Title
Natural history of diseases: Statistical designs and issues

Journal
Clinical Pharmacology & Therapeutics, 100(4)

ISSN
0009-9236

Author
Jewell, Nicholas P

Publication Date
2016-10-01

DOI
10.1002/cpt.423

Peer reviewed

HHS Public Access

Author manuscript

Clin Pharmacol Ther. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Clin Pharmacol Ther. 2016 October ; 100(4): 353–361. doi:10.1002/cpt.423.

Natural History of Diseases: Statistical Designs and Issues

Nicholas P. Jewell

School of Public Health & Department of Statistics, University of California, 107 Haviland Hall, MC 7358, Berkeley, CA 94611, 510-642-4627

Nicholas P. Jewell: jewell@berkeley.edu

Abstract

Understanding the natural history of a disease is an important prerequisite for designing studies that assess the impact of interventions, both chemotherapeutic and environmental, on the initiation and expression of the condition. Identification of biomarkers that mark disease progression may provide important indicators for drug targets and surrogate outcomes for clinical trials. However, collecting and visualizing data on natural history is challenging in part because disease processes are complex and evolve in different chronological periods for different subjects. Various epidemiological designs are used to elucidate components of the natural history process. We briefly discuss statistical issues, limitations and challenges associated with various epidemiological designs.

Posada de la Paz and his colleagues[1] define the natural history of a disease as the "natural course of a disease from the time immediately prior to its inception, progressing, through its presymptomatic phase and different clinical stages to the point where it has ended and the patient is either cured, chronically disabled or dead without external intervention." By "natural course" they mean here that no external intervention is applied that might change this pathway from health to disease to expression: possible interventions include preventative health measures (designed to stop or delay disease initiation), and treatment for symptoms (designed to change or delay the expression of a disease). Of course, scientists and policy-makers are usually acutely interested in what factors, including interventions, might alter the natural history of disease initiation and progression. In addition, with regard to the use of natural history information for drug development, natural history patterns may be of interest in the presence of standard current treatment strategies.

In this article, we wish to briefly discuss characteristics of the natural history of a disease that are of primary interest and discuss statistical and epidemiological considerations that are relevant to describe and estimate such properties from population samples. A previous general approach to these kinds of questions was given by Brookmeyer,[2] although we will use less technical statistical language here. Note that many of the citations provide considerable more complex and technical detail.

We will consider any possible disease outcomes here, whether rare or common, and without differentiation between infectious and chronic conditions. It is relevant to note, however, that

The author declares no financial conflict of interest.

there has been considerable recent interest about natural history studies in *rare* diseases, following a new grant program announced by the U.S. Food and Drug Administration (FDA) that includes background industry guidance regarding issues in drug development for rare diseases.[3] This agency's interest arises from the potential to use natural history information to inform treatment product and development, specifically in assisting effective clinical trial designs, and in determining intermediate and appropriate endpoints—including possible biomarker properties and surrogate outcomes. The needed natural history information for such purposes is often much less well understood for rare diseases than for common chronic conditions. Necessarily, rare outcomes raise issues in sampling populations, particularly in elucidating factors that influence the onset of diseases. At the same time, characterizing disease progression after onset for rare conditions requires identification of affected patients, often from multiple data sources.

Not surprisingly, the FDA's interest is most focused on disease progression rather than onset (in contrast to important preventive health programs). In this regard, the agency specifically has drawn attention to both (i) identification of various disease expression outcomes, and which pathways to these expression outcomes might be most responsive to an intervention, and (ii) understanding drug targets that will modify disease expression and how great an effect is needed to meaningfully alter or delay the latter. In addition, choosing the timing and frequency of patient assessments in determining changes in disease expression associated with intervention may considerably influence a clinical trial design. This is akin to determining screening schedules with regard to assessing onset rather than expression. For example, such questions underlie the scientific debate in the US regarding when to start using mammograms to screen for breast cancer[4], or colonoscopies to detect colorectal cancers[5], and how often to carry out repeat screens.

Schematic for Disease Initiation and Progression

Figure 1 displays a very simple schematic that represents the natural history of a disease. In some cases, the expression of a (fatal incurable) disease may coincide with death. Deaths that occur before disease initiation are necessarily not linked directly as an effect of that specific disease. More subtly, deaths that occur after disease initiation may not be associated with the disease or may be the direct result of the disease; in the latter case, we would count death as part of disease expression.

In any given example, it is essential to be quite explicit about such assumptions in describing properties of the disease process. After disease onset, Figure 1 illustrates what is commonly referred to as the *Illness-Death Model*. The latter is a simple example of a *multi-state model* for disease progression[6]. In most cases, natural history of disease progression focuses on that part of the schematic after disease initiation. Note that natural history studies attempt to elucidate factors that influence the overall risks of moving from state to state and the pace by which transitions occur (i.e. rates). Natural history parameters are therefore usually summarized by cumulative risks or (instantaneous) rates that may vary over time. The distinction and links between risks and rates are important and date back to our early understanding of the application of quantitative methods to the studies of diseases as exemplified by Farr's essay "On Prognosis"[7,8]—in reading this historical gem it may be

helpful to the modern reader to see how Farr's concepts relate to our modern vocabulary in statistical epidemiology.[9] The importance of comparative studies and the complexity of how a disease's natural history may change over time and geography is now a core component of our understanding. A classic example is provided by McKeown's graph showing that close to 90% of the decline in scarlet fever mortality occurred prior to the introduction of sulphonamides in 1935.[10] That this has led to a commonly held humility about the contribution of modern drugs to human health, this too is subject to considerable debate.[11]

Simple Questions about the Disease Course

For a simple example, we first consider dementia as a disease outcome. There are many important questions about the life course of dementia that are key to prevention, treatment, and policy decisions regarding current and future health resources. Such questions include determining both the prevalence and the age at incidence of dementia in a population. A related question is estimation of the lifetime risk of dementia (or risk by a certain age). Assessing what individual cofactors influence incidence, prevalence, lifetime risk, and disease expression are all key scientific questions. For example, how does incidence differ between men and women? What interventions delay specific clinical symptoms after disease onset? What is the impact of disease onset on subsequent mortality as compared to the general population? Does remaining life expectancy differ with age at onset? All of these questions can be posed in terms of the rates of progression between the various states in Figure 1 but often require complex mathematical formulations. In general terms, natural history parameters depend solely on the transition rates between the various stages represented by arrows in Figure 1 (and any intermediate disease stages not depicted in the simple schematic).

