PUBLIC HEALTH 153 (2017) 103–106

Available online at www.sciencedirect.com

Public Health

journal homepage: www.elsevier.com/puhe

Short Communication

Selection bias in ecological studies

V. Rousson, P.C. Rosselet*, F. Paccaud

Institute of Social and Preventive Medicine (IUMSP), Lausanne University Hospital and University of Lausanne, Route de La Corniche 10, 1010 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 27 January 2017
Received in revised form
18 June 2017
Accepted 1 September 2017

An ecological study is an observational epidemiological study which uses aggregated data instead of individual data, where the aggregation is usually done at some geographical level such as a town, a region, or a country.[1] They are typically carried out when an exposure of interest is fully defined at the group level.[2] A classical example is the association between air pollution and a specific health outcome. While it is difficult (or even meaningless) to measure individual exposure to air pollution, it is easier to measure air quality in a region. In order to investigate the association between air pollution and the occurrence of some disease, one may correlate the available measures of air quality with the prevalence (or incidence) of the disease over a set of regions. This approach faces several limitations; in the literature about ecological studies, one can find numerous warnings against the 'ecological fallacy', which states that a correlation at the group level is not of the same magnitude (and not always with the same sign) than a correlation at the individual level,[3] and about confounding.[4] The present communication aims to be a reminder about another serious limitation of ecological studies which is rarely mentioned in the literature: selection bias. Yet, as illustrated below, a deception due to a selection bias might be still more spectacular than any confounding issue.

In individual-based studies, selection bias refers to the situation where a sample of individuals is not representative (i.e. cannot be assumed to have been randomly selected) of a population of interest, such that the results of the study cannot be generalized to this population without further assumptions. While a definitive assessment remains difficult to achieve, it is often possible to evaluate whether and to what extent selection bias has occurred in an individual study, e.g. by comparing characteristics from the sample with known quantities from the population, such as age and gender distribution. However, this kind of evaluation is more difficult to undertake in an ecological study, since a population of aggregated data, e.g. a population of countries of interest (of which the sample of countries which is analyzed should be representative), is rarely defined.

One may consider as an example an ecological study which examined correlations between cardiovascular risk factors and stroke incidence over 18 regions from 10 countries: China, Denmark, Finland, Germany, Italy, Lithuania, Poland, Russia, Sweden, and Yugoslavia.[5] Here, a definition of what might be the population of regions or countries of interest is not obvious.[6] If a population cannot be defined, it is also not possible to evaluate or even discuss the presence and size of a potential selection bias. Similarly, a recent paper conducted an ecological study which showed a significant positive correlation between the average yearly chocolate consumption and the number of Nobel Prizes (per 10 million inhabitants) calculated over 23 countries.[7] Besides various and obvious confounding issues,[8] the countries included in this ecological study had not been selected at random from a prespecified population of countries of interest. In fact, only countries with at least one Nobel Prize were included. By including the many countries without any Nobel Prize (but with a non-zero chocolate consumption), the correlation is likely to become smaller.

* *Corresponding author*. Route de la Corniche 10, 1010 Lausanne, Switzerland.
E-mail addresses: valentin.rousson@chuv.ch (V. Rousson), patricia.rosselet@gmail.com (P.C. Rosselet), Fred.Paccaud@gmail.com (F. Paccaud).
http://dx.doi.org/10.1016/j.puhe.2017.09.001

In order to avoid a difficult discussion about bias selection in an ecological study, some researchers may be tempted to renounce making a generalization of the results to a broader population of countries by arguing that the countries included in the study are the only countries of interest. In contrast to individual studies, where there is absolutely no point in mentioning somewhere the names of the individuals who were selected in the sample, it is common to see the names of the countries included in an ecological study printed on scatter plots, a sign of the specific interest that researchers usually have in these countries. However, a correlation calculated from a sample of countries where no random selection has been carried out might be totally misleading.

