Statistical challenges in observational cohort studies

Hof, M.H.P.

Publication date
2015
Document Version
Final published version

1 Introduction

For over a century observational cohort studies have been used to study determinants of health and disease. Within a sample from the population, we can determine the relation between health outcomes (e.g. death) and a broad range of factors as genetic markers, environmental exposures, and lifestyle determinants. Since the early days of epidemiology, the cohort study has been an effective tool to investigate these relations. For instance, during the summer of 1854 a cholera outbreak struck Soho, a neighborhood in London. By mapping out the cases of cholera, physician John Snow discovered that most of the deaths were clustered around a pump on Broad Street. Snow also observed that none of the monks in the local monastery contracted cholera. Further investigation revealed that these monks did not take water from the pump since they only drank home-brewed beer. In consequence of Snow's investigation, the handle of the pump on Broad Street was removed. Later, researchers discovered that the Broad Street pump had been dug next to an old cesspit, which had begun to leak. In the cesspit the diapers of babies, who had contracted cholera from another source, had been washed [1, 2].

More recent examples of associations that were found with observational cohort studies are the increased risk of lung cancer death for smokers compared to non-smokers [3, 4], and the increased risk of developing the acquired immunodeficiency disorder (AIDS) among healthcare workers exposed to blood of AIDS patients [5]. Currently, many large cohort studies (i.e. > 5000 individuals) are being performed to investigate complex public health questions as ethnic differences in health and growth or the development of children [6, 7, 8, 9, 10, 11].

Unlike randomized controlled trials (RCT), in which the researcher randomly assigns an intervention (or characteristics) to individuals from the sample, the researcher only *observes* the research population in a cohort study. Because the assignment of intervention groups (e.g. intervention/placebo) is random for each individual in RCTs, all groups are comparable in determinants. We cannot control

this in cohort studies, and we have to take care that all groups are comparable. When the composition of all groups is different and we do not correct for this, subsequent analyses of the sample might be biased. This is referred to in the epidemiological literature as selection bias or confounding [12]. Therefore RCTs are generally believed to give a higher quality of evidence compared to cohort studies. However, well-executed cohort studies that minimize this problem can have a large impact on science and policy making [13, 14, 15].

Cohort studies can be retrospective or prospective. Retrospective cohort studies are based on characteristics and exposures that have been present in the past and have led to health-related events, while in prospective cohort studies a sample of the population is taken and followed over some time-period. At the start of these studies individual characteristics and exposures are measured and the related health-related events are assumed to happen during the follow-up period [16, 17, 18].

A cohort study involves the execution of three steps to obtain the data (see figure 1.1); (*i*) sampling, (*ii*) measurement of determinants, and (*iii*) measurement of the outcome variable(s). The success of a cohort study is highly depended on the generalizability of the study results, which is the extent to which the results from the sample are valid for the population it is sampled from [19]. In a well-executed cohort study, the data is of sufficient quality to guarantee that the effect of confounding factors are minimized [20].

This thesis is dedicated to the design and analysis of data from cohort studies. The first three parts of this thesis are focused on solving methodological problems in cohort studies. In the fourth part, data from a large prospective cohort study is analyzed.

Part I: Sampling

Before the recruitment period starts, careful consideration is necessary for choosing the sampling design. A sampling design involves assigning the probability of being included in a sample for each individual $i = 1, \ldots, n$ in the population, denoted as π_i [21]. For instance, when we desire a completely random sample of size m (where $m < n$), we have $\pi_i = m/n$. More complex sample designs, e.g. oversampling of rare subgroups, can be achieved by changing π_i for specific individuals. Because the sampling design determines the probability that individuals