To illustrate how complex it can be to answer even apparently simple natural history questions, consider, for example, an obvious question regarding the natural history of Alzheimer's Disease (AD) as to whether the disease is more virulent depending on a patient's age at diagnosis: this might be operationalized by examining whether age at diagnosis is related to mortality after diagnosis (although this can be problematic as discussed below). This question has been previously examined[12] by investigators who used data from the Baltimore Longitudinal Study of Aging. However, such a question is surprisingly difficult to quantify using standard methods. This is, in part because there is already an obvious relationship between age and subsequent mortality in any population followed prospectively: Older individuals suffer from increased mortality as compared to younger people. Thus, one needs statistical assessments and models that extend standard survival regression methods to determine whether early age at diagnosis contributes to *excess mortality differences* (compared to later ages at diagnosis) than one would expect in a healthy population. Even this comparative difference can be problematic; neither estimated mortality survival curves for different age groupings nor estimated medians of *remaining survival times* at various ages are straightforward to interpret as a response to the question posed. In particular, a simple test on an age at diagnosis regression coefficient in a survival regression model does not usually address the fundamental question.[13] Brookmeyer *et al.*[12] tackle this challenge by calculating percentage reductions in median remaining lifetimes in AD patients in comparison to other populations. However, this again does not address the

right question: for example, imagine a hypothetical disease that always leads to death one year after diagnosis *irrespective of the age at* onset. The percentage reduction in remaining lifetime is clearly much smaller for patients with early age at onset as compared to later even though the disease affects them identically. This thorny issue is related to similar questions in quantifying the impact of environmental exposures (such as smoking) on health as captured by concepts such as *years of potential life lost*.[14] Note that Johnson *et al*.[15] describe an additive increase of 8% to background mortality rates once an individual enters late stage Alzheimer's Disease, an effect shown to be independent of age at onset or gender.

Accepting these challenges, it has long been known that a comprehensive understanding of the natural history of a disease assists investigators in designing, implementing and interpreting intervention studies that are targeted to reducing disease onset, disease expression or both. We now discuss approaches to visualizing natural history data, and various design strategies to capture and analyze such information. My discussion is necessarily brief and eclectic, focusing on only a few major issues surrounding the design and analysis of natural history data.

Life History Visualization—the Lexis Diagram

There is a long and rich history of methods to graphically display natural history information for individuals. The most prominent of these ideas is the Lexis diagram dating back to the nineteenth century[16]. Lexis focused primarily on a simpler version of Figure 1 that eliminates any intermediate stages and simply displays time from "birth" to "death" as illustrated in Figure 2, essentially a reproduction of Figure 1 of Keiding (1998)[17]. Note that "birth here" might refer to disease initiation in Figure 1 so that one focuses on the natural history after disease onset. In such case, this simple Lexis diagram might indicate longer times to expression for individuals whose disease started later in chronological time, by comparing the lines at the right to those on the left.

Lexis also discussed a three-dimensional plot that allowed for an intermediate stage, in his paper illustrated by marriage. This is illustrated in Figure 2 of Brinks *et al*. (2014)[18] for two individuals, one who develops disease during their lifetime and one who does not—a similar version is reproduced here for convenience in Figure 3.

The Lexis diagram is particularly useful as a schematic to aid understanding of how various sampling schemes will affect estimation of various natural history parameters, as we discuss further below. Keiding (1990)[19 3] provides a definitive discussion of statistical inference issues associated with the Lexis diagram. The schematic also draws attention to a fundamental statistical challenge in interpreting the risks of events such as death as represented in Figure 1: consider the risk of cardiovascular disease at chronological time, t, for someone who is at age a, and was born in year y. It is entirely plausible that such a risk will depend on when you were born (the so-called cohort effect), your age, and current treatment and diagnosis patterns at the moment t (the period effect). For example, varying dietary patterns in early life might yield different risks for individuals of the same age depending on when they were born. The same phenomenon occurs when certain treatments are available now to patients with high cholesterol and/or high blood pressure that were not

available to individuals born in an earlier era. Thus, the risk of interest can depend on each of *t*, *a*, and *y* in different ways. Yet, it is well known that these separate effects cannot easily be separately disentangled: the non-identifiability of age, period and cohort effects.[20] Modern approaches to this challenge involve complex statistical models[21] that nonetheless include subtle assumptions that permit some resolution.

Screening Tests

Before turning to a brief discussion of various population sampling schemes to collect data on the natural history properties associated with a specific disease in a particular population, we note that measurement of such samples critically depends on screening tests to ascertain the stage at which an individual is in at one or more time points. In addition, measurement of the time until disease initiation fundamentally depends on being able to detect when that event occurs for an individual. For many infectious diseases, such a test may depend on the presence of antibodies reflecting prior infection and thus indicating that the disease process has been initiated and the individual is at least in Stage 1, although the latter rarely tell you when infection occurred in the past. Similarly, diagnostic tests are also necessary for non-infectious diseases to determine whether the disease has been initiated. Again, such tests, such as the ECG, may indicate heart damage without pinpointing the timing of its occurrence.

In many cases, screening tests are only effective some time after initiation and clinical symptoms are detectable. In addition, screening tests are not always perfect and thus are subject to misclassification. This is particularly relevant when tests produce false negatives meaning that individuals with sub-clinical disease initiation are missed. Unless misclassification rates are low (as determined by the sensitivity and specificity of the test being high), it is necessary to account for these effects in any statistical analysis. In summary, the development of powerful and accurate screening tests is an important tool in yielding useful natural history information.

