits to what pregnant women will do for research. Most people have a busy schedule and it is unrealistic to expect women and their families to spend many hours on data collection, especially in large-scale studies. In our experience, (the Danish National Birth Cohort), we believe that we came close to that limit by asking for two self-administered questionnaires and four 30-min computer-assisted telephone interviews during a time period of about 2.5 years; two interviews took place during pregnancy and two at 6–18 months after birth. Some of these data collection sessions required further preparation in terms of finding prescriptions, health records, and vaccination dates, and having these documents ready when we called. We also asked for two blood samples to be taken during pregnancy and one from the umbilical cord shortly after birth (16).

Collection of biological samples provides opportunities to add additional data on external exposures and genetic factors. Although exposure assessment based upon biological sampling usually gives limited information about the start and end of exposures measured by the biomarker, measurement of such biomarkers gives the possibility of assessing exposures that the participants know nothing about. Most specific environmental, dietary, and occupational
exposures are often unknown to the exposed. Therefore it seems advisable to sample biological materials such as blood, urine, hair and placenta tissues as much as possible.

Experience also shows that studies need to include both biological and environmental data to get funded. We still expect that a large part of disease prevention in the future needs to take genetic factors into consideration, although, so far we have seen only a few examples where this has played a major role.

Life course epidemiology aims to identify preventable component causes along the route to the clinical manifestation of diseases. Therefore, this research will also improve our ability to predict diseases, not with certainty but with higher probability than what we have today. The question of the way in which the participants of such a study can and must be informed about these predictions raises major ethical questions that need to be addressed. This may be trivial if there are established ways of reducing certain risks but will be much more complicated if we have no such information on risk factors. These issues will be further addressed in the next section.

**Ethics**

Using longitudinal data from conception to death requires careful attention to ethical problems. Most of these problems fall into five categories: protection of sensitive data, getting informed consent, providing data access to participants, providing data access to other researchers, and the possibilities of making predictions.

**Data protection**

Most cohort studies in health research include sensitive data, however, most cohorts would have limited scientific value if such data were not available for the research. The amount of sensitive data in most cases will increase over time and the scientific value of data will most likely be more and more valuable over time. Protecting these cohorts from unwanted disclosure of sensitive information must be given high priority. At the same time, we have to make sure that the data are used and the sources are open to other researchers. Under certain circumstances it may be unacceptable to give priority to data protection and thereby limit legitimate and important research for the community. This is not an easy task and requires resources and planning.

It is important that the data are kept within the research community. Most examples on misuse of personal data on a large scale are of people with political or economic interests, as was seen in countries occupied by the Nazis during the Second World War or in Eastern Europe until the end of the communist regime. Personal data have been abused to put pressure on people, limit their career opportunities, or even put their life in danger. Although the political system may seem reasonable at one point in history, these data sources are to be kept over decades and political structures may change within this period.

Most data will exist only in electronic forms and thus be easier to protect, by means of password control or encrypting, than the paper data. Some data sources will be difficult to disclose, like biological specimens, but still they nevertheless have to be protected. It is standard procedure in epidemiology never to store or analyze data with personal identifiers (www.ieaweb.org, document on Good Epidemiology Practice).

Data are identified by a study-specific ID number and the link between this number and personal identification is stored in a separate room, in a locked safe. Still, most cohorts have such highly detailed data on individuals that you would be able to find people you know in the data base by combining data on sex, age, exposures, diseases, etc. It is important that all researchers with access to data sign statements of confidentiality and allow only data analyses in work places that are kept locked. Research on sensitive data should not be performed outside research offices and data should not be sent by ordinary mail. If data are sent electronically, the files should be encrypted. Computers with personal data should also be password-protected and should log all data use.

The most likely disclosure of personal data is loss of data by accident or thefts, but attempts to retrieve personal data may also come from people who have interest in disclosing data. Insurance companies or employers may have an interest in obtaining data on health of those who seek insurance or of the employees. In most countries, the police may get access to databases in severe criminal cases, but such data should not be generally given out unless other options have been examined and should only be disclosed if there is a legal obligation to do so.

**Consent**

It is generally accepted that use of personal data from primary sources is based upon informed consent, especially if data come from biological specimens. The problems in long-term follow-up studies with multiple purposes lie in the word ‘informed’. Being informed means, you know what will happen with your data in the future and it is within the nature of most long-term follow-up studies that this cannot be foreseen. Most studies based on data from these cohorts were not planned as they were conducted in the beginning and usually cannot be planned in the long term, since science develops over time. There are two solutions to this problem: (i) to ask for informed consent every time a new project is being planned or (ii) to ask for consent related to the purpose of the study only. In active cohorts, option i is not a viable solution for the
participants or the researchers. Participants become annoyed by repeated requests and the researchers will find that more and more subjects will drop out either from particular studies or from the cohort altogether.

It must be possible to provide general consent to take part in research and to let the ethical committees take responsibility for most of the specific projects or to decide that a renewed consent is necessary. In some cases, it is necessary to review the informed consent, for example, if the study provides new data that the participants have a personal interest in knowing. However, most studies will not provide any information of particular value for the individual participant.

It is reasonable that when people regret their participation in a long-term follow-up study, they have the opportunity to withdraw their informed consent at any time. They will no longer take part in future parts of the study. On the other hand, it is not obvious that they can request all their previous data to be withdrawn from the study. Researchers are obliged to keep research data from published studies for a number of years in order to give other researchers a chance to evaluate their findings. Therefore, these limitations should of course be stated upfront when the participants sign up for the study.