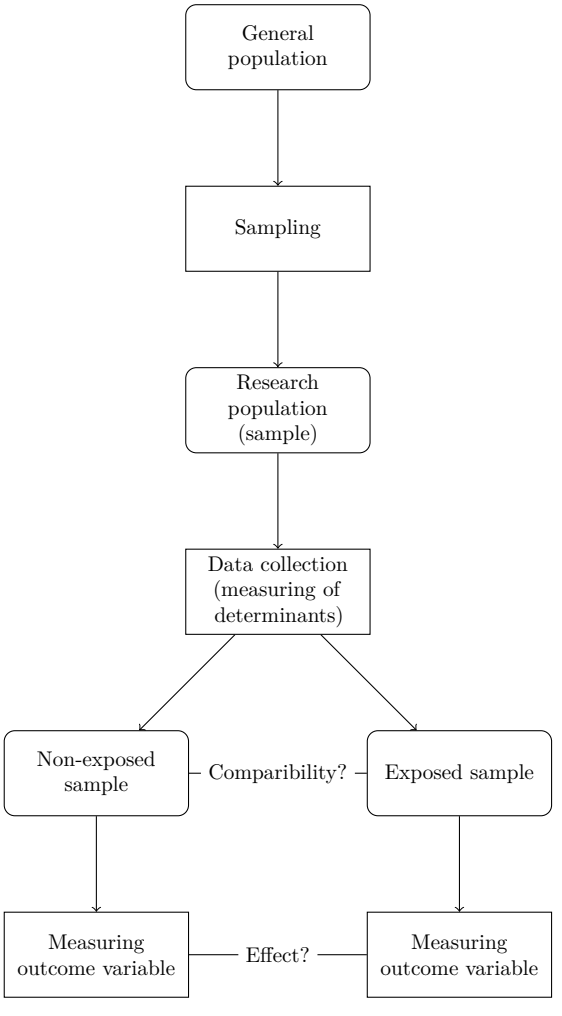

Figure 1.1: Structure of a basic cohort study, adapted from [16].

from the population are included in the sample, it has a large impact on the sample size and composition of the study. Note that both factors influence the the statistical power and costs of the study. Algorithms that can satisfy particular sampling designs are, among others, the cube method developed by Deville and Tillé [22] and the list sequential sampling method developed by Bondesson and Thorburn [23].

Two major problems during the recruitment period of cohort studies are heterogeneous non-participation and delayed response. Both problems may cause the sample composition to be different than expected. This is caused by the finding that the actual probabilities to be included in the sample were different than the expected values π_i. Consequently, we may invite too many or too few individuals from subgroups that are respectively related to a high and low participation probability. This might lead to unnecessary costs, decreased precision, or even biased results.

Usually, non-participation probabilities are often not known before the recruitment period starts. However, we might have some (inaccurate) prior knowledge, which could be incorporated in the sampling method. This could be done by using a Bayesian approach, in which we combine prior knowledge with new data that becomes available during the recruitment period. With this combined information we can estimate the participation probabilities of as yet non-invited individuals. In addition, delayed response can be dealt with by treating the response of an invited person who did not respond to the invitation yet as missing data. We can impute missing responses by replacing them with the expected participation probabilities. In chapter 2, we have extended the list sequential sampling method with such an approach. Under certain conditions (the participation probability is related to observed characteristics), our extended adaptive list sequential sampling method can deal successfully with non-participation and delayed response.

Part II: Record linkage

To reduce costs, individuals in cohort studies may be followed up through existing data sources to obtain determinants or outcome data [24]. For example, extra information could be obtained from municipal registries [25], medical data sources [26], healthcare insurance registries [27], or pharmacy data sources [28]. Combining data from two (or more) data sources requires a record linkage strategy; to identify of records from the different data sources as belonging of the same

individual. When an unique identifier (e.g. patient number) is registered in all data sources, record linkage is a cheap and easy way to enrich your data. Unfortunately, an unique identifier is often not available because of privacy regulations or practical reasons. In the absence of an unique identifier, we may use partially identifying variables that are registered in all data sources, such as gender, zipcode, name, and date of birth [29].