Note that comprehensive screening programs may necessarily disrupt the natural history of a disease if they are used as a motivation for early treatment for conditions. Thus, natural history summaries for populations may differ simply because of variation in prevalent screening policies. Similarly, screening tests may become more sensitive over time, essentially moving the detectable preclinical period earlier in the natural history of the disease (and equivalently potentially lengthening the time between disease initiation and diagnosis expression and diagnosis). This contributes to what is known as *lead time* bias in that the additional “time” in a particular stage is only due to screening properties and not to the underlying phenomenon.[21] Feinstein[23] discussed related issues in cancer related to improvements in diagnostic tests that identify stages of the disease, coined *stage migration*.

Note that the value of screening to effective treatment and improved prognosis is itself a notoriously difficult concept to quantify. Such difficulties underlie much of the debate of whether and when to screen for chronic diseases. Although it is commonly assumed that widespread screening will necessarily reduce mortality through early detection, this is far from a foregone conclusion. Based on a population-based study that used relevant control

groups, Woods *et al.* showed that screening for neuroblastoma in infants is unlikely to reduce mortality.[24]

Epidemiological Studies to Identify Risk Factors

Before turning to a broad description of epidemiological natural history study designs, it might be useful here to note that there often opportunities to understand the properties of a disease process using naturally available data include in population and medical registries. Several classic studies are of this kind: for example, David Barker and his colleagues noted a strong similarity between the geographic distribution of coronary artery disease in the 1970s and 1980s and maps of infant mortality in the 1920s. Using data from birth and death registries, a subsequent investigation showed a link between birthweight (and birth at one year of life) and subsequent adult risks of death from ischemic heart disease.[25] Similar findings were also based on individuals born during the Dutch famine of 1944–45 using a historical cohort assembled from records of an Amsterdam hospital.[26] It is likely that many future natural history studies will exploit the availability of "big data" sources to extract natural history information.

Returning to designed studies, one ideally needs a random sample of individuals who can be followed from the time origin through disease initiation (if it occurs) until death so that the time periods spent in each stage of Figure 1 are known precisely (along with any intermediate steps of interest). The sample need not be a simple random sample but knowledge of the sampling frame then becomes critical so that sampling weights can be applied appropriately. Further measurements during such follow-up periods on *biomarkers* provide key information on disease development and expression. In almost all cases, this is simply not practical for a variety of reasons. It may not be possible to follow individuals from the time origin. Estimating the transition time from one stage to another requires almost continuous screening with a perfect test. In actuality, such levels of continuous uninterrupted follow-up are expensive, unethical and practically impossible to implement in mobile populations. Thus, epidemiologists have identified many differing "short-cut" sampling schemes, all of which require careful statistical assessment as almost all designs include some form of sampling bias.

For example, consider the Lexis diagram illustration in Figure 2. If one samples a population at a fixed chronological time one necessarily misses individuals who have died before that date. (Think of such sampling as drawing a vertical line on the diagram at a fixed time on the X-axis.) That is, the sample of individuals favors those who were "born" recently and/or had long "lifetimes". A more obvious issue afflicts sampling individuals all of the same age (now, draw a horizontal line at a specific age on the Y-axis). Here, such a sample has no information on lifetimes that are shorter than the chosen sampling age. In addition, either of these kinds of "cross-sectional" samples requires subsequent follow up to determine the age "at death". Often such follow-up must be limited for practical reasons including a fixed length of follow-up after sampling or individuals ceasing observation because of migration or other reasons. Thus many "lifetimes' will necessarily be censored in that one has only partial information about the ultimate age at "death." Sometimes information on the timing of events such as disease initiation and expression can only be narrowed to a time interval

due to intermittent screening, a phenomenon known as *interval censoring*.[27] A classic example is estimation of the distribution of age at menarche in girls from New Guinea based on data from repeated annual surveys: some girls had completed development by the time of their first survey, some were lost before the last available survey before that had completed menarche, and some had not reached menarche at the time of their last survey.[28]

An additional complication arises when we allow for intermediate stages in the natural history as illustrated in Figure 1. Chronological cross-sectional sampling will determine which individuals do not have the disease in question and those for whom disease has already initiated, so-called *prevalent* cases. Again, this kind of sampling may miss individuals with shorter times between disease initiation and disease expression, obviously so if expression removes an individual from possible sampling (as is the case with death). In most situations, investigators are more interested in patterns describing *when* natural history events occur—such as the age-at-incidence distribution—rather than descriptions of prevalence properties, largely because prevalence depends on being alive to be sampled and thus introduces selection effects as discussed subsequently. The statistical links between incidence and prevalence are described in Keiding[29], Andersen & Keiding[30], and Keiding[31].

Careful consideration of the selection biases induced by sampling is thus required for almost all sampling schemes that are designed to measure natural history parameters.

Cross-Sectional Studies

Brookmeyer[2] defines the simplest sampling scheme as a random sample of individuals chosen at a fixed point in time, what we referred to earlier as a cross-sectional design. The simplest data collection system then determines which members of the sample have the disease and which do not, along with their ages at sampling, again assuming an appropriate perfect screening test. Covariates may also be measured. In its most basic form, this kind of cross-sectional information does not provide retrospective information on the time of disease onset (if the latter were available, the data structure would represent a basic form of *right-censored survival data*, for which there is a rich literature on appropriate methods of analysis[32]).

Is it possible for such data to determine the age at incidence distribution for the disease (ignoring covariates for the moment) from such cross-sectional data? Surprisingly, the answer is yes, at least with certain assumptions. First we have to ignore death as a competing risk as such individuals necessarily are unobservable. We thus can only estimate the age at incidence distribution for those individuals who do not die. This is usually reasonable in studies of disease that occur earlier in life where death is not common but may be an issue in studying elderly populations (see further subsequent discussion). Further, one must also assume that the death rate at a given age is not affected by disease initiation. That is, there is no difference in the chances of death for a 50 year old if they are disease free or if they got the disease starting at age 40; otherwise cross-sectional sampled individuals at a certain age will underrepresent those who have already contracted the disease in question. This is only reasonable if the disease in question has little effect on mortality. In addition, it is standard to assume that the age at screening is independent of the age at initiation.

