Epidemiology / Épidémiologie

Biostatistics and epidemiology: measuring the risk attributable to an environmental or genetic factor

Jacques Benichou

Unité de biostatistique, CHU de Rouen & Inserm U 657, Institut hospitalo-universitaire de recherche biomédicale, Université de Rouen, 1, rue de Germont, 76031 Rouen cedex, France

Abstract

Disease frequency is measured through estimating incidence rates or disease risk. Several measures are used for assessing exposure–disease association, with adjusted estimates based on standardization, stratification, or more flexible regression techniques. Several measures are available to assess an exposure impact in terms of disease occurrence at the population level, including the commonly used attributable risk (AR). Adjusted AR estimation relies on stratification or regression techniques. Sequential and partial ARs have been proposed to handle the situation of multiple exposures and circumvent the associated non-additivity problem. Despite remaining issues in properly interpreting AR, AR remains a useful guide to assess prevention strategies. ***To cite this article: J. Benichou, C. R. Biologies 330 (2007).***
© 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

Résumé

Biostatistique et épidémiologie : la mesure du risque attribuable à un facteur environnemental ou génétique. La mesure de la fréquence d'une maladie repose sur les concepts d'incidence et de risque. Plusieurs mesures d'association entre exposition et maladie existent. Leur estimation ajustée repose sur la standardisation, la stratification, ou les méthodes plus flexibles de régression. Le risque attribuable (RA) est une mesure d'impact populationnel d'une exposition sur la survenue de nouveaux cas. Son estimation ajustée repose sur la stratification ou la régression. Les RAs séquentiels et partiels permettent de prendre en compte plusieurs expositions et le problème associé de non-additivité. Malgré certaines questions d'interprétation, le RA demeure un guide utile à l'évaluation de stratégies de prévention. ***Pour citer cet article : J. Benichou, C. R. Biologies 330 (2007).***
© 2007 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

Keywords: Incidence; Hazard; Risk; Measure of association; Measure of impact; Attributable risk; Sequential attributable risk; Stratification; Regression; Causality

Mots-clés : Incidence ; Risque instantané ; Risque ; Mesure d'association ; Mesure d'impact ; Risque attribuable ; Risque attribuable séquentiel ; Stratification ; Régression ; Causalité

1. Introduction

A major aim of epidemiologic research is to measure disease occurrence in relation to various character-

istics such as exposure to environmental, occupational, or lifestyle risk factors, genetic traits or other features. The generic term exposure will be used throughout this chapter to denote these characteristics. We will start with reviewing various measures that are at the root of quantitative epidemiologic thinking. These include measures that quantify disease occurrence, associations between disease occurrence and exposures as well as their consequences in terms of disease risk (Section 2). Emphasis will be placed on measures based on occurrence of new disease cases, referred to as disease incidence. Measures based on disease prevalence, i.e., considering previously existing disease cases as well as new cases, will be mentioned only in passing.

In Section 3–7, we will focus on the main measure of impact at the population level, namely attributable risk. This measure will be introduced in some detail in Section 3. Then, we will successively review three specific problems regarding attributable risk. First, we will consider adjusted attributable risk estimation from epidemiologic study data in Section 4, an issue that has generated intensive methodological research in the last 20 years, resulting in essentially satisfactory solutions. Second, we will discuss the lack of additivity of attributable risk contributions for separate exposures and present a possible solution in Section 5. Third, we will examine conceptual issues involved in interpreting attributable risk estimates in Section 6. Final remarks will follow in Section 7.

2. Rates, risks and measures of association

2.1. Incidence and hazard rates

The incidence rate of a given disease is the number of persons who develop the disease (number of incident cases) among subjects at risk of developing the disease in the source population over a defined period of time or age. Incidence rates are not interpretable as probabilities. While they have a lower bound of zero, they have no upper bound. Units of incidence rates are reciprocal of person-time, such as reciprocals of person-years or multiples of person-years (e.g., 100 000 person-years). For instance, if five cases develop from the follow-up of 50 subjects and for a total follow-up time of two years per subject, the incidence rate is $5/100 = 0.05$ cases per person-year (assuming an instantaneous event with immediate recovery and all 50 subjects being at risk until the end of the observation period). Usually, incidence rates are assessed over relatively short time periods compared with the time scale for disease development, e.g., intervals of five years for chronic diseases

with an extended period of susceptibility, such as many cancers.

Synonyms for incidence rate are average incidence rate, force of morbidity, person-time rate, or incidence density [1], the last term reflecting the interpretation of an incidence rate as the density of incident case occurrences in an accumulated amount of person-time [2]. Mortality rates (overall or cause-specific) can be regarded as a special case of incidence rates, the outcome considered being death rather than disease occurrence.

Incidence rates can be regarded as estimates of a limiting theoretical quantity, namely the hazard rate, $h(t)$, also called the incidence intensity or force of morbidity. The hazard rate at time t, $h(t)$, is the instantaneous rate of developing the disease of interest in an arbitrarily short interval Δ around time t, provided that the subject is still at risk at time t (i.e., has not fallen ill before time t). It has the following mathematical definition:

$$h(t) = \text{limit}_{\Delta \downarrow 0} \Delta^{-1} \Pr(t \leqslant T < t + \Delta \mid t \leqslant T) \qquad (1)$$

where T is the time period for the development of the disease considered and Pr denotes probability. Indeed, for time intervals in which the hazard rate can be assumed constant, the incidence rate as defined above represents a valid estimate of the hazard rate. Thus, this result applies when piecewise constant hazards are assumed, which can be regarded as realistic in many applications, especially when reasonably short time intervals are used, and leads to convenient estimating procedures, e.g., based on the Poisson model.

Strictly speaking, incidence and hazard rates do not coincide. Hazard rates are formally defined as theoretical functions of time, whereas incidence rates are defined directly as estimates and constitute valid estimates of hazard rates under certain assumptions (see above).

From the definitions above, it ensues that individual follow-up data are needed to obtain incidence rates or estimate hazard rates. The cohort design that incurs follow-up of subjects with various profiles of exposure is the ideal design to obtain incidence or hazard rates for various levels or profiles of exposure, i.e., exposure-specific incidence or hazard rates. In many applications, obtaining exposure-specific incidence rates is not trivial, however. Indeed, several exposures are often considered, some with several exposed levels and some continuous. Moreover, it may be necessary to account for confounders or effect-modifiers. Hence, estimation often requires modelling. Alternatively to the cohort design, in the absence of individual follow-up data, person-time at risk can be estimated as the time period width times the population size at midpoint. Such estimation makes