When size matters

Hans-Werner Hense

Section of Clinical Epidemiology, Institute of Epidemiology and Social Medicine, Medical Faculty, Westfälische Wilhelms-Universitaet Muenster, Domagkstr. 3, 48129 Muenster, Germany. E-mail: hense@uni-muenster.de

It is longstanding epidemiology textbook wisdom that follow-up (or cohort) studies come at an expense. For one, most conditions of interest, including chronic diseases, occur in the population at low rates and only small proportions of a study cohort are affected per unit time of follow-up. On the other hand, the frequency and distribution of the exposures of interest, in particular when they are rare, pose further challenges to the informative value of a prospective study. Within a given study setting, there are essentially two options by which an optimized accumulation of the required person-time experience may be achieved: either by extending the time span of following the cohort or by increasing the number of individuals sampled into the study base. Commonly, a combination of both options is applied to adapt the strategies of study conduct to the limitations of feasibility, practicability and financial constraints.

In fact, the definition and enumeration of a cohort is often laborious and time consuming, and it needs proper planning and preparation; but it is often only the first, and not necessarily the most demanding, in a long sequence of steps which follow. A variety of factors affect the course—and thus the tangible and intangible expenses—of a cohort. Adding to the problems of rare exposures and incidences mentioned above, the lag-time or induction periods for the occurrence of different endpoint conditions may belong and have to be accommodated in the study plan. Also, the intensity and frequency of follow-up examinations, participant contacts and record validations, which are required to ascertain these endpoints have to be taken into consideration. Furthermore, the migratory mobility of the cohort may adversely affect the ability to effectively trace all individuals of the study sample, resulting in high proportions of loss to follow-up.

Therefore, the planning and execution of cohort studies poses frequently a major challenge to epidemiological research. The numbers of prospective cohort studies that are presently conducted is nevertheless considerable. Numerous modifications of the prospective design have also been proposed and applied, mostly with the aim of making a more efficient use of study data, such as, for example, nested case–control or case–cohort studies. Also, cohorts are often restricted to individuals with special exposures, for example, to certain settings of potential occupational or environmental risk, or to patients with certain clinical conditions, like cohorts of HIV patients. Readers who are interested to get to know more about the details of specific studies are referred to the Cohort Profiles that have formed an established feature over the recent years in this Journal.

Despite, or rather because of, the prevalent diversity and heterogeneity of individual cohort studies, researchers have started to join forces to pursue other strategies that offer promising perspectives when attempting to deal with the contradiction between the demand for estimate precision and statistical power (i.e. numbers of events and person-time) and the inevitable obstacles consequent on limited resources. The solution is pooling of individual cohorts, either by meta-analysis of individual study estimates or by pooled analyses of individual study data. The idea is straightforward and tries to make optimal use of what prospective data are available in the epidemiological research arena. This development has gained momentum and attracted supporters in many fields. There are, however, also trade-offs and limitations that need to be considered. In this regard, the methodological differences between individual cohorts are of particular relevance because they may not be easily discounted. For example, variations in laboratory methods or in the definitions of endpoints may be too marked to permit pooled analyses. Thus, ignoring such incongruence between studies, the simplicity of the approach may occasionally be seductive and result in misleading analyses. The need for quality data of large quantity with good harmonization has been recognized, however, and recently a group of investigators joined in proposing the DataSHaPER approach to integrating data across studies. This approach, with the aim of creating vast sample sizes for bio-clinical studies, consist of two components that support the preparation of congruent protocols for data collection and provide a central reference to facilitate harmonization. The approach may be used prospectively, as a source and guide for creating harmonized questions for new studies, or retrospectively, as a structured framework for harmonizing existing studies. Of note, the benefit of having specific
methods and questions in each individual cohort, and the advantage of being able to investigate such specific issues on a smaller scale, is hence combined with the broader perspective of additional options for a synthesized analysis with other, similar cohorts. Whereas traditional cohorts, due to the limited compatibility of their methods and procedures, may be criticized as often implicating elements of inefficiency, studies with harmonized protocols may come closer to a win–win situation: while maintaining their individual profile (and individual study output), they can concurrently raise their scientific impact by contributing in an optimized manner to answering study questions which are beyond their individual statistical power.

Although the use of a large number of studies for synthesized analyses has become rather common now and opens new avenues for accumulating huge databases with enormous potentials, this last decade has also seen an unprecedented rise in the sample size of individual epidemiological studies. Specifically, the call for ever larger cohort studies has been brought forward very explicitly in recent months.10 Following the UK Biobank model,11 numerous countries in Europe and the world have set out to establish literally mega-cohorts with sample sizes of several tens to hundreds of thousands of participants. Behind the gigantic size of these studies—beating traditional cohort sizes by orders of magnitude—stands the wish to be able to study the role of genetic factors—each commonly very rare—in the development and course of complex traits and diseases.12–15 Furthermore, the interplay of environmental and genetic factors, and their impact on the causation of disease, have become the focus of these studies. Manolio and Collins10 note therefore, that the costs and inefficiencies of simply 100-fold expansion of the traditional and standard cohort models would be prohibitive. They conclude that fundamentally different approaches are required to maximize efficiency. They propose to adhere to the paradigm of the UK Biobank. This was founded on a centralized model of recruitment, data collection, sample processing and follow-up. Between April 2007 and July 2010, more than 500 000 individuals in the age range of 40–69 years were recruited into the study—a most impressive and, frankly, almost unbelievable number of people.