With such assumptions, one can use the data on disease prevalence at sampling, along with the ages of the sampled individuals, to estimate the age at incidence distribution. This cross-sectional sampling scheme yields what is known as *current status data*, a data structure that is common in demography, economic, epidemiology and medicine. The first known use of such data that I am aware of is due to Hajnal who described how to use data on the proportions of individuals at differing ages who are already married to estimate the mean age at marriage[33]. Jewell & Emerson[34] provides a survey of statistical methods used to interpret current status data. These ideas were used in early studies of HIV transmission to sexual partners of HIV infected individuals to ascertain *infectivity* properties of the virus, particularly how time, or more specifically the number of unprotected sexual contacts influences accumulating risk of acquiring infection in the partner[35]. McKeown & Jewell[36] modify current status methods to allow for misclassified screening with tests of known specificity and sensitivity. A further example is provided by Jewell & Petito[37] who show how to estimate the age incidence distribution of Hepatitis C for women of child-bearing age using birth certificate screening data from the U.S. Birth Data Files from the National Center of Health Statistics.

Methods for estimation of age at incidence distributions from current status data can be extended simply to examine the effects of covariates on age at incidence. These are particularly simple when the covariates are fixed like gender. For time-varying covariates, the situation becomes more complex, particularly if the covariate is also only measured cross-sectionally. In such cases, there is the immediate problem of *temporality* for individuals for whom disease is present at sampling. In such cases, the data cannot determine whether the level of the covariate effected age at incidence or whether the presence of disease altered the covariate level.

Prevalent Cohort Studies

Prevalent cohorts are constructed from cross-sectional sampling of individuals, retaining only those with disease present and then following up such cases for a specified period of time. In such studies, we are no longer interested in age at incidence and its determinants but are instead focused on understanding the progression of disease from onset to expression. In rare disease settings, prevalent cohorts are more useful than incident cohorts as they avoid the very large samples of individuals that would be required to locate a sufficient number of incident cases. On the other hand, the sampling of individuals “mid-stream” during disease progression necessarily introduces a sampling bias that needs to be understood and addressed. For example, individuals with very short disease duration are undersampled by cross-sectional identification of prevalent cases[38].

Sometimes the time of sampling corresponds with a change in covariates such as treatment and so it is tempting to use time from sampling (until expression) as an outcome variable to compare treatment regimens. The biases introduced in doing so are discussed in Brookmeyer[2] and Wang, Brookmeyer & Jewell[39].

More Informative Cohort Studies

As noted previously, “ideal” cohort studies follow a population from “birth” and routinely assess individuals for both onset of disease (using a screening test) and subsequently disease outcomes. These studies are expensive and impractical in the case of rare diseases. Nevertheless, they provide substantially more information than either cross-sectional studies or prevalent cohort studies. Sometimes such studies may be unethical if a known treatment exists so that individuals receive an intervention after disease onset is detected. An alternative to cohort studies that nevertheless provides longitudinal information is the use of repeated cross-sectional studies over time. Magnus & Jaakola[40] critically appraise the use of such designs in assessing changes in prevalence of asthma and obstructive lung disease in children and young adults. Caplan *et al.*[41] compare the use of cohort versus repeated cross-sectional sampling in monitoring trends in breast cancer screening practices.

Handling Deaths in Natural History Cohort Studies

As noted earlier, deaths may occur during the follow up of cohort members being screened for either the onset or expression of a disease. For example, follow-up of patients with orthoplastic surgery involving total hip replacement is of considerable interest to determine the long-term efficacy of alternative procedures and materials. A suitable endpoint in this case may be revision surgery deemed necessary to repair or replace the original implant. In such follow-up studies, participants may be lost of to follow-up for a variety of reasons, not the least the cessation of the study at a moment where many study subjects have not yet experienced a revision. For these individuals the time from original surgery to revision will not be known, only a lower bound—technically these observations are (right-) censored as noted earlier. To take advantage of the information contained in these incomplete times to event we usually assume that censoring is independent of the time of interest (in the example, time to revision).

In some cases, censoring is caused by the death of the subject, and this might not be uncommon in follow-up of elderly cohorts. It is tempting to treat death as simply another form of censoring. This is often not appropriate for two fundamental reasons: first, it is unlikely to be the case that death is independent of the time to event. In the hip replacement example, it is possible, for instance, that a subject’s mobility may affect their risk of death and also their risk of revision, thereby violating the independence assumption. In and of itself, this issue requires that death be treated differently from other censoring events. Second, even if one was convinced that death was independent, standard methods for estimating time to the event are in fact targeted to natural history parameters *in a world where death does not occur*. In the example, this means that procedures such as the Kaplan-Meier estimator[19] estimate the population distribution of the time from hip surgery to revision, *assuming no subject dies in the meantime*. Some have thus reasonably argued that physicians and their patients are more interested in the probabilities of revisions in a certain time period after surgery and *before death*. The latter quantity requires treatment of death as a *competing risk*. The treatment of death as a censoring event or as a competing risk necessarily estimate *different population quantities*, and this is sometimes confusingly referred to as a bias in the Kaplan-Meier estimator[42]. In almost all cases, it is better to model

the effects of a covariate on a specific outcome in the presence of competing risks through cause-specific hazards[43] rather than attempting to directly model the hazard associated with the sub-distribution function that measures cumulative incidence of a specific outcome event[44]. While the latter approach has grown popular it faces fundamental challenges in interpretation as the sub-distribution functions do not relate to any mechanistic aspect of the problem as do risks acting on a (cause-specific) hazard rate, and the oft used proportionality assumptions are often not reasonable. See Haller *et al.*[45] for a comparison of the two approaches.

A disadvantage of the competing risk approach is that this estimator estimates a natural history parameter that necessarily depends on the underlying risks of death (as a competing risk) which in turn are influenced by population characteristics unrelated to the fundamental natural history process (in the example, the reliability properties of the hip replacement device), specifically the age distribution amongst other factors. Thus, this natural history parameter is not immediately *transferable* from population to population. On the other hand, the underlying natural history distribution of hip revision times (after surgery), assuming death is removed, is a parameter that may be compared meaningfully from population to population, at least with certain reasonable assumptions. The bottom line is that both parameters are potentially of interest and the context must determine which of the two parameters (or both, for that matter) are of fundamental concern. These issues carry forward into regression studies where one is interested in the comparison of times to events across subgroups of individuals. There is a rich literature on these topics[46]. Varadhan *et al.*[47] provide an introduction to some of the relevant ideas and methods.

