Annals of Epidemiology 25 (2015) 147–154

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

A history of the population attributable fraction and related measures

Charles Poole MPH, ScD *

Department of Epidemiology, University of North Carolina, Chapel Hill

ARTICLE INFO

Article history:
Received 9 November 2014
Accepted 16 November 2014
Available online 11 December 2014

Keywords:
Population attributable fraction
Exposed attributable fraction
Rate difference
Population rate difference
Caseload difference

ABSTRACT

Purpose: Since Doll published the first PAF in 1951, it has been a mainstay. Confusion in terminology abounds with regard to these measures. The ability to estimate all of them in case-control studies as well as in cohort studies is not widely appreciated.
Methods: This article reviews and comments on the historical development of the population attributable fraction (PAF), the exposed attributable fraction (EAF), the rate difference (ID), the population rate (or incidence) difference (PID), and the caseload difference (CD).
Results: The desire for PAFs to sum to no more than 100% and the interpretation of the complement of a PAF as the proportion of a rate that can be attributed to other causes are shown to stem from the same problem: a failure to recognize the pervasiveness of shared etiologic responsibility among causes. A lack of appreciation that "expected" numbers of cases and deaths are not actually the numbers to be expected when an exposure or intervention appreciably affects person-time denominators for rates, as in the case of smoking and nonnormal body mass, makes many CD estimates inflated. A movement may be gaining momentum to shift away from assuming, often unrealistically, the complete elimination of harmful exposures and toward estimating the effects of realistic interventions. This movement could culminate in a merger of the academic concept of transportability with the applied discipline of risk assessment.
Conclusions: A suggestion is offered to pay more attention to absolute measures such as the rate difference, the population rate difference, and the CD, when the latter can be validly estimated and less attention to proportional measures such as the EAF and PAF.

Introduction

The population attributable fraction (PAF) answers the following question: suppose every member of a population who was not in the most favorable level of an exposure or some other condition or event with regard to an adverse outcome had been shifted into that level. By what proportion would the entire population's rate, hazard, risk, prevalence, or caseload have been reduced?

The brief history to follow of the PAF and related measures begins in the early 1950s and ends, not altogether arbitrarily, at the close of the 1980s. The primary focus is on matters of interpretation, with some attention paid to the influence of rather basic features of study design and data analysis. Some statistical issues, such as those pertaining to sampling error and covariate adjustment, are not addressed.

1951 to 1953

The first PAF

In 1951, Doll [1] estimated what appears to be the first published PAF in the epidemiologic literature. He used the cases from his preliminary (1948–1949) case-control study with Hill [2] to form the numerators for lung cancer incidence rates in Greater London. To obtain the denominators, he apportioned census figures by the smoking distribution in the study's control group. Within each age stratum, he multiplied the total person-time at risk by the rate among the nonsmokers "to estimate the number of cases that would have been expected to occur if the entire population were nonsmokers" [1]. Summing across the strata, Doll "calculated that 41 cases would have been expected to occur against the 533 which were actually observed. It is therefore estimated that more than 90% of the cases of carcinoma of the lung can be attributed to smoking" [1].

The second and third PAFs

Two years later, Doll [3] generated another PAF using results from the full case-control study (1948–1952) [4]. This time, the

* Corresponding author. Department of Epidemiology, University of North Carolina, Chapel Hill, NC 27599-7435. Tel.: +1 919 966 9294; fax: +1 919 966 2089.
E-mail address: cpoole@unc.edu.

http://dx.doi.org/10.1016/j.annepidem.2014.11.015
1047-2797/

target population consisted of all adults aged 25 to 74 years in England and Wales in 1950. Within strata of age, gender, and locale, Doll used the controls from the case-control study as before to break down the person-time at risk by smoking. Then he applied the same approach to the stratified lung cancer deaths across England and Wales, apportioning them according to the smoking distribution of the study cases. An illustrative set of Doll's calculations for one subgroup is listed in Table 1.

After tallying the results across subgroups, Doll estimated that 1875 lung cancer deaths throughout England and Wales "would have been expected, in the absence of smoking" in comparison with the observed figure of 11,189: PAF = (11,189 − 1,875)/11,189 = 83%. This time, he focused his interpretation on the complement of the PAF, stating that "it is, therefore, concluded that about one in five of the lung cancer deaths in persons aged 25 to 74 in 1950, were attributable to causes other than lung cancer" [3].

More early PAFs

Later that same year, Sadowsky et al. [5] showed that Doll had independently rediscovered the same approach Cornfield had described in 1951 [6] for using exposure distributions in a case-control study to estimate smoking-specific rates in an external target population (i.e., in a population other than the source population for the cases and controls). Sadowsky et al. and subsequent authors [7,8] referred to this approach as "the method of Cornfield." In using several case-control studies to generate smoking-specific lung cancer rates, they chose as their target population, as he had done, the combined population of the 10 metropolitan areas of the United States in the study [9–11] that later came to be known as the First National Cancer Survey [12].