How could this be achieved? Obviously, in this sector epidemiology is about to enter the arena of serial and highly standardized participant acquisition. Around 100 participants per day passed through each of six assessment centres that were opened concurrently in the country. The assessment protocol was fairly basic, consisting of a self-administered touch-screen questionnaire on diet and lifestyle (about 30–40 min), some interview questions on own and family medical history, and medications, and several rather basic body measurements. High-response rates were not a focus of the recruitment activities, such that ‘diversity rather than participation rates’ were emphasized: a response rate of only ~10% was achieved.10,11

The benefits, in terms of optimized efficiencies, of such an approach are almost instantly obvious as is the selective focus on specific exposures: whereas respondents to the UK Biobank invitation could refuse examinations and questionnaire components, the drawing of blood was compulsory. Individuals not willing to agree to have 50 ml of blood drawn were not admitted, emphasizing that the research focus of such mega-studies rests on the identification of genetic and biochemical characteristics, and on their interactions. Dietary factors and lifestyle, medical history and psycho-social factors are, of course, also included in the study protocols; however, the depth of information is, not surprisingly, rather limited compared with the more traditional cohort studies. Likewise, the baseline body examinations are often basic rather than innovative, and provide only a crude assessment of somatic alterations or subclinical disease. Clearly, the inverse relation between the time spent on an individual’s examination and the desire for a high participant throughput per day will inevitably affect the dimensions of the study protocol.11

Therefore, against a background of thrust and enthusiasm of the proponents,10 it may be wise to also consider the potential trade-offs that will accompany such mega-cohorts. For one, investigators will have to struggle to really balance quantity (i.e. the quest for statistical precision) with quality (i.e. the scope, detail and depth of information accrued beyond the vials). Thus, too simple a somatic assessment may eventually lack the resolution that is necessary to precisely identify specific novel phenotypes. For example, intermediate phenotypes are targeted measures of disease progression that may be more directly related to a specific disease aetiology. Likewise, short and crude assessments of an individual’s condition at baseline will be unable to ascertain pre- or subclinical alterations. More importantly, their prospective identification during follow-up will be highly improbable given the sheer size of the cohort, which makes elaborated methods of examination almost prohibitive. In stroke research, for example, imaging represents an important tool for phenotype definition of stroke studies but can also be used directly as a trait, such as progression of white matter lesions, before the onset of overt disease.16 The identification of such alterations is presently a focus of bio-clinical research: their association with genetic and molecular traits is presumably closer than for clinically manifest end-stage disease, and the prospects for early intervention seem more promising. However, the identification of such intermediate phenotypes—beyond genetic and biochemical factors—their incidence or the age at onset, as well as their progression rates will be difficult to assess from mega-cohorts.
Clearly, a tracing of cohort members for mortality will not suffice and clinical manifestations of disease alone may be too comprehensive and lack the specificity of the various causal pathways and mechanisms involved.\(^{17}\) Add to this the logistic difficulties of identifying the onset of morbid events in vast cohorts if they are not embedded into registries with a wide coverage of the population which provide systematic notifications of, for example, cancer, diabetes or stroke incidence.

Manolio and Collins\(^{10}\) correctly point out that the amount of funding that goes into one such mega-cohort is indeed enormous. For the UK Biobank, the total cost for setting up the study was approximately $100 million, with another $7 million per year for maintenance and IT system supports. Different studies planned in different countries will certainly come up with different projected costs—how much it is going to cost is actually not clear. Thus, the National Helmholtz Cohort in Germany planned to enrol about 200,000 individuals, and started its phase of pilot and feasibility studies with a budget of over 20 million Euros. One can be sure that the visibility of these lighthouse projects of epidemiology will be enormous. On the other hand, their beacons may tend to overshadow the rest of the scenery. When economists talk of opportunity cost, they imply that such costs play a crucial part in ensuring that limited or scarce resources are distributed fairly and used efficiently. Thus, opportunity costs are not restricted to monetary or financial costs, but alternatives underutilized, output foregone or time lost should also be considered opportunity costs. There must be no doubt that vast sample size projects are needed and meaningful in the rapidly emerging field of bio-clinical studies. And we will have to learn how the rapidly evolving dynamics of their implementation, involving large-scale utilization of staff, time and general resources, will not result in compromising or stifling concurrent plans for traditional cohorts of smaller size and with a more focused approach.

The need for high-throughput technologies and the advent of quasi-industrial study procedures\(^{18}\) marks a significant change in the modes of action in modern epidemiological research. Epidemiologists of all kinds are given a wide choice of options now and it will be their task, conditional on their study objectives and research focus, to create a delicate balance between ‘bigger is better’ and ‘small is beautiful’.

**References**


16. Lanktree MB, Dichgans M, Hegele RA. Advances in genomic analysis of stroke what have we learned and where are we headed? *Stroke* 2010;41:825–32.


18. The UK Biobank sample handling and storage validation studies. *Int J Epidemiol* 2008;37(Suppl. 1).