The same issues also arise, of course, when *longitudinal changes* in disease biomarkers and other factors are being studied rather than, or in addition to the time to a natural history event. Much of the research here was developed in investigations of the natural history of HIV disease although the ideas have now been extended to studies of biomarkers of disease development in a wide variety of applications including cancer, cardiovascular disease and kidney transplantation. A quite different application occurs in the study of sexual intercourse data in prospective pregnancy studies[48]. Various statistical approaches have been suggested for joint modeling of longitudinal and survival data. Early general surveys of these ideas in this context are given by Tsiatis & Davidian[49] and Kurland *et al.*[50]. Asar *et al.*[51] provide a recent introductory tutorial on the topic. A book length treatment is given by Rizopoulos[52].

A parallel literature exists for a specific type of longitudinal data where interest focus on the timing of *recurrent events* that occur as a disease progress. Examples of recurrent events include cancer reoccurrences, repeated hospitalizations and the like. Amorim & Cai[53] provide a recent tutorial of this kind of data and analysis.

Immortal Time Bias in Cohort Studies

Sometimes interest in natural history leads to a natural comparison of the disease patterns of individuals who experience an intermediate event between disease initiation and an expression outcome. Levesque *et al.*[54] consider the example of whether the use of statins affects the onset of insulin treatment in diabetics. Statin users were identified by the start of

a new (long-term) prescription at any point after diabetes diagnosis or other suitable cohort entry time. Immortal time bias arises because the statin user group counts the follow-up time between diagnosis and statin prescription even though no outcome event can possibly happen in this period (otherwise they would have experienced the outcome earlier as non-statin users). A similar example focuses on the mortality experience of individuals with a single primary melanoma as compared to those who experience multiple primary melanomas. Using the diagnosis time of the first occurring melanoma as the time origin to measure subsequent survival necessarily introduces immortal time bias and will make the survival experience of the multiple primary melanoma group appear more favorable.[55] This kind of bias is closely related to lead time bias associated with screening tests as noted earlier, and is usually best accommodated through use of time-varying covariates in a time to event analysis.

Case-Control Studies

As noted earlier, for rare diseases most forms of cohort studies are prohibitively expensive because of the very large sample sizes required in order to observe a reasonable number of events and because follow-up to ascertain the timing of events is itself costly. In such cases, different definitions of what is meant by a “case” allow investigation of either progression from a healthy state to disease onset or progression from disease onset to disease expression. Of course, there are significant issues raised by case-control sampling. The first is that it is not possible for case-control data to provide estimates of population incidence due to oversampling of cases, and this necessarily limits this design with regard to a basic description of the natural history of a disease in a population. However, relative comparisons of natural history parameters across subgroups are possible with careful analysis. Although relative changes in incidence across subgroups (as defined by individual covariates) are estimable[56, 57], there is increased potential risk for bias in measuring relevant covariate information in that this may be ascertained retrospectively, often many years prior to the event that defines case status. It is important to note that identification of “cases” may depend on the severity of the disease and may not reflect the entire spectrum of clinical disease; this is particularly true if hospitalizations, for example, are used to ascertain case status. Of course, this may be successfully addressed if there is external knowledge of the populations from which cases and controls are selected, in particular the natural incidence proportion for the disease in question and thus knowledge of case and control sampling fractions.[58,59]

Case-Only Studies

The *case-only* or *case distribution* design exploits comparison of an exposure distribution amongst a random sample of cases to a theoretical or known population distribution. See Greenland[60] for an overview of these and other related designs and their analysis. The data from such studies can be used to compare risks across different levels of exposure. These methods have been widely used in case-genotype studies where genotypic information on a series of cases is compared to a known or theoretically supported genotype distribution in the population from which the cases arose. The methods are closely related to techniques associated with the design and analysis of case-cohort studies[61] where the primary

difference is that, in the latter situation, only a sample of the source population is available from which to extract information on the population exposure distribution. McKeown, Yan & Jewell applied similar methods for estimation of age-specific attack rates during the 2009 epidemic of H1N1 influenza.[62] The design is widely used to investigate multiplicative effect modifications of genetic effects by environmental exposures,[63] although the approach requires independence between the two risk factors and has other limitations.[64] A variant of the case-only design is the case-crossover design where exposure is measured on the same cases at two or more different points in time and used therefore to study transient risk factors.[65, 66] The approach is popular in pharamcoepidemiology to study the association of temporal drug exposures to acute outcomes.[67] It is important to stress that case-only studies usually require stringent assumptions to provide reasonable population natural history information. An enlightening example is a study of deaths only as way of elucidating long-term outcomes for individuals with non-metastatic prostate cancer[68], and the subsequent debate about its validity.[69]

Hybrid Studies

There are a considerable number of variants of these basic types of studies that are all motivated by specific natural history questions of interest. For example, the Canadian Study of Health and Aging[70] was based on a cross-sectional sample of Canadians, aged 65 or older, in 1991 with sampled individuals followed forward solely for mortality (exploiting death certificate searches). Of course, the study was completed before almost half of the sample had died so that their ages at death were right censored. However, at recruitment, subjects were screened for the presence of dementia with age at onset retrospectively ascertained using appropriate methods. On the other hand, as noted, the onset of dementia after recruitment was not observed. Figure A.1 of Carone, Asgharian and Jewell[71] schematically describes this complex sampling method. In addition, as might be expected, the sample was based on stratification factors, including age, and these effects also need to be accommodated.

Discussion

This brief commentary introduces some of the key statistical challenges to collecting data on natural history information for samples of individuals and relating such to key parameters of interest. Clearly, it barely scratches the surface of current research on such topics. This is a rich area of study that cries out for a much longer synthesis of methods and ideas. It is important to note that many major statistical software packages now include methods to address some of the statistical questions raised here, although great caution should be used due to the complex sampling and bias questions associated with most natural history studies. Specific topical challenges include methods to exploit disease registries and other large medical care databases to obtain natural history trends and variability.