**Providing access to data**

It appears to be very reasonable that participants have access to their own data; that they provided themselves or are generated from analyses of the provided specimens. Several organizations and committees have proposed this as a general rule, but as a rule it is not without caveats.

Providing access to data will increase the risk of unwanted disclosure of personal data. We then have to organize the database in a way that makes it easy to look for individuals. In order to secure our databases for unwanted disclosure we do the opposite.

We will also have to make sure that people who request personal data are those who they claim they are, and that is not easy. Research departments are not specialized in the correct identification of individuals by their ID cards or similar.

A lot of research data do not make sense at the individual level and the time of collection. If this would make sense then there would be no need for the research. We cannot just provide data without providing explanations and this may be a time consuming and rather frustrating experience for both the researchers and study participants. Providing information without subsequent counseling may be unethical. However, if the research produces data that could be of benefit to the participants, then the researcher should make this information available if the participants want it.

**Providing data to other researchers**

People provide data to research in the hope that their data will be used in the best possible way for this purpose. Researchers who collected data do not own the data. Personal data can be owned only by people who provide the data. For this reason, it is necessary to make sure that there is open access to data for researchers with legitimate research aims within the purpose of the cohort. All experience shows that this is possible only if an external and independent committee administers the access to data. Many researchers who have spent years on a project feel that the data belongs to them and they will no longer be able to make rational decisions on this issue. They may want to be co-authors on all papers coming from the cohort data, or may think they have to check the quality of all publications coming from the data source, but neither of these conditions is justified. However, it is common sense that they should be given credit for their years of work in establishing such a cohort, like people are credited who developed a new laboratory test used by others later on.

The best cohorts will continue to generate important information for decades serving several generations of scientists and the open access policy will often facilitate to transfer data from one generation to the next.

**Making predictions**

Science is about uncovering and, at best, understanding the laws of nature in order to make predictions. In public health, the intention is to identify causes of ill health and remove these causes if avoidable, in order to prevent future diseases. When the EU first launched its human-gene program it was named ‘predictive medicine’. We now know that this was overselling the idea. Knowing about genetic factors is only quantitatively adding to our ability to make predictions and these still come with substantial uncertainties. However, in some cases, we may be able to make quite strong predictions for some individuals. Knowing genetic factors and environmental exposures (medicine, smoking, air pollution, etc.) may place individuals at an unacceptable high disease risk and they may not know about this. Are researchers obliged to provide this information to cohort participants even if they did not explicitly ask to be informed? If they can reduce their risk by making changes in their life by taking treatment or being medically checked at regular intervals, we probably should provide the information, but perhaps not do so if nothing can be done, unless participants explicitly have requested to get this information. When asking for informed consent at the start of follow-up, this problem should be addressed, although it can only be dealt with in a hypothetical form.

We expect that the ability to make important predictions and problems, or the opportunities they will produce, will become more frequent with time.
when both the cohort and the science mature. At present, no one will be able to envision fully the situation we will face in 20 years time.

Some of these problems are of particular relevance for studying childhood cancer. What should we tell parents who give birth to a child who carries chromosomal translocations that we know increase the risk of childhood leukemia, or who has testicular cancer in situ at birth?

### Power considerations

Childhood cancers are very rare diseases as their causes, or the combination of causes that onset the disease—the so-called causal field—are rare. If the disease is rare because its causes are rare, these causes (exposures) may be strongly associated with the disease (have high relative risks (RR)). If the disease is rare because its causation depends upon the outset of frequent component causes to act together with a rare exposure (environmental or genetic), then the frequent component causes produce weak associations (low RR). If all the component causes are rare, low RRs result. These considerations demonstrate our dilemma in using cohort studies to study rare diseases. If we focus on a single exposure, it is rare to have a high RR and thus be detectable. Unless the cohort is set up to recruit all (or many) with this exposure, the study will have low detection levels. If we focus upon more common exposures, the RRs are expected to be low.

Table 1 displays results on statistical power for single exposures in birth cohorts with a size from 100 000 to 700 000. With a sample size of 300 000 the relative risk would have to be 12 or more to study most frequent childhood cancer (leukemia) for exposure that affects less than 10% of the population. To study the rarest of the childhood cancers, like bone cancer, not even a sample size of 700 000 and a relative risk of 12 can produce more than 10–12% power for exposures that occur with a prevalence of 5 to 20% in the population.

Therefore, the strategy must be to identify a large part of the causal field, that is, to look at combinations of exposures. These combinations may be environmental or a combination of environmental and genetic factors. Fortunately, for some childhood cancers, good candidate genes are available.

Furthermore, it is of importance to study exposures in a longitudinal perspective, as some of the combined factors are expected to operate in a sequence, rather than at the same point of time.

### Conclusions

Cohort studies provide new opportunities to address hypotheses or exposures that cannot be reconstructed with the necessary precision back in time. This includes exposures which people know nothing about, or exposures that vary largely over time like dietary habits, use of medicine, and common diseases. Such exposures are often forgotten and the inborn asymmetry in data collection from cases and controls make the case-control study vulnerable to bias, especially if exposures have to be recalled. On the other hand, follow-up studies need to be very large and probably so large that several cohort studies from many different centers must be combined. Therefore, an international consortium for these cohort studies is needed.

### References

1. Olsen J. Prenatal exposures and long-term health effects. Epidemiologic Reviews 2000 22 76–81


5 Visfeldt J, Jorgensen N, Muller J, Moller H & Skakkebaek NE. Testicular germ cell tumours of childhood in Denmark, 1943–1989: incidence and evaluation of histology using immunohistochemical techniques. University Department of Pathology, Rigshospitalet, Copenhagen, Denmark.


Received 26 April 2006

Accepted 4 July 2006