Generally, there are two record linkage approaches; deterministic record linkage and probabilistic record linkage. In the deterministic approach, all linkage variables or a subset of linkage variables in all data sources have to agree to decide whether certain records belong to the same individual. With the probabilistic approach, we calculate the probability that records belong to the same individual. This probability is then used in subsequent analyses of the combined data, for example with the linear regression model of Lahiri and Larsen [30], or the mixture model of Chipperfield [31]. The problem of these statistical methods is that their applicability to real data is limited. The Lahiri and Larsen method requires strong assumptions on the data structure and only works for linear regression. The mixture model of Chipperfield is developed for a set of identified pairs of records that belong together and requires quite arbitrary decisions on classifying a set of records as belonging to the same individual or not. In this part of the thesis we have developed two new methods to analyze data derived from record linkage, which require less assumptions on the data structure as the currently existing methods. In chapter 3, we consider a weighted least squares approach to analyze data obtained from record linkage. In chapter 4, a flexible mixture model has been developed. With simulations and a real data illustration, we show that these newly proposed methods can be used in a large number of linkage situations (e.g. different types of outcome data, and when determinants are spread across different data sources).

Part III: Joint modeling

Accurate information about determinants and outcomes is necessary to obtain unbiased and accurate study results. Therefore, during the follow-up period of *prospective* cohort studies, determinants are often measured at multiple times. Using updated information, inferences on associations between determinants and outcomes can be improved. The frequency and the times at which determinants are measured is of great importance [24]. The larger the gap between the last determinant measurement and the time of an event, the lower the prognostic value

of the determinant for the event of interest. Moreover, when we are interested in the determinant value at a particular time, we often have to interpolate the values of the determinant. Therefore we have to use missing data techniques (e.g. mixed effect models for continuous determinants, frailty models for event data) to estimate the prognostic values of determinants at desired time points.

In chapter 5, we propose a joint model for the analysis of multiple recurrent events and multiple longitudinal markers. In recent years, joint modeling of longitudinal markers and events is increasingly used in medical research [32]. When longitudinal marker is predictive for an event, modeling the marker and the event simultaneously is necessary to produce unbiased estimates of the prognostic value of the marker trajectory for the event process.

Different joint models have been proposed for (*i*) multiple longitudinal outcomes and a single terminal event type [33, 34, 35], (*ii*) a single longitudinal outcome and multiple terminal event types (e.g. a competing risks situation) [36, 37, 38], (*iii*) multiple longitudinal outcomes and multiple terminal event types [39], and (*iv*) multiple longitudinal outcomes, a single recurrent event type, and a single terminal event type [40]. When we wish to model multiple markers and multiple recurrent events simultaneously we quickly run into computational problems because of the dimensionality of the random effects in the models of the markers, and of the frailty terms in the models for the recurrent events. Therefore, joint models for this situation have not received much attention yet. In this chapter, we propose a simulated maximum likelihood approach to deal with the computational burden for these types of large, complex models.

Part IV: Childhood growth

Childhood growth has received a lot of attention in medical research. Many studies have found a strong relation between the growth pattern during childhood and health problems in adulthood [41, 42, 43, 44]. In chapter 6, we re-investigate differences between the growth between the ages 0-3 years of native Dutch children and children from immigrant groups in Amsterdam. In the Netherlands, different growth curves have been developed for native Dutch children and children from Moroccan and Turkish origin [45, 46]. However, despite known differences, no growth charts have been developed for Surinamese children. To measure the differences in growth patterns of native Dutch children and children with an immigrant background (i.e. Surinamese, Moroccan, Turkish), we constructed new

growth charts using the longitudinal data from the Amsterdam Born Children and their Development (ABCD) study.

For the average child, the body mass index (BMI) changes substantially in early childhood and has two characteristic points; from birth till around the age of nine months the BMI increases to a maximum, referred to as the BMI peak [47]. After this peak the BMI decreases to a minimum around the age of 6 years, which is often referred to as the adiposity rebound [48, 49]. Although a substantial amount of research has been performed to investigate the relation between the adiposity rebound and health measurements at later age (i.e. early rebound is related to an increased probability of obesity and high blood pressure [50, 51]), almost no research has been performed to check whether the timing and height of the peak are associated with (adverse) health status at later age. In chapter 7, we identified the BMI peak with the use of mixed effect models, and measured the association between the estimated BMI peak and later health outcomes (blood pressure and anthropomorphic measures).