Acknowledgments

Thanks are due to Professors Niels Keiding, Marco Carone, and James Hanley for drawing my attention to a number of salient and important references. Funding support was provided through an NIH grant #UWSC7526.

References

1. Posada de la Paz M, Villaverde A, Alonso V, János S, Zurriaga Ó, Pollán M, Abaitua-Borda V. Rare diseases epidemiology research. Adv Exper Medic And Biol. 2010; 686:17–39.
2. Brookmeyer R. Statistical problems in epidemiologic studies of the natural history of disease. Envir Health Persp. 1990; 87:43–49.
3. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM458485.pdf
4. Nelson, HD.; Cantor, A.; Humphrey, L.; Fu, R.; Pappas, M.; Daeges, M.; Griffin, J. Evidence Synthesis. Agency for Healthcare Research and Quality; Rockville: 2016. Screening for breast cancer: A systematic review to update the 2009 U.S. Preventive Services Task Force recommendation. AHRQ Publication No. 14-05201-EF-1
5. Whitlock, EP.; Lin, J.; Liles, E.; Beil, T.; Fu, R.; O'Connor, E.; Thompson, RN.; Cardenas, T. Evidence Synthesis. Agency for Healthcare Research and Quality; Rockville: 2008. Screening for colorectal cancer: An updated systematic review. AHRQ Publication No. 08-05124-EF-1
6. Andersen PK, Keiding N. Multi-state models for event history analysis. Statist Meth Med Res. 2002; 11:91–115.
7. Farr W. On prognosis. British Medical Almanack. 1938; (Supplement 199–216):199–208. reproduced in *Soz. Praventiv. Med.* 48(4), 219–224 (2003).
8. Farr W. On prognosis. British Medical Almanack. 1938; (Supplement 199–216):208–216. reproduced in *Soz. Praventiv. Med.* 48(5), 279–284 (2003).
9. Gerstman B. Comments regarding "On prognosis" by William Farr (1838), with reconstruction of his longitudinal analysis of smallpox recovery and death rates. Soz Praventiv Med. 2003; 48(5): 285–289.
10. Silverman, WA. Human Experimentation. Vol. Chapter 4. Oxford University Press; Oxford: 1985.
11. Jayachandran S, Lleras-Muney A, Smith KV. Modern medicine and the twentieth century decline in mortality: Evidence on the impact of sulfa drugs. American Economic Journal: Applied Economics. 2(2):118–146.
12. Brookmeyer R, Corrada MM, Curriero FC, Kawas C. Survival following a diagnosis of Alzheimer Disease. Arch Neurol. 2002; 59:1764–1767. [PubMed: 12433264]
13. Carone, M. PhD Thesis. Vol. Chapter 5. Johns Hopkins University; 2010. Statistical Analysis of Cross-Sectional Survival Data.
14. Gardner JW, Sanborn JS. Years of potential life lost (YPLL)—What does it measure? Epidemiology. 1990; 1(4):322–329. [PubMed: 2083312]
15. Johnson E, Brookmeyer R, Ziegler-Graham K. Modeling the effect of Alzheimer's Disease on mortality. Int J of Biostatistics. 2007; 3(1) Article 13.
16. Lexis, W. Einleitung in die Theorie der Bevölkerungsstatistik. Trübner; Strassburg: 1875.
17. Keiding, N. Encyclopedia of Biostatistics. Vol. 3. John Wiley & Sons; Chichester, England: 1998. Lexis diagram.
18. Brinks R, Landwehr S, Fischer-Betz R, Schneider M, Giani G. Lexis diagram and illness-death model: simulating populations in chronic disease epidemiology. PLoS ONE. 2014; 9(9):e106043. [PubMed: 25215502]
19. Keiding N. Statistical inference in the Lexis diagram. Phil Trans Royal Soc Lond A. 1990; 332:487–509.
20. Nielsen, B.; Nielsen, JP. Identification and forecasting in mortality models. The Scientific World Journal. 2014. Article ID 347043: http://dx.doi.org/10.1155/2014/347043
21. Weedon-Fekjær H, Romundstat PR, Vatten LJ. Modern mammography screening and breast cancer mortality: population study. BMJ. 2014; 348:g3701. [PubMed: 24951459]
22. Hutchinson GB, Shapiro S. Lead time gained by diagnostic screening for breast cancer. JNCI. 1968; 41:665–673. [PubMed: 5677315]
23. Feinstein AR, Sosin DM, Wells CK. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer. NEJM. 1985; 312(25):1604–1608. [PubMed: 4000199]