Levin's PAF formula

The year 1953 also saw the publication of Levin's [13] novel approach to estimating PAFs from rate ratios rather than from rates and rate differences. As his derivation was somewhat cryptic, the clearer one Leviton [14] gave two decades later is reiterated here.

Let I represents the overall rate in a population, p the proportion of the person-time at risk in the level of a binary exposure variable with the higher rate, I_1 the rate in that level, and I_0 the rate in the level with the lower rate. If the outcome measure is a hazard, risk, or prevalence, I may be replaced by H, R, or P, respectively. The overall rate is a weighted average of the exposure-specific rates, with weights proportional to the person-time at risk: $I = pI_1 + (1 - p)I_0$. This formula simplifies to $I = p(ID) + I_0$, where $ID = I_1 - I_0$ is the rate difference comparing the two exposure levels.

The population rate (incidence) difference is $PID = I - I_0 = p(ID)$. The PAF is the PID expressed as a proportion of I, $PAF = (I - I_0)/I$, or $PAF = p(ID)/[p(ID) + I_0]$. Dividing the numerator and denominator of this expression by I_0, we obtain $PAF = p(IR - 1)/[p(IR - 1) + 1]$, where $IR = I_1/I_0$. Illustrative calculations are shown in Table 1.

The closest Levin [13] came to naming the PAF was in a table legend, where he called it a "proportion attributable." He interpreted it, in his substantive example, as the "maximum proportion of lung cancer attributable to smoking." From the range of PAF estimates, he calculated with results from several case-control studies, Levin concluded that "tobacco smoking may be responsible for from 56% to 92% of lung cancer. If the latter figure is correct, elimination of smoking would almost eliminate lung cancer (other factors remaining the same). If 56% is nearer the true figure, then elimination of smoking would reduce lung cancer by about one half, if smoking is a truly causative agent" [13]. He interpreted the complement of the PAF similarly to Doll [3], observing that a PAF of 50% to 75% "would mean that environmental causes, other than smoking, should be looked for in 25% to 50% of the cases" [13].

Levin [13] also defined, for what seems to have been the first time, the exposed attributable fraction: $EAF = (IR - 1)/IR$. He gave no name to this measure and no derivation, but it is easily derived as the rate difference expressed as a proportion of the exposed rate: $EAF = ID/I_1$. Division of the numerator and denominator by I_0 yields $EAF = (IR - 1)/IR$. Levin described the EAF as "the proportion of lung cancer in smokers attributable to smoking." A calculation of this measure is illustrated in Table 1 as well.

1954 to 1959

Gefeller wrote that the PAF "fell into oblivion" [15] after 1953. Indeed, the citation to Levin's 1953 article [13] in a 1958 review of statistical methods in cancer research [16] was for another contribution entirely.

Another sign of early indifference to the PAF came in 1954, when the authors of two articles [7,8] used several case-control studies to break down an external target population's overall lung cancer rate

Table 1
Incident lung cancer cases and controls and lung cancer mortality in England and Wales, women aged 45 to 64 years, urban areas other than London, 1950, by smoking status [1]

Group	Measure	Value
Case-control study		
Cases	Proportion smokers	$p_c = 13/23 = 0.57$
Controls	Proportion smokers	$p = 36/125 = 0.29$
England and Wales		
Nonsmokers	Lung cancer deaths	$A_0 = 0.43\ (533) = 232$
	Person-years at risk	$T_0 = 0.71\ (3{,}507{,}000) = 2{,}496{,}984$
	Rate*	$I_0 = 9.3$
Smokers	Lung cancer deaths	$A_1 = 0.57\ (533) = 301$
	Person-years at risk	$T_1 = 0.29\ (3{,}507{,}000) = 1{,}010{,}016$
	Rate*	$I_1 = 29.8$
	Rate difference*	$ID = 29.8 - 9.3 = 20.5$
	"Expected" lung cancer deaths	$A_1^* = 1{,}010{,}016\ (0.000093) = 94$
	Exposed attributable fraction	$EAF = 20.5/29.8 = (301 - 94)/301 = 69\%$
All	Lung cancer deaths	$A = 533$
	Person-years at risk	$T = 3{,}507{,}000$
	"Expected" lung cancer deaths	$A^* = 3{,}507{,}000\ (0.000093) = 325$
	Rate*	$I = 15.2$
	Population rate difference*	$PID = 15.2 - 9.3 = 5.9$
	"Expected" lung cancer deaths	$3{,}507{,}000\ (0.000093) = 325$
	Population attributable fraction	$PAF = 5.9/15.2 = (533 - 325)/533 = 39\%$

* Lung cancer deaths per 100,000 person-years.