To illustrate this point, one can consider the relationship between military expenditure (expressed as fraction of Gross Domestic Product, or GDP, calculated in 1984) and life expectancy in 30 non-European countries (except USSR), using data from the Canadian website http://home.ca.inter.net/paulye/GEODATA95.htm#SURVIE. This is a typical example of an ecological study which could not have been undertaken at the individual level, since military expenditure is a variable defined only at the national level. The corresponding scatter plot is shown on Fig. 1a. One gets a pretty large positive (and

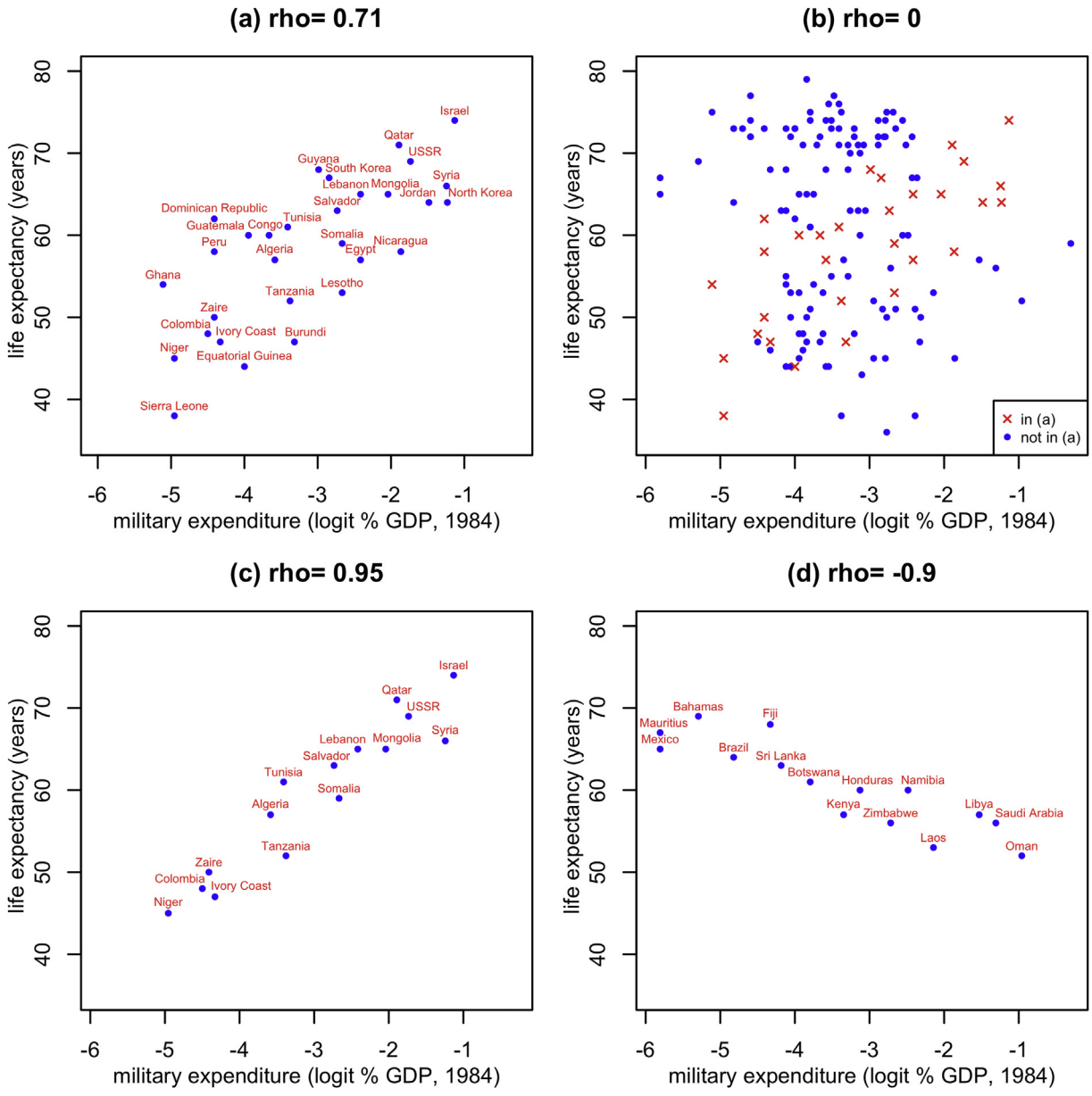

Fig. 1 – Life expectancy vs military expenditure in different selections of countries: (a) in a subset of $n = 30$ countries; (b) in all $n = 146$ countries with an army and available data (where the countries selected in panel (a) are indicated by a cross, other countries by a dot); (c) and (d) in two disjoint subsets of $n = 15$ countries. Military expenditure is expressed as the logit of the fraction of Gross Domestic Product (GDP) calculated in 1984. Spearman's correlation's rho is also provided in each panel. Source: http://home.ca.inter.net/paulye/GEODATA95.htm#SURVIE.