24. Woods WG, Tuchman M, Robison LL, Bernstein M, Leclerc J-M, Brisson LC, Brossard J, Hill G, Shuster J, Luepker R, Byrne T, Weitzman S, Bunin G, Lemieux B. A population-based study of the usefulness of screening for neuroblastoma. Lancet. 1996; 348(9043):1682–1687. [PubMed: 8973429]
25. Barker DJP, Osmond C, Winter PD, Margetts B, Simmonds SJ. Weight in infancy and death from ischaemic heart disease. Lancet. 1989; 334:577–580. DOI: 10.1016/S0140-6736(89)90710-1 [PubMed: 2570282]
26. Roseboom TJ, van der Meulen JHP, Ravelli ACJ, Osmond C, Barker DJP, Bleker O. Effects of prenatal exposure to the Dutch famine on adult disease in later life: an overview. Molec Cellular Endocrinology. 2001; 185:93–98.
27. Sun, J. The Statistical Analysis of Interval-censored Failure Time Data. Springer-Verlag; New York: 2006.
28. Peto R. Experimental survival curves for interval-censored data. J Royal Stat Soc C. 1973; 22(1): 86–91.
29. Keiding N. Age-specific incidence and prevalence: A statistical perspective (with discussion). J Royal Statist Soc. 1991; 154:371–412.
30. Andersen PK, Keiding N. Multi-state models for event history analysis. Stat Meth Med Res. 2002; 11:91–115.
31. Keiding N. Event history analysis and the cross-section. Statistics in Med;. 2006; 25:2343–2364.
32. Collett, D. Modelling Survival Data in Medical Research. 3. Chapman & Hall/CRC Press; Boca Raton, Florida: 2014.
33. Hajnal J. Age at marriage and proportions marrying. Population Studies. 1953; 7(2):111–136.
34. Jewell, NP.; Emerson, R. Handbook of Survival Analysis. Chapman & Hall/CRC Press; Boca Raton, Florida: 2013. Current status data: An illustration with data on avalanche victims; p. 391-412.
35. Jewell NP, Shiboski S. Statistical analysis of HIV infectivity based on partner studies. Biometrics. 1990; 46(4):1133–1150. [PubMed: 2085629]
36. McKeown K, Jewell NP. Misclassification of current status data. Lifetime Data Analysis. 2010; 16:215–230. [PubMed: 20157848]
37. Jewell NP, Petito LC. Misclassified group tested current status data., under revision. Biometrika. 2016
38. Brookmeyer, R.; Gail, MH. Biometrics. Vol. 43. Wiley; Hoboken: 1987. Biases in prevalent cohorts; p. 739-749.2002
39. Wang M-C, Brookmeyer R, Jewell NP. Statistical models for prevalent cohort data. Biometrics. 1993; 49:1–11. [PubMed: 8513095]
40. Magnus P, Jaakola JK. Secular trend in the occurrence of asthma among children and young adults: critical appraisal of repeated cross sectional surveys. BMJ. 314:1795. http://dx.doi.org/10.1136/bmj.314.7097.1795. [PubMed: 9224081]
41. Caplan LS, Lane DS, Grimson R. The use of cohort vs repeated cross-sectional sample survey data in monitoring changing breast cancer screening practices. Prev Med. 1995; 24(6):553–556. [PubMed: 8610077]
42. Keurentjes JC, Fiocco M, Schreurs BW, Pijls BG, Nouta KA, Nelissen RGHH. Revision surgery is overestimated in hip replacement. Bone Joint Res. 2012; 1:258–262. [PubMed: 23610656]
43. Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data.
44. Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. J Amer Statist Assoc. 94:496–509. DOI: 10.2307/2670170
45. Haller B, Schmidt G, Ulm K. Applying competing risks regression models: An overview. Lifetime Data Analysis. 2013; 19:33–58. DOI: 10.1007/s10985-012-9230-8 [PubMed: 23010807]
46. Xu J, Kalbfleisch JD, Tai B. Statistical analysis of illness-death processes and semicompeting risks data. Biometrics. 2010; 66:716–725. [PubMed: 19912171]
47. Varadhan R, Xue Q-L, Bandeen-Roche K. Semicompeting risks in aging research: methods, issues, needs. Lifetime Data Analysis. 2014; 20(4):538–562. [PubMed: 24729136]

48. McLain AC, Sundaram R, Louis GMB. Joint analysis of longitudinal and survival data measured on nested timescales by using shared parameter models: an application to fecundity data. Appl Statist. 2015; 64(2):339–357.
49. Tsiatis AA, Davidian M. An overview of joint modeling of longitudinal and time-to-event data. Statistica Sinica. 2004; 14:793–818.
50. Kurland BF, Johnson LL, Egleston BL, Diehr PH. Longitudinal data with follow-up truncated by death: Match the analysis method to research aims. Statist Sci. 2009; 24(2):211–222. DOI: 10.1214/09-STS293
51. Asar O, Ritchie J, Kalra PA, Diggle PJ. Joint modeling of repeated measurement and time-to-event data: An introductory tutorial. Int J Epidemiol. 2015; 44:334–344. DOI: 10.1093/ije/dyu262 [PubMed: 25604450]
52. Rizopoulos, D. Joint Models for Longitudinal and Time-to-Event Data. Chapman & Hall/CRC Press; Boca Raton, Florida: 2012.
53. Amorim LDAF, Cai J. Modelling recurrent events; A tutorial for analysis in epidemiology. Int J Epidemiol. 2014; 44:324–333. DOI: 10.1093/ije/dyu222 [PubMed: 25501468]
54. Lévesque LE, Hanley JA, Kezouh A, Suissa S. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. BMJ. 2010; 340:b5087. [PubMed: 20228141]
55. Doubrovsky A, Menzies SW. Enhanced survival in patients with multiple primary melanoma. Arch Dermatol. 2003; 139(8):1013–1018. [PubMed: 12925389]
56. Breslow, NE.; Day, NE. Statistical Methods in Cancer Research. Vol. #32. International Agency for Research on Cancer; 1980.
57. Jewell, NP. Statistics for Epidemiology. Chapman & Hall/CRC Press; Boca Raton, Florida: 2004.
58. Miettinen O. Estimability and estimation in case-referent studies. Amer J Epidemiology. 103(2): 226–235.
59. Rose S, van der Laan MJ. Simple optimal weighting of cases and controls in case-control studies. Int J Biostatistics. 2008; 4(1) Article 19.
60. Greenland S. A unified approach to the analysis of case-distribution (case-only) studies. Statist Med. 1999; 18:1–15.
61. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika. 1986; 73:1–11. DOI: 10.1093/biomet/73.1.1
62. McKeown K, Yan P, Jewell NP. Age-standardized measures of attack rates for confirmed cases of the 2009 H1N1 influenza pandemic with implications for planning. 2011 Unpublished manuscript.
63. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Statist Med. 1994; 13:153–162.
64. Albert PS, Ratnasinghe D, Tangrea J, Waholder S. Limitations of the case-only design for identifying gene-environment interactions. Amer J Epidemiology. 2001; 154(8):687–693.
65. Maclure M. The case-crossover design: a method for studying transient effects on the risk of acute events. Amer J Epidemiology. 1991; 133:144–153.
66. Maclure M, Mittleman MA. Should we use a case-crossover design? Ann Rev Public Health. 2000; 21:193–221. [PubMed: 10884952]
67. Delaney JA, Suissa S. The case-crossover study design in pharmacoepidemiology. Stat Methods Med Res. 2009; B18(1):53–65. [PubMed: 18765504]
68. Aus G, Hugosson J, Norlén L. Long-term survival and mortality in prostate cancer treated with noncurative intent. J Urology. 1995; 154(2):460–465.
69. Abrahamsson PA, Adami H, Taube A, Kim K, Zelen M. Re: Long-term survival and mortality in prostate cancer treated with noncurative intent. J Urology. 1996; 155(1):296–297.
70. McDowell I, Hill G, Lindsay J, Helliwell B, Costa L, Beattie B, Tuokko H, Hertzman C, Gutman G, Parhad I. Canadian study of health and aging: Study methods and prevalence of dementia. Canadian Medical Association Journal. 1994; 150:899–912. [PubMed: 8131123]

71. Carone M, Asgharian M, Jewell NP. Estimating the lifetime risk of dementia in the Canadian elderly population using cross-sectional cohort survival data. J Amer Statist Assoc. 2014; 109:24–35.

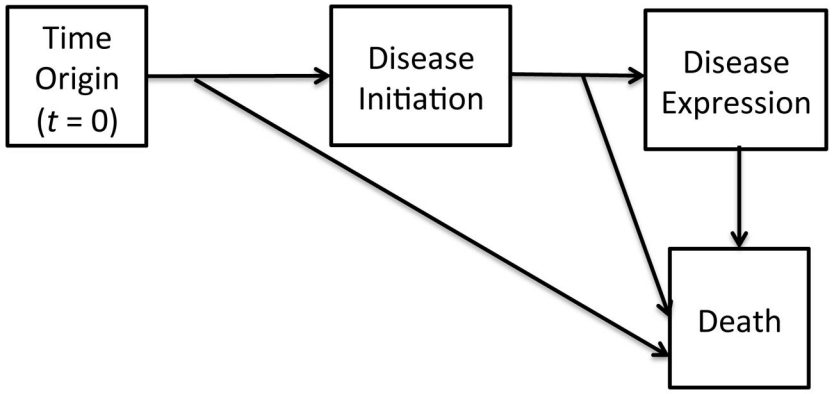

Figure 1.
Multi-state schematic for disease progression from being disease–free to death from various causes.

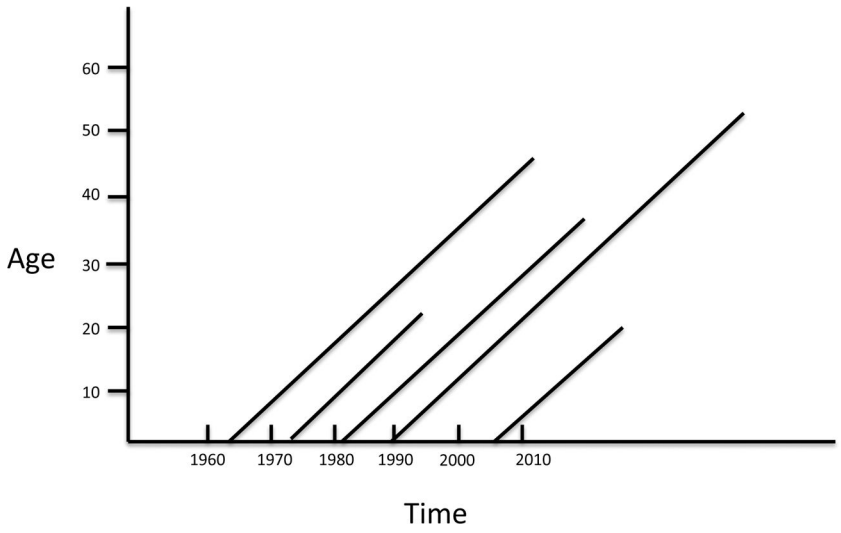

Figure 2.

Lexis diagram showing five ‘lives’ from ‘birth’ to ‘death.’ Each life is represented by a line of unit slope that begins at a point in time on the X-axis (‘birth’) until ‘death’ at a certain age represented on the Y-axis. Note that ‘birth’ might refer to disease diagnosis/onset, for example. Adapted from Keiding (1998).[17]

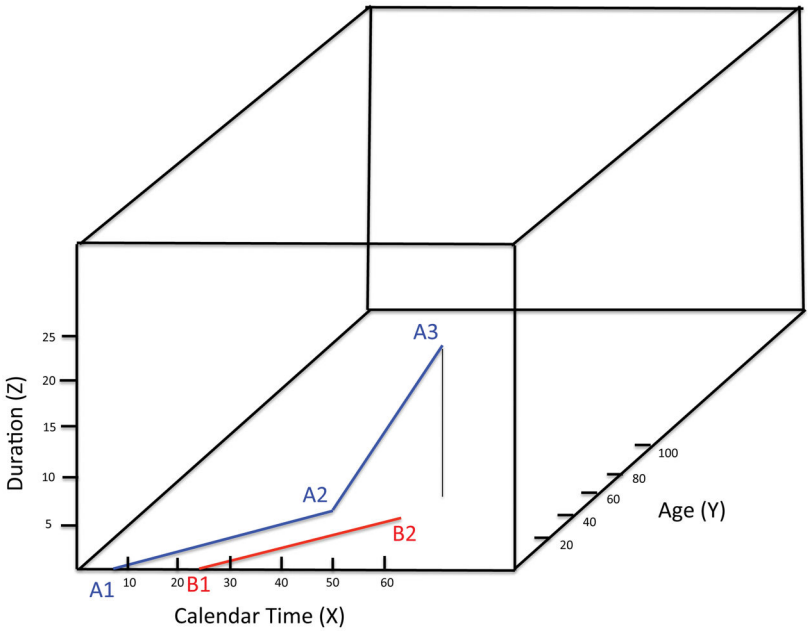

Figure 3.

Three-dimensional Lexis diagram representing ‘lives’ from ‘birth’ in calendar time represented on the X-axis, until an intermediate stage (e.g. disease onset) at age represented on the Y-axis, where the line now rises in the third-dimension until ‘death’ at a certain age having experienced a disease duration represented on the Z-axis. For example, the individual represented in blue is ‘born’ at time A1, contracts the disease at B1 at the noted age, and dies at time A3. On the other hand, the individual represented in red never contracts the disease and therefore spends their entire life in the X–Y plane. Adapted from Brinks *et al.*[18]